Control and Surveillance of Partially Observed Stochastic Epidemics in a Bayesian Framework

Hola Kwame Adrakey

Submitted for the degree of
Doctor of Philosophy

Heriot-Watt University

Department of Actuarial Mathematics and Statistics,
School of Mathematical and Computer Sciences.

December 20, 2016

Abstract

This thesis comprises a number of inter-related parts. For most of the thesis we are concerned with developing a new statistical technique that can enable the identification of the optimal control by comparing competing control strategies for stochastic epidemic models in real time. In the second part, we develop a novel approach for modelling the spread of Peste des Petits Ruminants (PPR) virus within a given country and the risk of introduction to other countries.

The control of highly infectious diseases of agriculture crops, animal and human diseases is considered as one of the key challenges in epidemiological and ecological modelling. Previous methods for analysis of epidemics, in which different controls are compared, do not make full use of the trajectory of the epidemic. Most methods use the information provided by the model parameters which may consider partial information on the epidemic trajectory, so for example the same control strategy may lead to different outcomes when the experiment is repeated. Also, by using partial information it is observed that it might need more simulated realisations when comparing two different controls. We introduce a statistical technique that makes full use of the available information in estimating the effect of competing control strategies on real-time epidemic outbreaks. The key to this approach lies in identifying a suitable mechanism to couple epidemics, which could be unaffected by controls. To that end, we use the Sellke construction as a latent process to link epidemics with different control strategies.

The method is initially applied on non-spatial processes including SIR and SIS models assuming that there are no observation data available before moving on to more complex models that explicitly represent the spatial nature of the epidemic spread. In the latter case, the analysis is conditioned on some observed data and inference on the model parameters is performed in Bayesian framework using the Markov Chain Monte Carlo (MCMC) techniques coupled with the data augmentation methods. The methodology is applied on various simulated data sets and to citrus canker data from Florida. Results suggest that the approach leads to highly positively correlated outcomes of different controls, thus reducing the variability between the effect of different control strategies, hence providing a more efficient estimator of their expected differences. Therefore, a reduction of the number of realisations required to

compare competing strategies in term of their expected outcomes is obtained.

The main purpose of the final part of this thesis is to develop a novel approach to modelling the speed of Pest des Petits Ruminants (PPR) within a given country and to understand the risk of subsequent spread to other countries. We are interested in constructing models that can be fitted using information on the occurrence of outbreaks as the information on the susceptible population is not available, and use these models to estimate the speed of spatial spread of the virus. However, there was little prior modelling on which the models developed here could be built. We start by first establishing a spatio-temporal stochastic formulation for the spread of PPR. This modelling is then used to estimate spatial transmission and speed of spread. To account for uncertainty on the lack of information on the susceptible population, we apply ideas from Bayesian modelling and data augmentation by treating the transmission network as a missing quantity. Lastly, we establish a network model to address questions regarding the risk of spread in the large-scale network of countries and introduce the notion of 'first-passage time' using techniques from graph theory and operational research such as the Bellman-Ford algorithm. The methodology is first applied to PPR data from Tunisia and on simulated data. We also use simulated models to investigate the dynamics of spread through a network of countries.

Acknowledgements

I feel extremely lucky that I met Bernd Schroers during my postgraduate studies in African Institute for Mathematical Sciences (AIMS). Apart from being a wonderful teacher, Bernd was the one that linked me to Kokouvi Gamado who later advised me to do a PhD and in particular to come to UK and be supervised by Gavin Gibson and George Streftaris. I will always be grateful to Kokouvi for this suggestion and his valuable guidance all these years.

Working with Gavin and George has been really exciting. Apart from their energy and their enthusiasm for research, their support and guidance throughout my PhD have been invaluable. I very much enjoyed discussions on the intuition behind the work in this thesis and many other subjects. I am grateful to Gavin for successfully securing funding for this thesis without which I would not be able to pursue my dream. Gavin and George have been encouraging and very patient while I was writing up this thesis and provided me with very constructive comments. I have to confess that I have never met such wonderful persons in my life.

I am grateful to the Actuarial Mathematics & Statistics department at Heriot-Watt University for providing funding.

I would like to thank the mathematical biology group in Plant Sciences, Cambridge University, especially Franco Neri who provided us with the data on 1995 Citrus canker outbreak in Florida. I would also like to thank EFSA for providing funding for the PPR modelling as part of the project MAPRA.

Last but not least, my dad and mum, Kodjo and Irene Adrakey, my brothers and sisters were extremely supportive throughout this period, despite me being abroad many thanks to them. A particular thank to my wife Afi Adetou Etonam for her unwavering encouragement: your belief in me is my first source of motivation.

My sincere appreciation to all my friends and colleagues at Heriot-Watt University in particular to Hove Kouevi, Ayawoa Dagbovie and Nadia Taou for the great moments we shared together.

There are so many people out there that made my life in Edinburgh very pleasant that I cannot name here. You all know who you are and I just want to say thanks for creating a family atmosphere around me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Throughout history, through the twentieth century until the present day, epidemics have been arguably the most important cause of suffering and death in human, animal and plant populations impacting on the politics and economics of many countries (Schumann, 1991, Chap 1). Therefore, finding the best understanding of the mechanism of epidemic spread and its control has become the biggest challenge that epidemiologists, statisticians and mathematicians face. Much work has been done in modelling and validating a variety of epidemiological models including models where hosts in the population are subdivided into different compartments. Relatively less, however, is known about how to design optimal controls given the available resources. The current Ebola epidemic in west Africa, the 2001 food and month disease in Britain (Ferguson et al., 2001; Chis-Ster et al., 2009) and the 1995 citrus canker outbreak in Miami (Graham et al., 2004) provide some illustrations of how problematic and controversial policy decisions for control might be.

Reducing the risk of propagation by targeting the best control strategies is therefore crucial for any disease. The aim of this thesis is to provide a general methodology for designing cost-effective control strategies so as to reduce or minimise the impact on a target population. In particular, we wish to answer two main questions:

1. How can we design computationally efficient approaches to assessing controls so that optimal approaches can be identified?

2. Given the available resources at our disposal, what is the best design that optimises their deployment across the population?

1.2 Structure of the Thesis

This thesis is divided into two parts. The principal part is mainly concerned with providing real-time control tools for any potential infectious disease outbreak based on removal/replacement of hosts in a spatially distributed population. This is particularly relevant to arboreal pathogens. Specifically, in the first chapter, we explain why in general it is important to provide mathematical models to explain the spread of an epidemic and ultimately help in identifying optimal controls. We briefly give the background of the type of models used in the literature covering deterministic and stochastic models. We then review how control has been carried out in previous work.

Chapter 2 begins by describing the specific models of interest in this framework. We mainly focus on stochastic models where hosts may or may not be spatially distributed. We then review two important approaches (the Gillespie algorithm and the Sellke construction) used to analyse and implement stochastic models for epidemics. We focus on representing different aspects of an epidemic process using the Sellke representation where the outcome is a deterministic function of the model parameters and the set of Sellke thresholds.

Chapter 3 is mainly focussed on reviewing the Bayesian approach. We review some of the previous applications of Bayesian data-augmentation methods particulary Markov Chain Monte Carlo (MCMC) using the general SIR epidemic model, and we consider a particular implementation of the approach that allows the imputation of additional events beyond the observation period.

Chapter 4 begins by motivating the coupling of epidemics using Sellke thresholds. In particular we develop and compare different control strategies based on removal or replacement of hosts using the idea of Sellke thresholds introduced in Chapter 2. This chapter is concerned with models taking account of the temporal dynamics of the epidemic process. We initially assume that no observation is available for inference and that the model parameters are known. Two different stochastic compartmental epidemics models are considered: Susceptible-Infective-Removed (SIR) and a not-typical Susceptible-Infective-Susceptible (SIS). We adopt a stochastic optimisation method for identifying the optimal control strategies among competing strategies using the Simulated Annealing algorithm. The methodology is applied on different simulated data using fixed parameters. We illustrate how such an algorithm could be applied to epidemic data. We show empirically the benefit of coupling epidemics when comparing competing control strategies.

Chapter 5 is an extension of Chapter 4 to the case where the control is based on some partially observed epidemic data. In addition, this chapter is concerned with models that take account of both time and space which are more applicable to arboreal

diseases. Inference is carried out in the Bayesian framework using the Markov Chain Monte Carlo (MCMC) coupled with data augmentation. Given that the intervention will always lie at some future time, we construct a number of alternative MCMC algorithms differing in the extent to which they impute explicit infection times beyond the observation period. Similar to the chapter 4, epidemics are coupled under different control strategies using the Sellke construction. At the level of the individual host, we construct different measures including the risk, the challenge, the threat defined as the expectation of the product of risk and challenge and other measures. These measures are then used for prioritising hosts for consideration under a putative control strategy. Different designs spaces are constructed and compared. The approach is applied to simulated data and on a population representative of citrus population in Florida. We compare the effectiveness of different prioritisation measures and quantify the benefit of using the Sellke construction in this context. We show that a control design based on the threat measure is more cost-effective compared to other measures.

Chapter 6 is concerned with modelling and designing control strategies using the coupling idea developed in Chapter 5 with an emphasis on real-life epidemic data regarding the spread of Citrus canker from Florida in 1995. A retrospective analysis on the performance of the control strategies is also performed. This chapter also includes an extension to a two-level mixing population where susceptibilities and infectivities of hosts vary.

In the final part of this thesis, we develop a mathematical model that can provide insight into likelihood of spread between countries of the disease Peste des Petits Ruminants (PPR) and risk of introduction across boundaries and that might be applicable more generally. In particular, we first describe how model parameters can be estimated in the Bayesian setting when there is uncertainty on both the transmission network between infected hosts and the susceptible population. Then, we establish a network model to address questions regarding the risk of spread in the large-scale network of countries and introduce the notion of "first-passage time" using techniques from graph theory and operational research such as the Bellman-Ford algorithm. This chapter also includes a first attempt to provide a real-time risk assessment tool for the spread of PPR at the continental scale. The methodology is also applied on simulated data. Moreover, we compare our findings with those presented in the literature.

1.3 Epidemic modelling

1.3.1 Motivation for modelling epidemics

To provide a best control during a disease outbreak, it is important to understand and be able to describe the underlying process. However, data from epidemic outbreaks

make the analysis difficult due to their incomplete nature (e.g. the exact times of infections or the number of infectious hosts are mostly unknown). Recent outbreaks of plant diseases such as Citrus canker (Gottwald and Timmer, 1995; Gottwald et al., 2002) and Cassava mosaic disease in Sub-Saharan Africa (Hillocks and Thresh, 2000; Alabi et al., 2011), human diseases like Ebola in west Africa and SARS (Severe Acute Respiratory Syndrome) (Lipsitch et al., 2003; Wallinga and Teunis, 2004), the world-wide emerging animal diseases like influenza H1N1 (Boender et al., 2007; Fraser et al., 2009) and FMD (Ferguson et al., 2001; Chis-Ster et al., 2009) and many more show how challenging it is to understand the mechanism underlying the spread of such diseases. However, relating models to epidemic data is crucial when it comes to understanding the processes that govern the dynamics of the epidemic. Models could be used to provide a better understanding of the epidemiological quantities of interest such as the infection process and the transmission dynamic, and to express scientific hypotheses (O'Neill, 2010). To facilitate a clear understanding of the disease dynamics at the population or individual level, it is vital to keep the model as simple as possible with a manageable number of parameters.

It is then apparent that not only are epidemic models essential for providing an adequate description of the mechanism that generate the data, they have also the ability to predict the future course of the outbreak conditioned on its current state. This contributes to determining the effect of different control strategies which lead to identification of effective control policies and therefore prevent major spread. In addition, models can help in designing optimal controls for future outbreaks by providing tools to target high-risk regions (Woolhouse, 2011). Models are therefore important in any data analysis and fundamental to answering important questions regarding the underlying process specially the question of epidemic control.

1.3.2 Compartment models

To represent the spread of disease through a finite population of size N, epidemiologists commonly use compartmental models. In one such model, individuals in the population are categorized by disease status. Individuals who are able to contract the disease are Susceptible (S) until they become infected. In this case, if they cannot yet transmit the disease, they are said to be Exposed (E). After a period of time, they may become infectious but not present any symptoms i.e. they are Cryptic (C). Then they become symptomatic and still Infectious (I). An infected host can recover from the disease or be immunized, or be isolated or removed through death or control. Such individuals play no further role in the epidemic spread. A range of commonly studied epidemic models can be represented in this framework including SEIR, SECIR, SCIR, SCI, SI and SIR depending on the biology of the disease under study. Although our focus is mainly on the SCI and SI, below we give a brief overview on how epidemics

can be modelled considering the SIR model. In addition, throughout this thesis, we assume that the population is closed i.e. the size N of the population is fixed at each time ($N = S(t) + E(t) + C(t) + I(t) + R(t)$), where $S(t)$, $E(t)$, $C(t)$, $I(t)$ and $R(t)$ are the numbers of individuals in the compartment S, E, C, I and R respectively at time $t \geq 0$.

There are two standard approaches to implementing compartment models of this form: stochastic and deterministic modelling.

1.3.3 Deterministic models

Early approaches to epidemic were focused on specific human diseases and adopted the deterministic approach. The first contribution by Bernoulli (1760) aimed to measure the effectiveness of a certain technique, namely variolation (a technique that usually confers immunity to patient), against smallpox. This analysis can be found in (Daley and Gani, 1999). Later, Ross (1911) used deterministic models to explain the transmission of Malaria. The first general mathematical model for studying epidemics was the deterministic model introduced in (Kermack and McKendrick, 1927). They assumed a closed population and used the homogeneous mixing principle for continuous time $t \geq 0$. That is, the size of population is fixed and individuals randomly mix together which, in terms of disease, means that any infected individual is equally likely to infect any currently susceptible individual.

Deterministic models, used since the 19^{th} century, considered a set of differential equations which describe the temporal evolution of the state of the system. This description ensures that the future state of the process can be precisely determined if the initial conditions are known. Considering the typical general epidemic (SIR) model with susceptible, infective and removal $S(t)$, $I(t)$ and $R(t)$ at time t respectively, the deterministic epidemic model (Bailey, 1975, Chap 6.2) is defined by the following set of differential equations:

$$\frac{dS}{dt} = -\beta S(t)I(t) \tag{1.3.1}$$

$$\frac{dI}{dt} = \beta S(t)I(t) - \gamma I(t) \tag{1.3.2}$$

$$\frac{dR}{dt} = \gamma I(t) \tag{1.3.3}$$

where β is the contact parameter and γ the removal rate. A common initial condition typically used is $(S(0), I(0), R(0)) = (s_0, i_0, 0)$ where i_0 is usually small. Note that in some literature, β is replaced by $\frac{\beta}{s_0}$ which is more appropriate when dealing with a varying initial susceptibility size. This parametrisation leads to the branching process

approximation for the initial stage of a stochastic epidemic in a large population (Andersson and Britton, 2000, Chapter 3).

It follows that

$$\frac{dS}{dR} = -\frac{\beta}{\gamma}S \tag{1.3.4}$$

therefore,

$$S(t) = S_0 \exp\left(-\frac{\beta}{\gamma}R(t)\right) \tag{1.3.5}$$

$$\text{so that } I(t) = N - R(t) - s_0 \exp\left(-\frac{\beta}{\gamma}R(t)\right) \tag{1.3.6}$$

where N is the size of the population. Furthermore, from 1.3.2, we have

$$\frac{dI}{dt} = \beta I(t)\left(S(t) - \frac{\gamma}{\beta}\right). \tag{1.3.7}$$

Kermack and McKendrick (1927) show that, if the epidemic is ever to take off, $s_0 > \frac{\gamma}{\beta}$. In other words the initial number of susceptibles must exceed the quantity $\frac{\gamma}{\beta}$.

The ratio $R_0 = \frac{s_0\beta}{\gamma}$ is indeed of fundamental importance in epidemic modeling and is defined as the expected number of secondary infections produced by a typical infected individual during its entire infectious period in a completely susceptible population (Heesterbeek and Dietz, 1996). This is referred to as the *basic reproduction number* originally used in demography to state the average number of individuals that one individual reproduces (Britton, 2010). By analogy to the threshold result of Kermack and McKendrick (1927) stated previously, when $R_0 > 1$ the epidemic takes off while there is a minor outbreak when $R_0 \leq 1$.

Note that this quantity is also useful when considering the stochastic version of the model.

1.3.4 Stochastic models

A first genuinely stochastic approach to epidemics models goes back to McKendrick (1926). But the one that received more interest was the unpublished work on the chain-binomial model of Reed and Frost presented in series of lectures in 1928. More details of this model can be found in (Wilson and Burke, 1942, 1943). Moreover, the model was extended by Greenwood (1946, 1949) and Bailey (1953).

Advances made in early $1940's$ in the application of mathematics to stochastic processes have led to some important developments in computational modelling. During that period, Barlett (1949) formulated the stochastic version of the model described by Kermack and McKendrick (1927) and thereafter followed a rapid increase of stochastic models for epidemics (Dietz and Schenzle, 1985), especially during the last 50 years mainly due to the increases in available computing power.

The stochastic version of the model considered in Section 1.3.3 in the deterministic case is defined as a continuous-time Markov process $\{(S(t), I(t)); t \geq 0\}$ and known as the standard (Markov) stochastic SIR model. Here, the transition between states occurs according to a random process. In particular, the transition probability from S to I i.e. an infection in a small time interval $(t, t + dt)$ is defined as

$$P[S(t + dt) = s - 1 | S(t) = s] = \beta S(t) I(t) dt + o(dt), \tag{1.3.8}$$

while the transition from I to R i.e a removal is governed by

$$P[I(t + dt) = i - 1 | I(t) = i] = \gamma I(t) dt + o(dt). \tag{1.3.9}$$

This description of the process implies that new infection occurs at a point of a non-homogeneous Poisson process (Grimmett and Stirzaker, 2001) with rate $\beta S(t) I(t)$. In addition, the infectious periods are defined to be independent exponentially distributed with mean $\frac{1}{\gamma}$. However, other prescribed distributions could be considered such as the Gamma or Weibull (O'Neill and Becker, 2001; Streftaris and Gibson, 2004). For now, we only focus on this brief definition of the stochastic model as the main body of the thesis is based on this approach. Additional details are given at the appropriate points.

The above definition of the basic reproduction number could be generalized as $R_0 = \beta s_0 E(I)$, where $E(I)$ is the expectation of the infectious period I. For stochastic model, the behaviour of R_0 for large populations (Whittle, 1955; Williams, 1971; Andersson and Britton, 2000) is analogous to that of the deterministic model. Indeed, from the threshold theorem (Whittle, 1955) in a population with a large number of susceptibles, $R_0 \leq 1$ implies that only a small number of susceptibles will become infected with probability one whereas if $R_0 > 1$ there is a positive probability of a major outbreak.

1.3.5 Deterministic versus stochastic

Deterministic models considered a set of differential equations which describe the future state of an epidemic process. This description ensures that the future state of the process can be precisely determined if the initial conditions are known. However, real-life epidemics are more complex. Indeed, in a large population, a real-life epidemic can go extinct or lead to a major outbreak infecting more or less a large number of the population. Also, in a small population with one individual who has a cold, some will catch the disease while others will not (Daley and Gani, 1999). In other words, chance plays a part in the spreading of the disease. Therefore, only stochastic models can describe effectively these processes underlying the transmission of an epidemic

and the consequent variability in the outcomes of epidemics that proceed under identical conditions. Moreover, one of the most important advantages of using stochastic models concerns estimation. In fact, there is little value in estimating an unknown quantity without considering its uncertainty. Stochastic models are appropriate in accommodating uncertainty in the epidemic outcomes therefore allowing a likelihood approach to estimate model parameters.

It is now widely known by researchers that, both deterministic and stochastic models have their strengths and are very important in modelling any given system (Renshaw, 1991; Isham, 2005). Although stochastic models are more natural and realistic models to be used when modelling epidemics their analysis can be difficult. In this case deterministic models may provide a useful alternative. In addition, deterministic models could be used as a starting point when dealing with a new pathogen (Andersson and Britton, 2000). In this thesis, however we focus on stochastic models in order to incorporate intrinsic stochasticity in many ways.

1.4 Background on epidemic control

The most practical reason for modelling an epidemic lies in the need to design an optimal control in order to reduce or to eradicate the disease impact. In this section, we review some approaches designed in that regard.

This goal was already evident in Bernouilli's pioneering work on smallpox (Bernoulli, 1760). He used epidemic models to demonstrate that inoculation with material taken directly from patients with a mild case of smallpox to immunize an individual, in the hope that it would reduce the death rate, ultimately increased the population of France. The same idea is shared by VanderPlank (1960, 1963) who introduced the concept of monocyclic (primary inoculum) and polycyclic (primary and secondary inoculum) pathogens and developed two simple models of plant diseases to model these two types of pathogens. This concept leads to a quantification of various parameters included in the models, to predict how disease will occur and identify controls to reduce the impact of the pathogen. As a result, he elaborated the concept of horizontal and vertical resistance (resistances conferred by genes) to show that coupling appropriately different varieties of plants could reduce both the initial and the secondary inoculum.

These approaches provide motivation for disease modelling. In this section, we present various approaches used to prevent or to control epidemics.

1.4.1 Control by educational campaign

When there is no recognised cure or effective vaccine, the education campaign becomes the most reliable control program. During the course of an epidemic, a change in behaviour of the host population could have an impact on the transmission dynamics. These behaviours include susceptible hosts avoiding contact with infected individuals or avoiding public gatherings, in the case of human diseases for example, or by simply following some prescribed rules or guidance. These practices have been successful at times including the 1918 influenza epidemic (Bootsma and Ferguson, 2007), the "STOP-AIDS" campaign lunched by the Swiss AIDS foundation in 1987 (Daley and Gani, 1999, chapter 7), and the Ebola prevention campaign by differents health organisation (Centre for Disease Control-CDC, World Health Organisation-WHO and Médecin Sans Frontières-MSF).

Such campaigns always come with a cost in addition to the benefits. The cost can take the form of resources (money) or it can be the number of infections at the end of an epidemic. Therefore, controls that are optimal with regard to this cost are then sought. There is a comprehensive literature devoted to this topic. (Daley and Gani, 1999, chapter 7) showed that the education campaign is 4% of the cost of an HIV treatment by representing the number of new infectives as Poisson process and assuming that condoms provide total protection from HIV infection. On the other hand, Castilho (2006) designed optimal controls through education campaigns for an epidemic based on the simple SIR model by reducing the contact rate and increasing the removal rate simultaneously.

In the case of high within-herd transmission rate, education can advise avoiding contacts between flocks through common grazing for example. The strategy is effective in animal and plant epidemics where owners or actors are sensitized to take some actions either by reporting some suspect cases or by disinfecting equipment and storage facilities (Gottwald and Timmer, 1995).

While educating the population has proved to be a valuable strategy to limit the impact of an epidemic, it is necessary to couple it with other measures.

1.4.2 Control by immunization

This strategy is commonly used with human and animal diseases. It is certainly the most widespread control policy given that it creates an artificial immunity amongst the current susceptible population. A vaccination program against smallpox is known to be the most efficient control for eradicating the disease entirely (Fenner et al., 1988). However, the complete eradication is not always achieved for certain epidemics (measles, malaria, polio, avian influenza for instance) due to the fact that it is not always possible to vaccinate every individual for practical and ethical reasons. Various

control schemes for human and animal epidemics can be found in (Knox, 1980; Dietz, 1981) and (Anderson, 1982).

In controlling the spread of a disease by immunisation or other means, the question of how to distribute the resources (which and how many individuals to vaccinate) to minimise the risk of a large outbreak, remains of critical importance. These questions have been answered in various published papers. In the remainder of this section, we discuss different models considered for the purpose of answering these questions.

In a homogeneous population, the previous question becomes that of determining how many susceptibles need to be immunised so that an optimal reduction in the final epidemic size is obtained. In (Smith, 1964), it was shown that, for the general epidemic model, $v_c = 1 - 1/R_0$ is the level of immunisation required to avoid a major outbreak. The quantity v_c is known as the *critical immunisation level*, where R_0 is the basic reproduction number.

In general, Anderson (1982) shows that if θ is the threshold parameter, then the proportion vaccinated v satisfies $v \geq (1 - 1/\theta)r$, where r is relative reduction in the rate at which vaccinated individuals become infected, compared with the rate for unvaccinated individuals. This means that $v_c = \frac{1}{r}\left(1 - \frac{1}{\theta}\right)$. In particular, if the vaccine is perfect (i.e. giving 100% immunity ($r = 1$)) v_c is equivalent to the version in (Smith, 1964). When r is small and θ is large, $v_c > 1$; this implies that the population is not safe from a major outbreak.

However, assuming homogeneously mixing populations prior to determining the vaccination strategy is unrealistic since most strategies in practice are implemented on a heterogeneous community. In this case, the two questions stated previously are relevant, since the effect of vaccination depends not only on the sample to vaccinate but also on which groups of host are prioritised. In trying to answer these questions Anderson and May (1991, Chap. 12) describe an optimal immunisation strategy within a spatially heterogeneous population and Becker and Dietz (1995, 1996) have considered subgroups consisting of population with small size (Households, Schools). In their analysis, they considered four different strategies:

1. Households chosen at random and all their members are immunised
2. Random vaccination of individuals
3. Preferential selection of large households for immunisation
4. Strategies for households of equal size (imunisation of a fixed fraction of members in every household).

Anderson and May (1991) showed that in a two-level mixing population (where hosts mix at a higher level within a group than with hosts in other groups), the optimal vaccination strategy is such that the proportion of non-immunised susceptible members

in each group are identical. Becker and Dietz (1995), through considering a severe disease spreading in a population consisting of households (the whole household is infected once an individual is infected), showed that the optimal strategies in that case is the one that immunises the same fraction of individuals in each household.

Ball et al. (1997a) then generalised this result to a more realistic within-group distribution where a group with an infected host may still contain some susceptibles. They named it an "equalising" strategy which consists of leaving the number of susceptibles in each group as similar as possible. This is done by picking sequentially individuals from groups with the largest number of remaining susceptibles.

It is important to note that in a heterogeneously mixing population where infectivity varies, there is no simple solution for controls. The natural solution is to find a trade-off between high susceptibility and high infectivity when dealing with an optimal strategy.

1.4.3 Control by screening and quarantine

Another approach in controlling the spread of disease is by screening the at-risk hosts and quarantining those likely to propagate the epidemic. This approach gives a means of stopping the disease from entering a region believed to be disease free. This procedure has been adopted to prevent the introduction of malaria in countries free of malaria (Daley and Gani, 1999), and severe acute respiratory syndrome (SARS) adopted by the WHO (2003). It was recently applied to the Ebola outbreak where individuals flying from high-risk countries including Liberia, Guinea and Sierra Leone were screened and quarantined if there were any suspicion of Ebola symptoms by the CDC (2014, 2015).

However, Selvey et al. (2015) showed that this measure is not always cost-effective and used the 2003 SARS pandemic and the 2009 influenza pandemic as exemplar case studies to demonstrate this.

For diseases of plants and animals however, this strategy is shown to slow the spread from the source of infection for a given period of time but may potentially fail to stop the introduction of a pathogen into a previously disease-free region. This is due to the difficulties of controlling the importation of plant or animal via illegal trade as pointed out by EFSA (2015).

1.4.4 Control based on host diversity

The dynamic of infectious diseases can be impacted by host diversity within a population. Even over a century ago, there was a suggestion that species diversity and transmission of vector-borne diseases of humans were correlated (Service, 1991). This idea has long been appreciated by Macdonald (1957) who singled out the role of cat-

tle in diverting mosquitoes from transmitting malaria to humans. This concept is later referred to as Zooprophylaxis (WHO, 1982) and has been successfully used in environmental management of vector-borne diseases (Ault, 1994). Elton (1958) has extended this idea to plant diseases. He estimated that reducing the density of the host plant for a disease by introducing new resistant plant species could influence the behaviour of the pathogen vectors. This was supported by Burdon and Chilvers (1982) and Boudreau and Mundt (1982) through their empirical research on plant diseases and by Anderson and May (1981) who demonstrated the sensitivity of disease transmission to host density using epidemiological models.

In plant populations, removing hosts has been the common strategy in order to create a barrier between healthy and infected hosts. However, some plant varieties are self-evidently resistant to a pathogen, and therefore can be used to slow the spread of disease in a plant population. Previous work in crop diversification strategies for epidemic control (Finckh and Wolfe, 1997, 2006) showed that one way to slow the spread of disease in an agricultural crop is the construction and deployment of crop multilines and cultivar mixtures (Mundt, 2002).

1.4.5 Pre-emptive culling in livestock and plants disease

Preemptive culling refers to a procedure of culling or removing high-risk exposed hosts either by direct contact or at a certain geographical distance from a potential source (Keeling et al., 2001; Tildesley et al., 2009; Ferguson et al., 2001; Klinkenberg et al., 2003; Donnelly et al., 2006). This strategy has been widely used during the last decade in Europe, USA, Asia and elsewhere, to control highly infectious diseases in animal populations such as avian influenza, bovine tuberculosis, classical swine fever, FMD (Keeling et al., 2001; Morris et al., 2001), and in plant populations such as citrus canker, dutch elm disease, sudden oak death (Gottwald et al., 2001a, 2002).

Most often in practice, culling is carried out in a ring (ring culling) around a detected infection. However, implementing this form of control is invariably challenging given that it requires an accurate estimate of how far the epidemic is likely to spread ahead of symptomatic hosts. Nonetheless, there has been a comprehensive literature showing the effectiveness of such a control approach.

Using the approach followed by the joint USDA, APHIS/Florida Department of Agriculture and Consumer Services Citrus Canker Eradication Program (CCEP) in Florida between 1995 and 2005 (Gottwald et al., 2001a) whereby culling of all hosts around a detected infective was carried out, Parnell et al. (2009); Cunniffe et al. (2015) showed how this type of control can be optimized by matching the spatial scale of control with the spatial scale of epidemic spread. Dybiec et al. (2004); Stacey et al. (2004); Dybiec and Gilligan (2005); Gilligan et al. (2007); Dybiec et al. (2009) extended this idea to the process where spread of the disease occurs non-locally and

through a more complex network.

As stated above, once a ring is defined, all hosts in that ring are culled. However, culling all hosts simultaneously may not be feasible given the limited resources that are available during any outbreak. To this end, other strategies may focus on sampling hosts according to some measures or metrics. Theses measures are often used to construct a geographical map from which hosts may be prioritized for culling.

Boender et al. (2007) defined the basic reproduction number (R) as the number of new infections that each infected farm is expected to generate, which allow mapping out transmission geographically for the spread of the 2003 Netherlands Avian influenza (H7N7). A similar idea is used by Tildesley et al. (2009) where they developed spatial maps to quantify the effect of contiguous premises culling on outbreaks from each infectious source. In both, it was shown that culling based on this measure outperformed the most used random sampling strategy. Later, te Beest et al. (2011) introduced a risk-based culling to account for both the distance of susceptible farms to the infectious and the number of secondary cases that a susceptible was expected to cause if it became the focus of a new local epidemic. They defined an approximation of the expected number of infections caused by farm i as $E_i^* = R_i q_i^*$ where q_i^* is the approximate probability that farm i was infected. Although their model did not account for cryptic hosts, they showed that culling hosts with the highest expected number of infections is more cost-effective compared to the traditional random culling. However, approaches of this kind carry various limitations. First, cryptic hosts might not have been taken into account; second the methods always assume a fixed infectious period which is unrealistic in most real-life epidemics and finally this type of approximation is feasible only for epidemics with a relatively short infectious period. For an epidemic such as citrus canker where hosts can be infectious for a long period of time, this approximation of the probability of being infected will be close to unity, even for hosts located at a long range from the foci. Similar approaches have been used (DEFRA, 2013) to predict spread of the ash die-back by 2017 in the UK by jointly considering the likelihood of infection at any given location within the UK and the hazard value for a particular location. Handel et al. (2011) considered a similar approach to tackle the problem of post-epidemic surveillance; they combined the transmission model with an estimate of farm-specificity of being detected infectious. Again, their risk-targeted model is shown to be more efficient than random sampling.

In the next chapters of this thesis, we will extend this idea of prioritisation described previously. In particular, we will investigate how prioritisation could be done using a range of measures systematically in the Bayesian framework. In addition, we will look at how optimal controls might be identified with the minimum of computation using coupled epidemics.

Chapter 2

Statistical models: Structure and implementation

2.1 Structure of models

In epidemiology, the risk of a host contracting a disease often depends on three factors: location, time, and characteristics of the host, such as genetic factors. However, all this information is not always available when collecting epidemic data. Therefore the choice of models for making inference on the epidemic depends on the factors contained in the data or, more precisely, on what is essential to understand the relationship between exposure and contraction of the disease. Two common model structures are used in that regard, namely non-spatial (time-dependent) and spatio-temporal models.

2.1.1 Non spatial models

The term non-spatial model in epidemiology refers to models that do not explicitly represent the spatial pattern of the spread of the disease on the spatial distribution of the host population. These models are commonly used to analyse the impact of diseases in a community and at a particular location. The smallpox epidemic in Abakaliki (Nigeria) (Bailey, 1975), the October/November 1967 epidemic of respiratory disease on the island of Tristan da Cunha (Shibli et al., 1971) are some illustrations. The probability that a host is infected in an infinitesimal time interval is usually represented as

$$P\left(i \text{ infected in } [t, t+dt]\right) = \beta I(t)dt + o(dt). \tag{2.1.1}$$

where β is the contact rate and $I(t)$ the epidemic size at time t.

2.1.2 Spatial models

Epidemiological studies are very often spatial and temporal (Shaddick, 2013). One of the principal reasons for using a spatio-temporal epidemic model is to "*reduce the initial variability of disease risk by using and exchanging information between neighbouring spatial locations and proximal temporal moments*" (Iftimia et al., 2015). These spatial locations can be small areas or any geographical locations, networks of sampling positions or economic and social structures. However, the consideration of models involving time and space has been hampered by the difficulty in collecting, managing, processing and analysing data regarding space and time. In recent years, advances in technology, computing power and techniques such as Geographical Informations System (GIS) have made possible the collection and storage of epidemic data over time and geographical scales.

Analytic tools such as spatio-temporal distance class analysis (Nelson, 1995; Gottwald et al., 1996), spatio-temporal autoregressive integrated moving average (STARIMA) methods (Gottwald et al., 1992; Reynolds and Madden, 1988a,b) and ordinal regression (Besag, 1997; McCullach, 1980) have been used in the past. However, spatio-temporal stochastic models have been successfully used in recent years to represent the transmission from infection sources (individuals in the population or sources outside the population). In this framework, the inclusion of space and time into an epidemic model occurs through the introduction of a dispersal kernel characterising the challenge posed by an infected host to a susceptible. Considering a stochastic SIR model for instance, the probability of contracting the disease in an infinitesimal time interval can be modelled as:

$$P\left(i \text{ infected in } [t, t+dt]\right) = \lambda_i(t)dt + o(dt). \tag{2.1.2}$$

where λ_i is the hazard rate or the force of infection on host i at time t which depends on the transmission kernel characterising the spatial nature of the dispersal of the pathogen.

2.1.2.1 Dispersal kernel in two dimensions

The kernel characterises the spatial nature of the dispersal of the pathogen from infected hosts to susceptible hosts. It plays an important role in modelling any control strategy; especially when the control involves culling hosts in an appropriate range around a detected infection. Here, we define a kernel in two dimensions.

Assume that the inoculum emitted from a given source infects a susceptible individual located at a random distance from the source, drawn from a density $r \sim f(r|\theta, \alpha)$ with parameter α and placed at a uniformly distributed bearing θ (Figure 2.1).

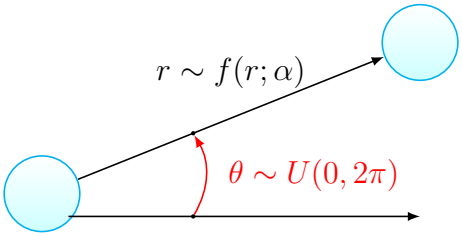


Figure 2.1: Schematic representation of the epidemic transmission.

Following Clark et al. (1999), the probability that an inoculum from a source falls in a small area at distance r and angle θ, delimited in polar coordinates by dr and subtending an arc of angle $d\theta$ is given by

$$P(dS) = \int_{r}^{r+dr} \int_{\theta}^{\theta+d\theta} f(r', \psi|\alpha) dr' d\psi \tag{2.1.3}$$

where $f(r, \psi)$ is the dispersal kernel and ψ is the direction.

$$P(dS) = \int_{r}^{r+dr} \int_{\theta}^{\theta+d\theta} f(r'|\psi, \alpha) g(\psi) dr' d\psi \tag{2.1.4}$$

where $g(\theta)$ is the density of θ assumed to be uniform over $(0, 2\pi)$. Assuming an isotropic kernel density,

$$P(dS) = \int_{r}^{r+dr} f(r'; \alpha) g(\theta) dr' d\theta$$

$$\propto \frac{f(r;\alpha)}{r} dS$$

since $dS = r dr d\theta$. Defining the two dimensional kernel $K(r;\alpha)$ as $P(dS) = K(r;\alpha)dS$, we can derive the kernel as

$$K(r;\alpha) \propto \frac{f(r;\alpha)}{r} \tag{2.1.5}$$

In this case, the parameter α is known as the dispersal parameter.

2.1.2.2 Representation of the hazard rate

The hazard rate $\lambda_i(t)$ defined in Equation (2.1.2) depends on the underlying assumption made. For example, when the transmission is assumed to occur from only individuals in the population i.e. by the process of a secondary infection as in the models considered in (Gibson, 1996; Gibson and Austin, 1996; Chis-Ster et al., 2009; Parnell et al., 2009; Demon et al., 2011; Cunniffe et al., 2014), the rate of infection is given

by

$$\lambda_i(t) = \left(\beta \sum_{j \in I(t)} K(d_{ji}, \alpha) \right) \tag{2.1.6}$$

whereas in a more complex setting where the disease transmission occurs from both individuals from the population (secondary infection) and sources outside the population (primary infection) (Gibson, 1997; Parry et al., 2014; Cook et al., 2008; Cunniffe et al., 2015), the rate of transmission is

$$\lambda_i(t) = \left(\beta \sum_{j \in I(t)} K(d_{ji}, \alpha) + \epsilon \right) \tag{2.1.7}$$

where ϵ is the rate at which the inoculum is transmitted from the external source to susceptibles.

2.2 Implementation of stochastic models

Let us denote by E the state space of the system. By trajectory of the epidemic $X : [0, \infty) \to E$ we mean the number of individuals in each compartment at a given time $t \in [0, \infty)$. Note that here, $E = \mathbb{N}^n$ where n is the number of compartments.

2.2.1 Gillespie algorithm

The Gillespie algorithm developed by Gillespie (1976, 1977) is an exact stochastic simulation approach initially designed for chemical kinetics to address the intractability of analytical solutions to stochastic time-evolution equations. In essence, the algorithm generates the time-evolution of trajectories of finite populations in continuous time. The algorithm has become the standard simulation approach for any continuous-time Markovian system including epidemic models. The implementation of the algorithm for an epidemic model is described below:

Consider n moves with respective rates $\beta_1, \beta_2, \ldots, \beta_n$ and let $X(t)$ denote the trajectory.

Algorithm 1

1. Initialise $t = t_0$ and $X = X_0$.
2. Evaluate the probabilities $P_i = \frac{\beta_i}{\sum_{j=1}^{n} \beta_j}$ $i = 1, \ldots, n$.
3. Draw u_1 and u_2 from $U(0, 1)$

4. The time Δt to the next event is given by

$$\Delta t = \frac{1}{\sum_{j=1}^{n} \beta_j} \log\left(\frac{1}{u_1}\right) \tag{2.2.1}$$

5. The next event is the jth event where

$$j = \inf\{k, \sum_{i=1}^{k} P_i > u_2\} \tag{2.2.2}$$

6. The system next state is $t = t + \Delta t$ and $X \to X'$, where X' corresponds to the updated state according to the chosen move.

7. Repeat $2 - 6$ till the stopping criterion is reached, usually when $t \geq T$.

It is important to note that the rates β could be a function of time. The parametrisation described here implies that the same set of parameters will lead to different outcomes; in other words the epidemic outcome is a stochastic function of the parameters.

2.2.2 Sellke construction

The Sellke construction is one way to specify stochastic dynamics for standard stochastic epidemic models, which preserves the sampling properties of the Gillespie algorithm. Sellke (1983) showed that each susceptible individual has a degree of tolerance $Q_i \sim Exp(1)$ to the disease. A susceptible individual i becomes infected at time t when the total infection pressure exerted on it up to t defined by $A_i(t) = \int_0^t \lambda_i(s)ds$ reaches its threshold, i.e. when $Q_i = A_i(t)$.

In practice, to generate a realisation of an epidemic process through the Sellke construction, we start by drawing random thresholds for each host. We then compute the time to the next infection by computing a potential infection time for each host. More specifically, if the latest infection occurs at t_k and Δt_i is the subsequent period it takes for the next individual i to become infected, the infection pressure on each individual in period $[t_k, t_k + \Delta t_i]$ is given as $A_i(t) = \lambda_i(t_k)\Delta t_i$. Then we change the disease status of relevant hosts and proceed until no event occurs or until a stopping criterion is reached. By doing so, the outcome of the epidemic is a deterministic function of the thresholds and the parameter. That is $\underline{x} = h(\boldsymbol{\theta}, \boldsymbol{Q})$ at least for an SI model where h is a one-to-one function. The application of the Sellke approach to an SIR epidemic model is now described as follows:

Assume that initially, there are $S(0)$ susceptibles, $I(0)$ infective and $R(0)$ removed individuals.

Calculation of $\underline{x} = h(\theta, \boldsymbol{Q}, I)$ for SIR model with only secondary infection.

Notation

- $I = (I_1, ..., I_N)$ the infectious periods for the N individuals in the population (all initially susceptible).
- $\boldsymbol{Q} = (Q_1, ..., Q_{N-n_I})$, the Sellke thresholds for the individuals (all initially susceptible).
- β, the secondary infection rate.

Calculation

- Set $t = 0$, number of infections $n_I = 1$, infection rate $\lambda(t) = \beta n_I$. Set initial state of each individual, $s_i = 0, i = 1, ..., N - n_I$, $s_i = 1, i = N - n_I + 1, ..., N$ and time till next transition $r_i = Q_i/\lambda(t)$. Set final time T.
$$s_i = \begin{cases} 0 & \text{if } i = 1, ..., N - n_I \\ \\ 1 & \text{if } i = N - n_I + 1, ..., N \end{cases} \quad \text{and } r_i = \begin{cases} Q_i/\lambda(t) & \text{if } s_i = 0 \\ \\ I_i & \text{if } s_i = 1 \end{cases}$$
- While $t < T$ do the following {

  Identify next event: Let $t^* = \min\{r_1, ..., r_N\}$ and $j = \arg\min\{r_1, ..., r_N\}$.

  Implement event.

  - case $s_j = 0$ (infection event), set $s_j = 1$, $n_I = n_I + 1$, $r_j = I_j$.
  - case $s_j = 1$ (removal event), $n_I = n_I - 1$, delete row j from tableau and move lower rows up, set $n_I = n_I - 1$.

  Finally update remaining Sellke thresholds and waiting times in the tableau.

  - $R' = \lambda(t)$, $t = t + t^*$, $\lambda(t) = \beta n_I$.
  - If $s_i = 0$, $Q_i = Q_i - R't^*$, $r_i = Q_i/\lambda(t)$.
    If $s_i = 1$, $r_i = r_i - t^*$ }

It is important to notice that the choice of $Exp(1)$ distribution for the threshold is motivated by the Equation (2.2.5). Indeed dealing with each individual separately, the transition probability of being infected in an infinitesimal time period dt is given

as:

$$
\begin{aligned}
P[i \text{ infected in (t+dt)} | i \text{ is susceptible at time t}] &= P[A(t+dt) \geq Q_i | Q_i > A(t)] \\
&= (1 - P[Q_i \geq A(t+dt) | Q_i > A(t)]) \\
&= \left(1 - e^{-(A(t+dt)-A(t))}\right) \\
&= \left(1 - e^{-A'(t)dt}\right) \\
&= (A'(t)dt + o(dt)) \\
&= \beta I(t)dt + o(dt)
\end{aligned}
$$

Hence we obtain the equivalence between the Sellke thresholds and the epidemic model. It is important to notice that the third line of the above equation is possible due to the memoryless property of the exponential distribution, while the fourth line is obtained merely by using a Taylor expansion. However, Ball and O'Neill (1999) discussed an extension to a Gamma(s, β) distribution when a given infective contacts a given susceptible according to Poisson process with rate β but a susceptible becomes infected after making a total of s contacts with infectious individuals. Ball and Britton (2005) assumed different exponentially distributed thresholds with rate not necessary equal to one and later Streftaris and Gibson (2012) showed that the choice of the thresholds could be extended to a more general distribution using the Weibull distribution.

It is worth noting that the Sellke construction is a particular example of a functional-model representation in the sense of Dawid and Stone (1982), or a non-centred parametrisation (Papaspiliopoulos and Roberts, 2013). The benefit of using such a construction lies in its application to design control strategies where epidemics can be coupled by matching latent processes, thereby inducing a strong correlation between the outcome of competing controls. We will discuss this throughout the thesis.

2.2.3 Simulation of an epidemic process using Sellke thresholds

2.2.3.1 Standard stochastic SIR model

The standard stochastic SIR model is defined as a continuous-time Markov chain. That is, given the whole history of the epidemic up to time t the probability distribution of the process at time $t' > t$ depends only on the state of the process at time

t. The transition probability from S to I i.e., an infection in a small time interval $(t, t+dt)$ is defined as follows:

$$P[S(t+dt) = s-1 | S(t) = s] = \beta S(t) I(t) dt + o(dt), \tag{2.2.3}$$

while the transition from I to R i.e. a removal is described as:

$$P[I(t+dt) = i-1 | I(t) = i] = \gamma I(t) dt + o(dt), \tag{2.2.4}$$

where β and γ are the contact and the removal rates respectively.

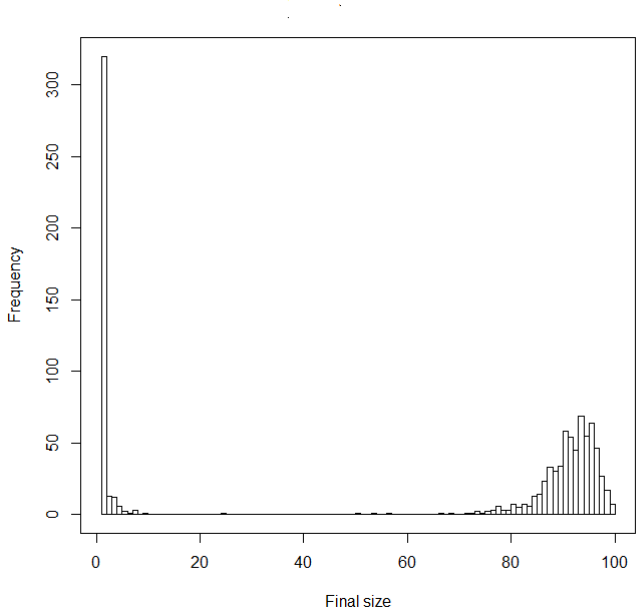


Figure 2.2: Final epidemic size after 1000 realisations (Sellke thresholds) of the standard stochastic epidemic model with $S(0) = 99$, $I(0) = 1$, $R(0) = 0$, $\beta = 0.003$, $\gamma = 0.1$ and $T_{max} = 60$.

We produce an histogram showing the final size of an epidemic in a population of size $N = 100$ where $S(0) = 99$, $I(0) = 1$, $R(0) = 0$, $\beta = 0.003$, $\gamma = 0.1$ after 1000 simulations (1000 Sellke thresholds for each host), observed up to a time $T = 60$ using the algorithm described in Section 2.2.2. The result is shown by the Figure 2.2. It can be seen that the epidemic is either minor or major, with no epidemics of final size between 10 and 70.

This description implies that individuals become infected (new infection occurs) as a Poisson process with rate $\beta S(t) I(t)$ and the infectious period follows an exponential distribution with rate $\gamma I(t)$. The latter has a lack-of-memory property, the reason why this model is called Markovian. Such an assumption is unrealistic, therefore other distributions for the infectious period might be investigated.

2.2.3.2 General stochastic SIR model

The general stochastic SIR model can be defined as a non-Markovian SIR model. In a non-Markovian SIR model, the infectious period is no longer exponentially distributed, but may follow an alternative distribution.

Non-Markovian SIR model with Weibull infectious period

Following Streftaris and Gibson (2004) we consider the more general case where the infectious periods of individuals follow a different prescribed distribution, In particular we consider the Weibull distribution since it has the flexibility to mimic a wide range of possible shapes for the distribution of infectious period.

The Weibull probability density function of a random variable X is given by

$$f(x,\nu,\lambda) = \begin{cases} \frac{\nu}{\lambda}\left(\frac{x}{\lambda}\right)^{\nu-1} e^{-(x/\lambda)^{\nu}} & \text{if } x \geq 0 \\ \\ 0 & \text{if } x < 0 \end{cases},$$

where $\nu > 0$ is the shape parameter and $\lambda > 0$ the scale parameter. This distribution has mean $\lambda\Gamma(1+\frac{1}{\nu})$ and variance $\lambda^2(\Gamma(1+\frac{2}{\nu}) - (\Gamma(1+\frac{1}{\nu}))^2)$; where Γ is the gamma function. Moreover, a reparameterisation of the Weibull distribution, replacing $\frac{1}{\lambda^{\nu}}$ by λ in the above equation, is given by

$$f(x,\nu,\lambda) = \begin{cases} \nu\lambda x^{\nu-1} e^{-\lambda x^{\nu}} & \text{if } x \geq 0 \\ \\ 0 & \text{if } x < 0 \end{cases}.$$

Thus the cumulative distribution function is given by

$$F_W(x,\nu,\lambda) = \begin{cases} 1 - e^{-\lambda x^{\nu}} & \text{if } x \geq 0 \\ \\ 0 & \text{if } x < 0 \end{cases},$$

It is worth noting that in the former case, the moments are given by $E(X^t) = \lambda^{-\frac{t}{\nu}}\Gamma\left(1+\frac{t}{\nu}\right)$ and the Weibull distribution is equivalent to $[Exp(\lambda)]^{\frac{1}{\lambda}}$. Note that this reparameterisation is particularly useful in Bayesian inference. With such a distribution, the infectious period loses the lack-of-memory property for $\nu \neq 1$ ($\nu = 1$ corresponds to the exponential case).

Therefore the probability that an infection occurs within a vanishing time window is given by

$$P[S(t+dt) = s-1 | S(t) = s] = \beta S(t) I(t) dt + o(dt), \tag{2.2.5}$$

and the transition from I to R of an individual is

$$\begin{aligned}
P[i \text{ is removed in (t+dt)} | i \text{ was infected at time t}] &= P[r_i \leq t + dt \mid r_i > t] \\
&= \frac{P[t < r_i \leq t + dt]}{P[r_i > t]} \\
&= \frac{P[t - s_i < r_i - s_i \leq t - s_i + dt]}{P[r_i - s_i > t - s_i]} \\
&= \frac{F_W(t - s_i + dt) - F_W(t - s_i)}{1 - F_W(t - s_i)} \\
&= \frac{f(t - s_i)dt}{1 - F_W(t - s_i)} + o(dt) \\
&= \nu\lambda(t - s_i)^{\nu-1}dt + o(dt); \quad (2.2.6)
\end{aligned}$$

where s_i and r_i are respectively the infection time and the removal time of the individual i.

2.2.3.3 Simulation study

We perform simulation of the non-Markovian SIR model using the Sellke construction. The benefit of using such a model is that it allows for either increased or decreased variability in infectious period, and also allows the infectious period distribution to peak around a selected value which may reflects the reality for many real-world epidemics such as FMD. To perform our simulation, we consider again a population of size $N = 100$ wherein $S(0) = 99$, $I(0) = 1$ and $R(0) = 0$; along with Weibull distribution, with scale parameter $\lambda = 9.94$ and shape $\nu = 1.1$, for the infectious period. We observe the epidemic process up to time $T_{max} = 60$. We simulate 1000 realisations of the process and at each simulation we observe the final size of the epidemic.

As we can see on Figure 2.3, the histogram of the epidemic final size after simulation shows two kind of epidemics: either it is minor or major. It can be seen that, the epidemic is behaving as in the Markovian case. This is due to the fact that the shape parameter of the Weibull distribution is close to unity and therefore the distribution is close to the exponential distribution.

Furthermore, to illustrate the impact of the contact rate in the epidemic process, we fix the infectious period parameters λ and ν and alter the value of the contact rate β. Figure 2.4 illustrates the influence of the contact rate on a given epidemic evolution. We notice that for small values of β, the main part of the population does not contract the epidemic while for large rate almost all the individuals becomes infected. One can easily predict this result since an infectious disease spreads rapidly

within a population where individuals are in frequent contact.

In addition, using the same thresholds simulated previously we fix the contact parameter β. By scaling the thresholds, we observe that the epidemic spreads quickly in a population where individuals thresholds are small whilst a population wherein individuals have large thresholds resists the epidemic as shown in Figure 2.5. This illustrates the fact that with fixed set of thresholds, different comparisons can be carried out.

We perform another simulation by fixing the mean of the infectious period at 9.59 ($\nu = 1.1$ and $\lambda = 9.94$) and altering the variance. The histograms on Figure 2.6 show that the epidemic size is increasing with the variance of the Weibull distribution. This result suggests that a larger epidemic is more likely when the variance is small since possibility of rapid removal of the initial infection becomes much smaller.

It is worth noting that on the Figure 2.6b, the spread of the epidemic is similar to the Markovian case (see Figure 2.2). This result is not surprising since for a shape parameter equal to 1, the Weibull distribution is merely the exponential distribution.

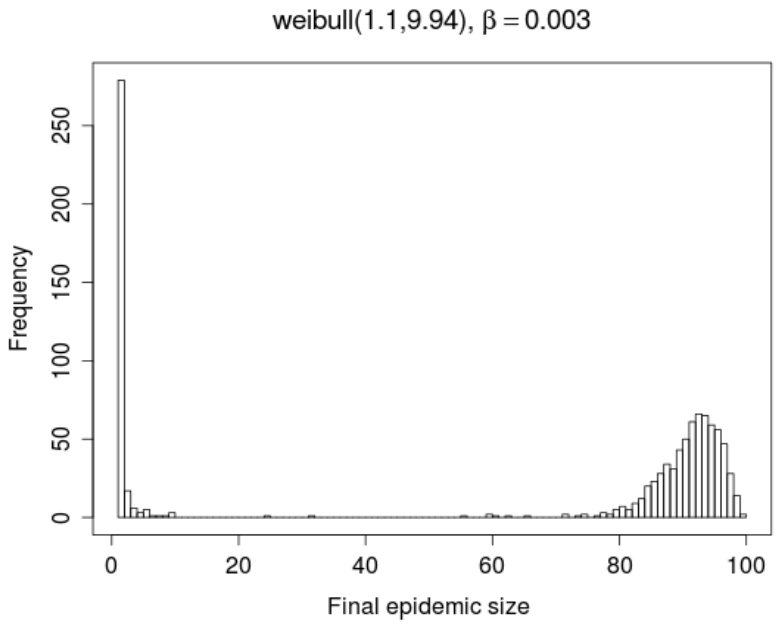


Figure 2.3: Final epidemic size after 1000 realisations of the non-Markovian process (Weibull infectious period) where $S(0) = 99$, $I(0) = 1$, $R(0) = 0$, $\beta = 0.003$, $\nu = 1.1$, $\lambda = 9.94$ and $T_{max} = 60$.

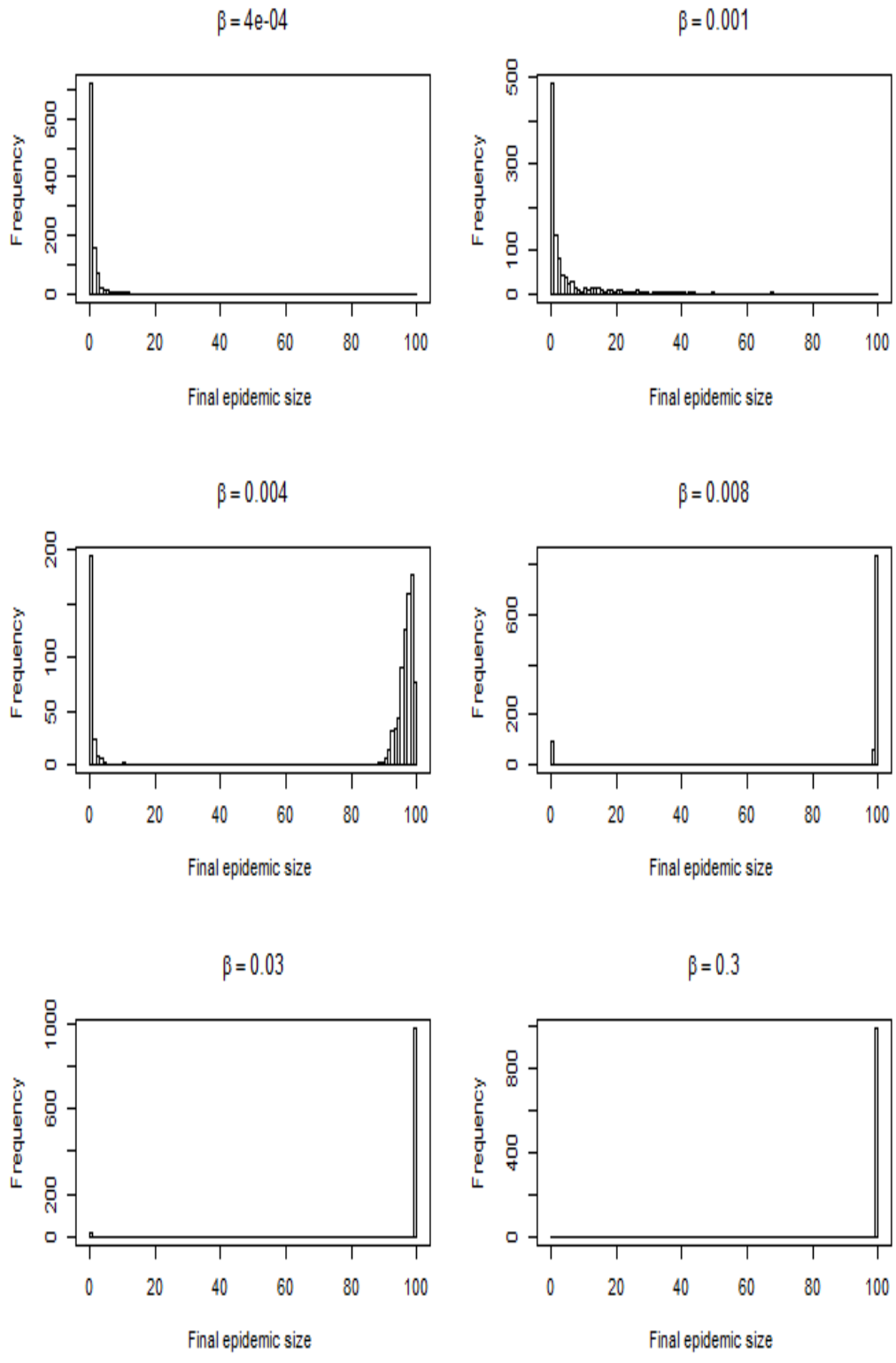


Figure 2.4: Final epidemic size after 1000 realisations of the non-Markovian (Weibull infectious period) with $S(0) = 99$, $I(0) = 1$, $R(0) = 0$, $\nu = 9.94$, $\lambda = 1.1$, $T_{max} = 60$ and different values of β.

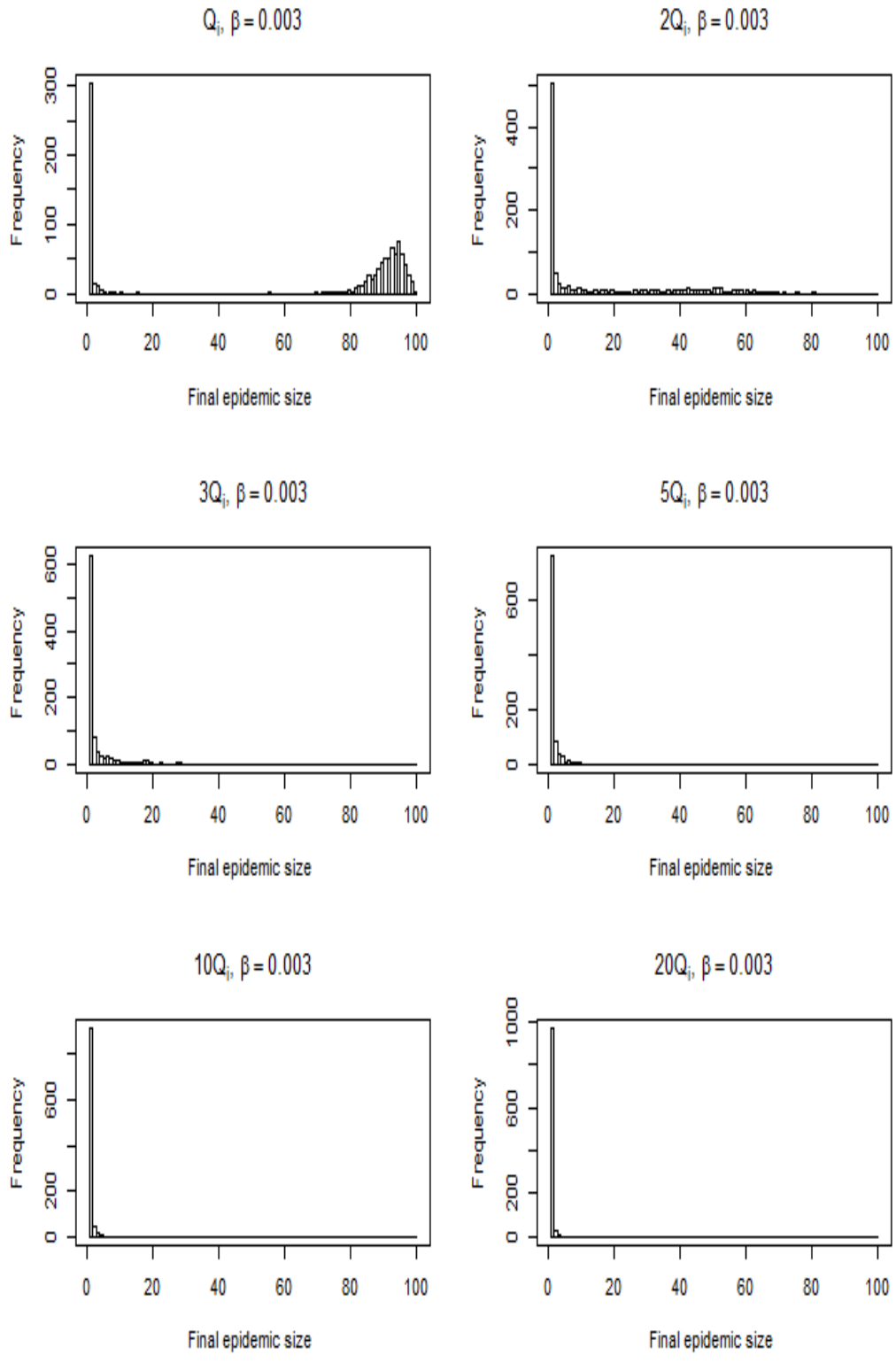


Figure 2.5: Final epidemic size after 1000 realisations of the non-Markovian (Weibull infectious period) with $S(0) = 99$, $I(0) = 1$, $R(0) = 0$, $\lambda = 9.94$, $\nu = 1.1$ and $T_{max} = 60$, fixed $\beta = 0.003$ and different thresholds.

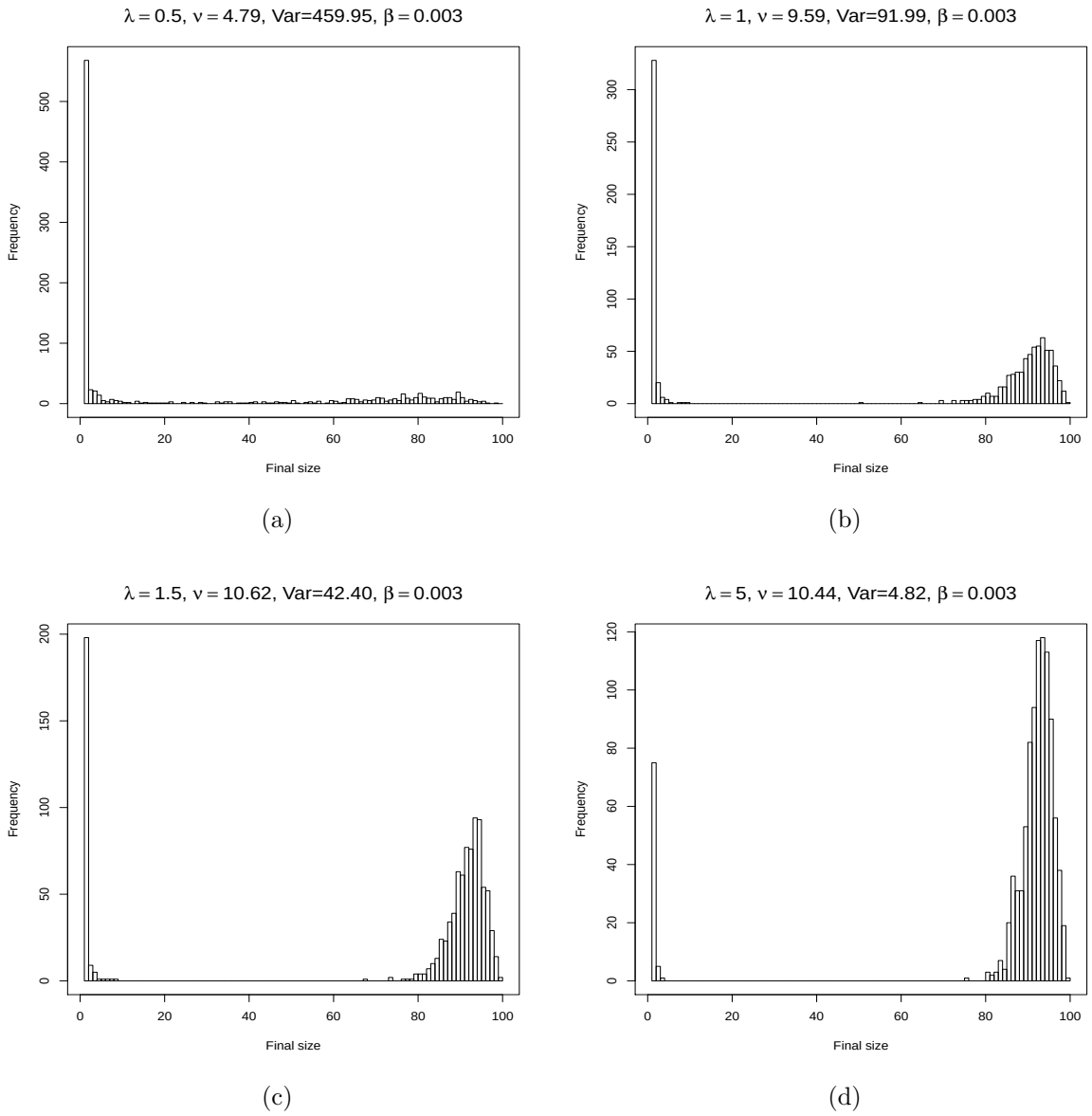


Figure 2.6: Final epidemic size after 1000 realisations of the non-Markovian process (Weibull infectious period) model with $S(0) = 99$, $I(0) = 1$, $R(0) = 0$, $\beta = 0.003$, $T_{max} = 60$ and fixed mean(9.59) with different variances of the infectious period .

2.2.4 Conclusions

This chapter reviews two fundamental approaches to specify the stochastic process for epidemic models, the Gillespie algorithm and Sellke construction. Clearly, the use of Sellke thresholds preserves the sampling properties of the Gillespie algorithm. One of the most significant features to emerge from the Sellke construction in the representation of the dynamics of an epidemic process, at least for an SI model, is that the outcome $\underline{x}$ of the epidemic is obtained deterministically as a function h of the model parameters given the thresholds ($\underline{x} = h(\boldsymbol{\theta}, \boldsymbol{Q})$). A direct benefit of using

this non-centered parametrisation is that different analysis could be applied on a given realisation of Sellke thresholds $\boldsymbol{Q}$, observing the effect of changing the infectious period distribution (see Figure 2.6) or the effect of a different contact rate (see Figure 2.4) for instance.

Taken together, this non-centered parametrisation suggests that different control strategies could be compared on the same set of Sellke thresholds and the outcome of controls are expected to be a deterministic function of controls given the thresholds and the parameters and possibly the infectious period. We will discuss this more thoroughly in the remaining part of the thesis.

Chapter 3

Statistical inference

3.1 Fundamentals of Bayesian approach

In recent decades, the Bayesian paradigm has become the most popular statistical approach used in epidemic modelling partly due to its flexibility and that, in addition to specifying a statistical model based for the observed data $\mathbf{y}$, given a fixed vector of parameters $\boldsymbol{\theta}$, it allows incorporation of prior information on $\boldsymbol{\theta}$. Contrary to frequentist inference, the model parameter $\boldsymbol{\theta}$ is treated as a random variable with a prior distribution $\pi(\boldsymbol{\theta})$.

Let us define $\pi(\boldsymbol{\theta}|\mathbf{y})$ to be the probability density function of the unknown quantity $\boldsymbol{\theta}$ given what we observe, $\mathbf{y}$. The key idea in Bayesian framework can be summarised as follows. We have a certain belief regarding $\boldsymbol{\theta}$ before the data were obtained. This knowledge is then updated through the data using Bayes rule to obtain the so-called *posterior distribution* $\pi(\boldsymbol{\theta}|\mathbf{y})$. Any estimation is then based on the posterior distribution of the parameter to estimate, given by

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\boldsymbol{\theta})\pi(\mathbf{y}|\boldsymbol{\theta})}{\int \pi(\boldsymbol{\theta}')\pi(\mathbf{y}|\boldsymbol{\theta}')d\boldsymbol{\theta}'}, \tag{3.1.1}$$

where $\pi(\mathbf{y}|\boldsymbol{\theta})$ is called the likelihood function of the parameter $\boldsymbol{\theta}$.

The above equation is known as Bayes formula. Since the denominator in (3.1.1) is independent on the parameter $\boldsymbol{\theta}$, it can be omitted from the Equation (3.1.1) subject to the integral being finite. We then have

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\boldsymbol{\theta})\pi(\mathbf{y}|\boldsymbol{\theta}). \tag{3.1.2}$$

In other words, Bayes' idea can be merely seen as

$$posterior \propto prior \times likelihood.$$

3.2 Likelihood

The likelihood plays a central role in statistical inference including classical and Bayesian statistics. Given an observation $\mathbf{y}$, the likelihood of a parameter $\boldsymbol{\theta}$ can be seen informally as the probability of the data given the parameter, more precisely $\pi(\mathbf{y}|\theta)$. In statistics, distinction is made between probability and likelihood since the likelihood is used to described a function of a parameter given an observation whereas the probability refers to a function of the data given a fixed parameter. However, the likelihood principle specifies the real importance of the likelihood. The principle states that all the information about $\boldsymbol{\theta}$ in an observed outcome is contained in the likelihood. In addition, if $\mathbf{y_1}$ and $\mathbf{y_2}$ are two observations such that

$$\pi(\mathbf{y}_1|\boldsymbol{\theta}) = k(\mathbf{y}_1, \mathbf{y}_2)\pi(\mathbf{y}_2|\boldsymbol{\theta}) \text{ for all } \boldsymbol{\theta}, \tag{3.2.1}$$

where $k(\mathbf{y}_1, \mathbf{y}_2)$ is a constant independent of $\boldsymbol{\theta}$, then the conclusion regarding $\boldsymbol{\theta}$ drawn from $\mathbf{y}_1$ and $\mathbf{y}_2$ should be identical. That is, if two samples have proportional likelihoods, then they contain the same information about $\boldsymbol{\theta}$. Full details of the likelihood principle could be found in (Casella and Berger, 2002). Therefore, Bayesian statistics respects the likelihood principle since the data enter the calculations *via* the likelihood.

3.3 Prior distribution

The subjectivity in the Bayesian approach arises in the assignment of the prior knowledge known as the prior distribution, or simply "prior", of the parameter $\boldsymbol{\theta}$ when conducting inference. This is the main source of controversy in this approach in the frequentist point of view. Indeed, in Bayesian inference, the parameter is treated as a random variable (because it is unknown) and the belief about $\boldsymbol{\theta}$ is quantified by a distribution $\pi(\boldsymbol{\theta})$ before looking at the observed outcome $\mathbf{y}$. There are several approaches to specifying a prior distribution. Here, we focus on the common ones.

3.3.1 Conjugate priors

In many situations, some choices of priors are more computationally efficient than others. It may be possible to derive the posterior distribution so that it is of the same family of distributions as the prior density function. Such a prior is known as a *conjugate prior*. This type of prior is useful in that it can allow one to compute the posterior distribution without numerical integration. In many situations, the likelihood function is often drawn from the exponential family of models. Morris (1983) showed that models in this family possess a conjugate prior distribution.

3.3.2 Noninformative priors

The central point of Bayesian inference is that it offers the possibility to include prior knowledge of a system into an analysis. However, such knowledge is not always available. The Bayesian approach reflects the lack of information on the parameter by assigning a 'vague' or a noninformative distribution to the model parameter $\boldsymbol{\theta}$. The aim in this case is that all of information about $\boldsymbol{\theta}$ in the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$ should come from the observed data $\mathbf{y}$ and hence any inference is based entirely on the likelihood. In the particular case where the prior is improper, i.e. $\int \pi(\boldsymbol{\theta})d\boldsymbol{\theta} = \infty$, inference is still possible in the Bayesian framework whenever the denominator of the posterior distribution in Equation 3.1.1 is defined or is finite. In this case, a proper posterior distribution with $\int \pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} = 1$ can be defined.

3.4 Posterior distribution

As mentioned at the beginning of this chapter, inference in the Bayesian setting is achieved *via* the posterior distribution of the parameter of interest $\pi(\boldsymbol{\theta}|\mathbf{y})$ given in Equation 3.1.1 which contains all information regarding the parameter. A density plot of the posterior distribution, for instance, could give a useful representation of belief on the parameter. However, its interpretation may be difficult. In this section, we discuss some of the common methods for summarising Bayesian posterior distributions.

3.4.1 Bayesian point estimation

To obtain point estimates, the Bayesian approach often uses the posterior mean ($\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta}|\mathbf{y})$, median ($\hat{\boldsymbol{\theta}} : \int_{-\infty}^{\hat{\boldsymbol{\theta}}} \pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} = 1/2$) or mode ($\hat{\boldsymbol{\theta}} : \pi(\hat{\boldsymbol{\theta}}|\mathbf{y}) = \sup_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta}|\mathbf{y})$) as a summary feature of the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$ which can be used as a point estimate for $\boldsymbol{\theta}$. Among these three estimates, the mode is the simplest to compute since it does not involve any integration. Note that for a constant prior, the posterior mode is identical to the maximum likelihood estimate. However, the choice depends on the shape of the posterior distribution, though the median might be used since it represents better the center of an asymmetrical distribution. There is, however a vast body of literature e.g. Carlin and Louis (2000) where this topic is discussed in more details.

3.4.2 Interval estimation

In contrast to a point estimate, an interval estimation identifies a range of possible values of the parameter. In Bayesian inference, the $100(1-\alpha)\%$ credible interval, analogous to the confidence interval (CI) in frequentist statistics, is a set $\mathbf{C} = (a, b)$

such that

$$1 - \alpha \leq \pi(\mathbf{C}|\mathbf{y}) = \int_{\boldsymbol{C}} \pi(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \tag{3.4.1}$$

The above interval is often used when the posterior distribution is symmetric and unimodal. However, the highest posterior density (HPD) credible set is preferred in case of a non-symmetric or multi-modal posterior distribution. The HPD is obtained by taking values of $\boldsymbol{\theta}$ that have a posterior density higher than some constant

$$\mathbf{C} = \{\boldsymbol{\theta} : \pi(\boldsymbol{\theta}|\mathbf{y}) \geq K(\alpha)\}, \tag{3.4.2}$$

where $K(\alpha)$ stands for the largest constant such that

$$1 - \alpha \leq \pi(\mathbf{C}|\mathbf{y}). \tag{3.4.3}$$

As the computation of such an interval is not analytically straightforward, therefore, numerical approaches are often used.

3.5 Bayesian computational methods

The Bayesian approach in statistical modelling has been historically difficult to implement due to the high-dimensional integration that it required. However, progress in computing in recent years has made these integrations, once cumbersome, feasible. Several approaches have been proposed. In this section, we only focus our discussion on, arguably, the most widely used: Markov Chain Monte Carlo (MCMC).

3.5.1 Stochastic simulation

Assume that we are interested in the following integral

$$I = \int g(x) dx \tag{3.5.1}$$

A traditional approach would be to use the Riemann summation, which consist of evaluating $g(x)$ at n points $(x_1, \ldots, x_n)$ in a regular grid and therefore

$$I \approx \frac{1}{n} \sum_i g(x_i) \tag{3.5.2}$$

However, if there is a probability density $\pi(x) \geq 0$ such that

$$I = \int \frac{g(x)}{\pi(x)} \pi(x) dx \tag{3.5.3}$$

and if we are able to draw samples x_i i.i.d from π, then by the law of large numbers, Equation 3.5.3 could be approximated as

$$\hat{I}_n = \frac{1}{n} \sum_i \frac{g(x_i)}{\pi(x_i)} \tag{3.5.4}$$

Also, the variance could be estimated using the central limit theorem as

$$v_n = \frac{1}{n(n-1)} \sum_{j=1}^{n} \left(\frac{g(x_j)}{\pi(x_j)} - \hat{I}_n \right)^2$$

Sampling from π remains the key challenge. Techniques such as MCMC provide valuable tools to draw samples from such a distribution. The key idea in MCMC is to design a Markov Chain such that the target distribution π is the stationary distribution. In Bayesian analysis we are particularly in the case where the stationary distribution is the posterior distribution of a parameter of interest.

3.5.2 Markov Chain Monte Carlo (MCMC)

If there is one method that has revolutionized statistics in these recent years, then it is Markov Chain Monte Carlo (MCMC). MCMC is a stochastic simulation extensively used in Bayesian inference. MCMC generates, by using a Markov chain, samples from a target distribution π, where π is designed to be the equilibrium distribution of the Markov chain used.

3.5.2.1 Introduction to Markov Chain

A Markov chain is basically a stochastic process that undergoes transitions from one state to another on a state space S where the distribution of the next state depends only on the current state. More formally, if $(X_1, \ldots, X_n)$ is a sequence of random variables where X_t denotes a random value at time t then

$$\pi(X_{n+1} \in A | X_0 = s_0, \ldots, X_n = s_n) = \pi(X_{n+1} \in A | X_n = s_n) = \int_A p(s_n, x) dx, \tag{3.5.5}$$

with $s_0, \ldots, s_n \in S \supseteq A$.

The above definition ensures that the random variable possesses the memoryless property, i.e. the future state of the chain only depends on its history through its current state. For fixed x in S, $p(s_n, x)$ is the density function of a move to state x given that currently we are the state s_n $(X_{n+1} | X_n = s_n)$. The $n-$steps transition

density $p^{(n)}(s_0, x)$ of $(X_n|X_0 = s_n)$ is defined as

$$\pi(X_n \in A|X_0 = s_0) = \int_A p^{(n)}(s_0, x)dx. \tag{3.5.6}$$

In the case where S is discrete, a Markov chain is defined by its transition probabilities p_{ij}, which is the probability of moving from state s_i to the state s_j in a single step i.e.

$$p_{ij} = \pi(X_{n+1} = s_j|X_n = s_i) \tag{3.5.7}$$

Indeed, if we define $\pi_j(n+1)$ as the probability of being at the state s_j at time $n+1$, then from the Chapman-Kolomogrov equation,

$$\begin{aligned}\pi_j(n+1) &= \pi(X_{n+1} = s_j) \\ &= \sum_i \pi(X_{n+1} = s_j|X_n = s_i)\pi(X_n = s_i) \\ &= \sum_i p_{ij}\pi(X_n = s_i)\end{aligned} \tag{3.5.8}$$

By defining the probability transition matrix as the matrix whose i, j^{th} element is p_{ij}, Equation 3.5.8 becomes

$$\pi(n+1) = \pi(n)P \tag{3.5.9}$$

where $\pi(n)$ is the vector of the $\pi_j(n), \forall j$ as defined in Equation (3.5.8). Iterating the Equation 3.5.9 we obtain $\pi(n) = \pi(0)P^n$. In other words, if the initial distribution is q, then $q^{(n)} = qP^n$ with the $ij-$th element of P^n, $p_{ij}^{(n)}$ defining the n-step transition probability that the process is at state j given that n steps ago it was at state i i.e.

$$p_{ij}^{(n)} = \pi(X_n = s_j|X_0 = s_i) \tag{3.5.10}$$

If the transition matrix P has a unique eigenvalue of modulus one (which must be equal to one) there is a unique distribution $\bar{\pi}$ that satisfies the following relation

$$\bar{\pi} = \bar{\pi}P \tag{3.5.11}$$

$\bar{\pi}$ is referred to as the stationary distribution. Note that for any initial distribution, a Markov chain would, in effect, reach its equilibrium after some suitably large number of iterations.

Properties 1. A Markov chain is said to be **irreducible** if from any state it is possible to reach any other state. This implies there exists an integer n such that $p_{ij}^{(n)} > 0$.

2. An **aperiodic** Markov chain is a chain in which the state space cannot be partitioned into disjoint subsets such that the chain is constrained to cycle through these subsets.

For the chain to reach its stationary distribution from any initial starting state, it has to be irreducible, aperiodic and possess a stationary distribution. In practice, the aperiodicity is mostly straightforward to show, but the irreducibility is not always trivial to show. However, to show the stationary property, a stronger property which states that P satisfies the **detailed balance** property with respect to the distribution $\bar{\pi}$ is often preferred. Recall that the **detailed balance** property states that for all i, j, if $\bar{\pi}$ denotes the distribution and P a transition matrix,

$$\bar{\pi}_i p_{ij} = \bar{\pi}_j p_{ji} \tag{3.5.12}$$

It then follows that $\bar{\pi}P = \bar{\pi}$. Details of the proof along with other properties of the Markov Chain could be found in Roberts (2012).

Two particular MCMC methods are widely used in Bayesian analysis: Metropolis-Hastings and Gibbs sampling.

3.5.2.2 Metropolis-Hastings algorithm

Initially introduced in the chemical physics literature to integrate over very complex functions by simply drawing random samples (Metropolis and Ulam, 1949; Metropolis et al., 1953; Hastings, 1970), the Metropolis-Hastings (M-H) algorithm is particularly appropriate in scenarios where a posterior distribution is known only up to an unknown normalisation constant.

We consider a target distribution π from which we wish to draw samples on a finite state space. For this purpose, the M-H algorithm generates random draws as follows:

1. Start with an arbitrary initial point $\boldsymbol{\theta}^0$ and choose a probability density $q(y, x)$ from which to propose the next candidate y given the current value x. This probability density q is referred to as the proposal distribution.

2. Suppose that the state after i iterations is θ^i. We propose a move to a new candidate $\boldsymbol{\theta}^*$ generated from $q(\theta^*, \boldsymbol{\theta}^i)$. Then find the acceptance ratio

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = min\left\{1, \frac{\pi(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^*, \boldsymbol{\theta}^i)}{\pi(\boldsymbol{\theta}^i)q(\boldsymbol{\theta}^i, \boldsymbol{\theta}^*)}\right\} \tag{3.5.13}$$

   The new candidate is accepted with probability $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$. This is done by simulating a random number u uniformly distributed between 0 and 1. If $u \leq \alpha$ the new point is then accepted and we set $\boldsymbol{\theta}^{i+1} = \boldsymbol{\theta}^*$, otherwise the new candidate

is rejected and we set $\boldsymbol{\theta}^{i+1} = \boldsymbol{\theta}^i$. This generates a Markov Chain $(\boldsymbol{\theta}^0, \ldots, \boldsymbol{\theta}^n)$ which converges after a large number of iterations to a stationary distribution that coincides with the target distribution.

Moreover, the transition probability $p(x, y)$ to move from state x to another state y, $x \neq y$ is given by

$$p(x, y) \propto q(x, y)\alpha(x, y) = q(x, y) \min\left\{1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right\} \tag{3.5.14}$$

It can be shown that the M-H satisfies detailed balance.

The efficiency of the algorithm depends on the choice of the proposal distribution. Therefore caution must be taken in this regard.

3.5.2.3 Metropolis algorithm

The most popular algorithm introduced by Metropolis et al. (1953) is a particular case of the Metropolis-Hastings algorithm described above, where the proposal distribution q is chosen to be symmetric, so that $q(x, y) = q(y, x)$. In this case the acceptance ratio reduces to

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}^i)}\right\} \tag{3.5.15}$$

a common choice to ensure a symmetric proposal distribution is to set $q(x, y) = f(|x - y|)$. This results in an algorithm known as a *random-walk* Metropolis since in the absence of the acceptance/rejection steps the Markov-Chain would perform a random walk.

3.5.2.4 Gibbs sampler

Originally used in statistical physics and later introduced in image processing (Geman and Geman, 1984), Gibbs sampling is a MCMC algorithm for obtaining a sequence of random draws. It has been applied extensively in Bayesian statistics. Moreover, Gibbs sampling is a specific case of Metropolis-Hastings wherein the acceptance ratio $\alpha = 1$. This means we accept all proposed moves.

Suppose that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ is the parameter vector. The usual form of the algorithm is as follows:

1. Start with an initial value $\boldsymbol{\theta}^0 = (\theta_1^0, \ldots, \theta_n^0)$ arbitrarily chosen.

2. Assuming that after k iterations we reach the state $\boldsymbol{\theta}^k = (\theta_1^k, \ldots, \theta_n^k)$. Then the next draw, θ_i^{k+1} is obtained by drawing each component in turn from its full conditional distribution given state of other components. That is

$$\theta_i^{k+1} \sim \pi(\theta_i | \mathbf{y}, \theta_1^{k+1}, \ldots, \theta_{i-1}^{k+1}, \theta_{i+1}^k, \ldots, \theta_n^k), \qquad \forall i = 1, \ldots, n.$$

As the number of iterations is increasing, the distribution of the state converges to the target posterior distribution subject to irreducibility and aperiodicity.

3.5.2.5 Reversible-jump MCMC

Real-life epidemic data are typically rather less complete than would be desirable for modelling and inference. In an ideal world, one would have observed the time at which each individual contracts the epidemic (infection times), the source of infection (who transmits the disease to whom) and maybe the time taken for an infected individual to cease being infectious. In many particular situation, only the times at which symptoms are detected are observed. Moreover, neither the infection times nor the transmission network nor, frequently, the size of the epidemic are known.

In the case where the number of individuals who contracted the disease out of the initial susceptible population is unknown, the traditional stochastic integration methods such as Metropolis-Hastings and Gibbs sampling fail to incorporate the lack of knowledge of the unobserved or 'cryptic' infections. However, there exists an extension of the M-H algorithm, namely Reversible Jump Markov Chain Monte Carlo, which provides the tools to tackle this issue.

Reversible Jump Markov Chain Monte Carlo (RJMCMC) is a general form of the Metropolis-Hastings algorithm. Green (1995) extends the most popular Metropolis-Hastings algorithm to allow transitions between subspaces of the parameter space with different dimensionality. This approach is often used for Bayesian model choice but has found a wide applicability in epidemiology since, most of the time, there is no precise knowledge of the number of infected individuals during the course of the epidemic.

The algorithm involves the steps of proposing and accepting/rejecting a new state, as in the case of the M-H algorithm. We first propose to move to a different space and then accept that move with a certain probability or reject it. The details of the algorithm is as follows:

Assume that after k iterations we reach parameter $\boldsymbol{\theta}^k \in \mathbb{R}^m$. Then we propose to move to $\boldsymbol{\theta}^* \in \mathbb{R}^{m'}$. To do that, we generate two vectors u and u' of random variables respectively with density function $q(u)$ and $q'(u')$ such that there exist a bijective function g that satisfies the relation $(\boldsymbol{\theta}^*, u') = g(\boldsymbol{\theta}^k, u)$. It is worth noting that this condition is required for dimension-matching and to allow a reverse move. We accept the move with probability

$$\alpha = \min\left\{1, \frac{\pi(\boldsymbol{\theta}^*|\mathbf{y})j([\boldsymbol{\theta}^k]|[\boldsymbol{\theta}^*])q'(u')}{\pi(\boldsymbol{\theta}^k|\mathbf{y})j([\boldsymbol{\theta}^*]|[\boldsymbol{\theta}^k])q(u)}\left|\frac{\partial(\boldsymbol{\theta}^*, u')}{\partial(\boldsymbol{\theta}^k, u)}\right|\right\} \tag{3.5.16}$$

where $j([\boldsymbol{\theta}^k]|[\boldsymbol{\theta}^*])$ is the probability of choosing to move from the space $[\boldsymbol{\theta}^k]$, space generated by the vector $\boldsymbol{\theta}^k$ to $[\boldsymbol{\theta}^*]$, space generated by the vector $\boldsymbol{\theta}^*$, and the joint

posterior distribution of the parameter $\pi(\boldsymbol{\theta}|\mathbf{y})$ is given up to proportionality by the product of the likelihood and the prior; that is:

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\mathbf{y}|\theta)\pi(\boldsymbol{\theta}). \tag{3.5.17}$$

In practice, especially in epidemiology where moves are merely deletion, moving and addition of infection times (Gibson and Renshaw, 1998; O'Neill and Roberts, 1999; Streftaris and Gibson, 2004)), we do not need to generate both u and u'. For instance, we only generate u such that $|u| = 1$ if the move consists of inserting a new infection time u. As a result, the acceptance probability becomes:

$$\alpha = \min\left\{1, \frac{\pi(\boldsymbol{\theta}^*|\mathbf{y})j([\boldsymbol{\theta}^k]|[\boldsymbol{\theta}^*])}{\pi(\boldsymbol{\theta}^k|\mathbf{y})j([\boldsymbol{\theta}^*]|[\boldsymbol{\theta}^k]))q(u)}\left|\frac{\partial(\boldsymbol{\theta}^*)}{\partial(\boldsymbol{\theta}^k, u)}\right|\right\} \tag{3.5.18}$$

In this case, by setting $\boldsymbol{\theta}^* = (\boldsymbol{\theta}^k, u)$ the Jacobian $\left|\frac{\partial(\boldsymbol{\theta}^*)}{\partial(\boldsymbol{\theta}^k, u)}\right| = 1$, therefore the acceptance probability of a move that consists of adding a new infection time to the set $\boldsymbol{\theta}^k$ is given by:

$$\alpha = \min\left\{1, \frac{\pi(\boldsymbol{\theta}^*|\mathbf{y})j([\boldsymbol{\theta}^k]|[\boldsymbol{\theta}^*])}{\pi(\boldsymbol{\theta}^k|\mathbf{y})j([\boldsymbol{\theta}^*]|[\boldsymbol{\theta}^k])q(u)}\right\} \tag{3.5.19}$$

For the reverse move (deletion), we set $(\boldsymbol{\theta}^*, u') = \boldsymbol{\theta}^k$, and the acceptance probability is

$$\alpha = \min\left\{1, \frac{\pi(\boldsymbol{\theta}^*|\mathbf{y})j([\boldsymbol{\theta}^k]|[\boldsymbol{\theta}^*])q'(u')}{\pi(\boldsymbol{\theta}^k|\mathbf{y})j([\boldsymbol{\theta}^*]|[\boldsymbol{\theta}^k])}\right\} \tag{3.5.20}$$

It is important to note that here, the vector $\boldsymbol{\theta}$ represents the set of infection times. In practice $\boldsymbol{\theta}$ must incorporate both the parameters of the epidemic model and the infection times.

3.5.2.6 Practical implementation of MCMC

Having designed a Markov chain which converges to the target distribution (the posterior distribution in the Bayesian framework), it is important to consider some aspects before drawing conclusions on the target distribution π. If we assume that the chain has started at some initial state s_0, then the number of iterations required for the chain to reach its stationary distribution depends on s_0, the initial state, and the choice of the proposal distribution. A poor choice of s_0 or the proposal distribution will demand a large number of iterations. Therefore the first n' (depending on the choice of s_0 and the proposal distribution) samples are discarded. This is known as the *burn-in period*.

A further issue is the *mixing* problem. In general s_i and s_{i+k} are not independent unless k is very large. To approximate an i.i.d. sample from the stationary distribution we can *thin* the chain at every m^{th} iteration generating a sample of the form

$\{s_i, s_{i+m}, s_{i+m}, \ldots\}$. Furthermore, there exist a collection of convergence diagnostics methods in the MCMC literature including Geweke's method (Geweke, 1991), the method by Raftery and Lewis (Raftery and Lewis, 1992) and Gelman and Rubin's method (Gelman and Rubin, 1992) to assess the convergence of the chain. There is some disagreement on whether these techniques guarantee convergence has occurred mainly due to the fact that there are different type of convergence (Roberts, 2012).

Moreover, there are several packages that make these analyses feasible. Arguably the most popular is WINBUGS, the current windows-based version of BUGS described by Lunn et al. (2000), BOA (Bayesian output analysis) by Smith (2007) and CODA(Convergence Diagnosis and Output Analysis) by Plummer et al. (2006).

3.5.2.7 Data-augmentation method

As mentioned in previous sections, data analysis is most straightforward when the data are complete. In general, epidemic data are rarely fully observed. For instance the precise infection times are not observed and this renders the data incomplete. In the latter case, the likelihood is typically intractable and any inference on the model parameters becomes difficult to handle. The technique of data augmentation provides a solution to this problem.

Data augmentation is the most widely adopted computational method for performing Bayesian analysis of missing data. The basic idea is that the observed data $\mathbf{y}$ are augmented with latent data $\mathbf{z}$ representing unobserved quantities, such as the infection times for example, such that the likelihood $\pi(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ stands for the model parameter, is tractable. Note that Bayesian inference usually aims to estimate the marginal posterior distribution of the parameters $\pi(\boldsymbol{\theta}|\mathbf{y})$.

The method is motivated by the following relation:

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \int_Z \pi(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}) dz \tag{3.5.21}$$

The above integration is straightforward using MCMC methods described previously.

We sample from the joint distribution $\pi(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$ using a range of MCMC techniques. Thus, sampling both parameters and censored data depends on the form of each distribution $\pi(\boldsymbol{\theta}|\mathbf{y}, z)$ and $\pi(z|\boldsymbol{\theta}, \mathbf{y})$. Practical implementation of the approach tends to use a mixture of Gibbs sampling and Metropolis-Hastings. Reversible-jump is particularly valuable when the range of z has components of differing dimensions as is typically the case in epidemic modelling.

3.6 Bayesian epidemic modelling

We now illustrate how the data augmentation approach can be applied to fit a stochastic model to a partially observed epidemic.

3.6.1 Example of SIR model with complete epidemic

In this example we use data from a smallpox epidemic (Bailey, 1975) in a Nigerian village along with the method described by Streftaris and Gibson (2004). See also O'Neill and Roberts (1999).

The data consist of a total of 30 cases in a population of size 120. Indeed, the observation reveals 30 removal times given in the Table 3.1. Time 0 indicates the first removal time. Occurrence of identical times in the table indicates cases appearing the same day. The model chosen to describe this process is exactly the non-Markovian SIR

0	13	20	22	25	25	25	26	30	35	38	40	40	42	42
47	50	51	55	55	56	57	58	60	60	61	66	66	71	76

Table 3.1: Smallpox data.

model described in Section 2.2.3.2. Recall that in this model, a susceptible becomes infected only by contact with an infective. The transition probability that describes this process is given in Equations (2.2.3). Moreover, an infected individual remains in the infected state for a random time, known as the *infectious period* and here modelled by a Weibull distribution as described in Equation (2.2.6).

We denote by $\boldsymbol{r} = (r_1 = 0, r_2, \ldots, r_{n_R})$ the ordered removal times observed during the period $[0, t_{obs}]$ and $\boldsymbol{s} = (s_1, s_2, \ldots, s_{n_I})$ the ordered infection times, $\mathcal{R}$ and $\mathcal{I}$ the set of removals and the set of infected individual respectively. We assume that the epidemic is complete so that no other infections have occurred during its course. Following O'Neill and Roberts (1999); O'Neill and Becker (2001); Britton and O'Neill (2002), the likelihood is given as:

$$L(\beta, \nu, \lambda; \boldsymbol{s}, \boldsymbol{r}) = \prod_{j \in \mathcal{I}_{-1}} \{\beta I(s_j^-)\} \exp\left\{ - \int_{s_1}^{t_{obs}} \beta S(t) I(t) dt \right\}$$

$$\times \prod_{j \in \mathcal{R}} [\nu \lambda (r_j - s_j)^{\nu - 1} \exp\{-\lambda (r_j - s_j)^{\nu}\} \tag{3.6.1}$$

The integral in Equation 3.6.1 needs to be discretised by transformation into a sum over successive events. To that end, an alternative representation is derived by Britton

and O'Neill (2002); Neal and Roberts (2005) as

$$\int_{s_1}^{t_{obs}} S(t)I(t)dt = \sum_{i=1}^{n_R}\sum_{j=1}^{N}(r_i \wedge s_j) - (s_i \wedge s_j), \tag{3.6.2}$$

where $s_{n_I+1} = s_{n_I+2} = \ldots = s_N = \infty$. Equivalently, Streftaris and Gibson (2012) proposed the form

$$\int_{s_1}^{t_{obs}} S(t)I(t)dt = \sum_{i=1}^{n_I+n_R} S(t_i^-)I(t_i^-)(t_i - t_{i-1}) \tag{3.6.3}$$

where $(t_1, t_2, \ldots, t_{n_I+n_R})$ is the ordered times of all events (infection and removal).

We shall suppose that β, ν and λ follow, *a priori*, gamma distributions $\Gamma(a, b)$, $\Gamma(c, d)$ and $\Gamma(m, p)$ respectively (O'Neill and Becker, 2001). The joint posterior distribution is then obtained using the Equation (3.1.1) as follows:

$$\pi(\beta, \nu, \lambda, \boldsymbol{s}|\boldsymbol{r}) \propto L(\beta, \nu, \lambda, \boldsymbol{s}; \boldsymbol{r}) \times \beta^{a-1}\nu^{c-1}\lambda^{m-1}\exp(-b\beta - d\nu - p\lambda) \tag{3.6.4}$$

This leads us to the full conditional distribution of the contact rate and the scale parameter of the Weibull distribution :

$$\beta|\boldsymbol{r}, \boldsymbol{s}, a, b \sim \Gamma\left(a + n_I - 1, b + \int_{s_1}^{t_{obs}} S(t)I(t)\right) \tag{3.6.5}$$

$$\lambda|\nu, m, p, \boldsymbol{r}, \boldsymbol{s} \sim \Gamma\left(n_R + m, \sum_{j\in\mathcal{R}}(r_j - s_j)^\nu + p\right) \tag{3.6.6}$$

We can therefore sample directly from these distributions using Gibbs sampling. However, the full conditional distribution of the shape parameter is given as:

$$\begin{aligned}\pi(\nu \mid \lambda, c, d, \boldsymbol{r}, \boldsymbol{s}) \propto \nu^{n_R+c-1}\prod_{j\in\mathcal{R}}(r_j - s_j)^{\nu-1} \\ \times \exp\left[-\lambda\left(\sum_{j\in\mathcal{R}}(r_j - s_j)^\nu\right) - d\nu\right]\end{aligned} \tag{3.6.7}$$

Since this is not in a closed form we adopt a Metropolis-Hastings step to update ν. We use the same approximation used by Streftaris and Gibson (2004) for the Gamma proposal distribution. The shape u and the scale v parameters of the Gamma proposal are solutions of the following equations:

$$\frac{u}{v} = \hat{\nu} \text{ and } \frac{u}{v^2} = -\left[\frac{\partial^2 \log\{\pi(\nu|\lambda, c, d, \boldsymbol{r}, \boldsymbol{s})\}}{\partial\nu^2}\right]^{-1}_{\nu=\hat{\nu}} \tag{3.6.8}$$

where $\hat{\nu}$ is the mode of the density $\pi(\nu|\lambda, c, d, \boldsymbol{r}, \boldsymbol{s})$. We use the method of Brent (Brent, 1977, Ch 3-4) for the maximisation. The new parameter ν' is then accepted with probability

$$\alpha = \min\left\{1, \frac{\pi(\nu'|\lambda, c, d, \boldsymbol{r}, \boldsymbol{s})/q(\nu'|\lambda, c, d, \boldsymbol{r}, \boldsymbol{s})}{\pi(\nu|\lambda, c, d, \boldsymbol{r}, \boldsymbol{s})/q(\nu|\lambda, c, d, \boldsymbol{r}, \boldsymbol{s})}\right\} \tag{3.6.9}$$

where $q()$ is the Gamma proposal obtained from the approximation.

The updates to the infection times $\boldsymbol{s}$ are carried out using a Metropolis-Hastings step by randomly selecting an individual j. If such a host is infected, its infection time s_j is moved uniformly in $[T_0, t_{obs}]$ with T_0 being a lower time prior to the first infection. $T_0 = -\infty$ but in practice must be chosen to be finite. Note that in this framework, any individual can potentially become the initial infection. The new vector of infection times $\boldsymbol{s}'$ is accepted with probability

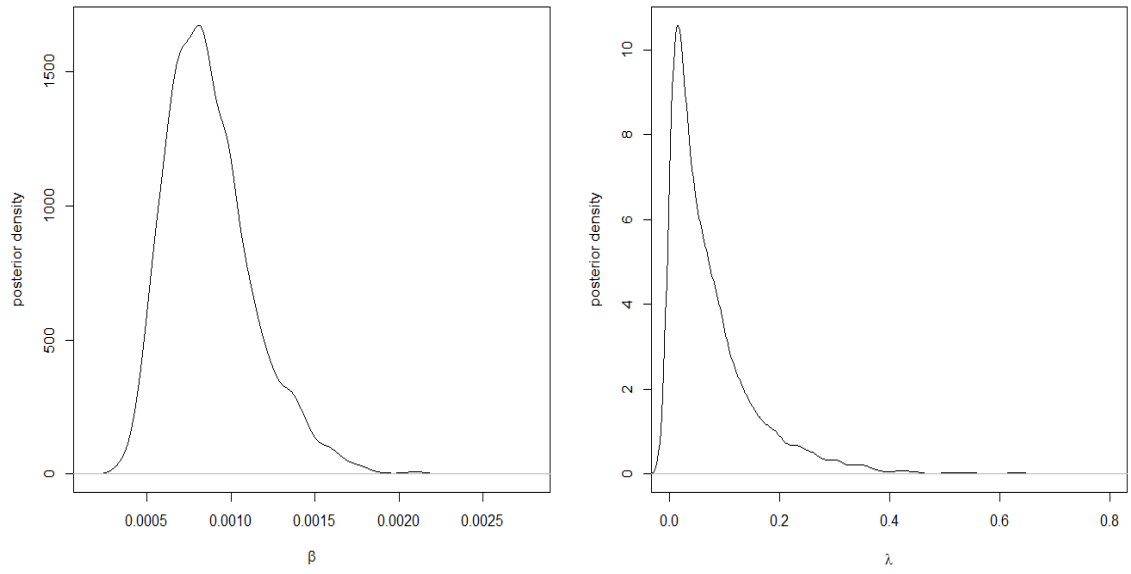
$$\alpha = \min\left\{1, \frac{L(\beta, \lambda, \nu; \boldsymbol{r}, \boldsymbol{s}')}{L(\beta, \lambda, \nu; \boldsymbol{r}, \boldsymbol{s})}\right\}. \tag{3.6.10}$$

We set $a = b = c = d = 0.01$ to represent a vague prior knowledge on parameters β and λ whereas we consider a more informative prior for the shape parameter ν by setting $m = 1$ and $p = 10$.

Figures 3.1, 3.2, 3.3 show the posterior densities for the model parameters β, λ and ν along with the mean of the infectious period (μ), its standard deviation σ and the basic reproduction number R_0 defined to be the expected number of secondary infections produced by a typical infected individual. The summary statistics are shown in Table 3.2. It can be seen that the posterior mean of the basic reproduction number is greater than unity; with a quarter of the population contracting the disease. It is evident that R_0 is probably not too much above 1 since this is a relatively small epidemic. In addition, it is believed that symptoms takes on average a period of $12 - 14$ days after infection to become visible Bailey (1975). This should correspond to the infectious period.These values are consistent with the posterior distribution of μ as shown by the 95% credible interval in Table 3.2.

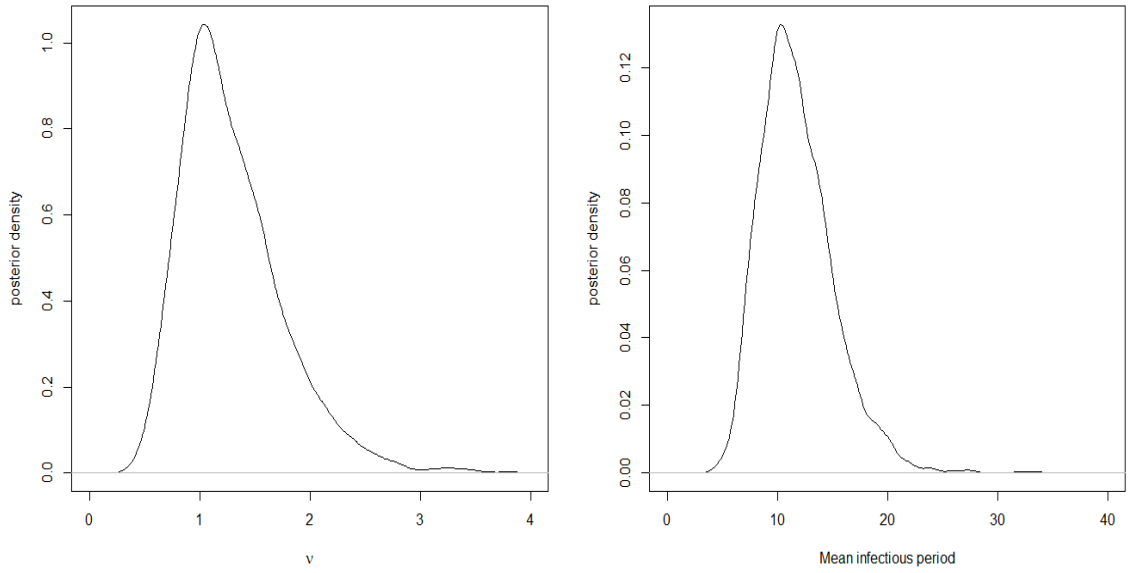
	Mean	SD	MC Error	0.025	Median	0.95
β	0.0008745458	0.0002630835	4.051929e-06	0.0004740828	0.0008355719	0.001486654
λ	0.0759245244	0.0825633211	1.770927e-03	0.0016837871	0.0498433817	0.301120996
ν	1.2974132138	0.4794788040	7.983310e-03	0.6215517380	1.2035599136	2.414046176
μ	11.81044	3.476584	0.05153583	6.72077	11.3632	19.66279
σ	10.49288	6.056765	0.1083857	4.362082	9.286176	23.81925
R_0	1.182915	0.3427175	0.005831918	0.7011448	1.140111	1.927861

Table 3.2: Summary Statistic for the model parameters.



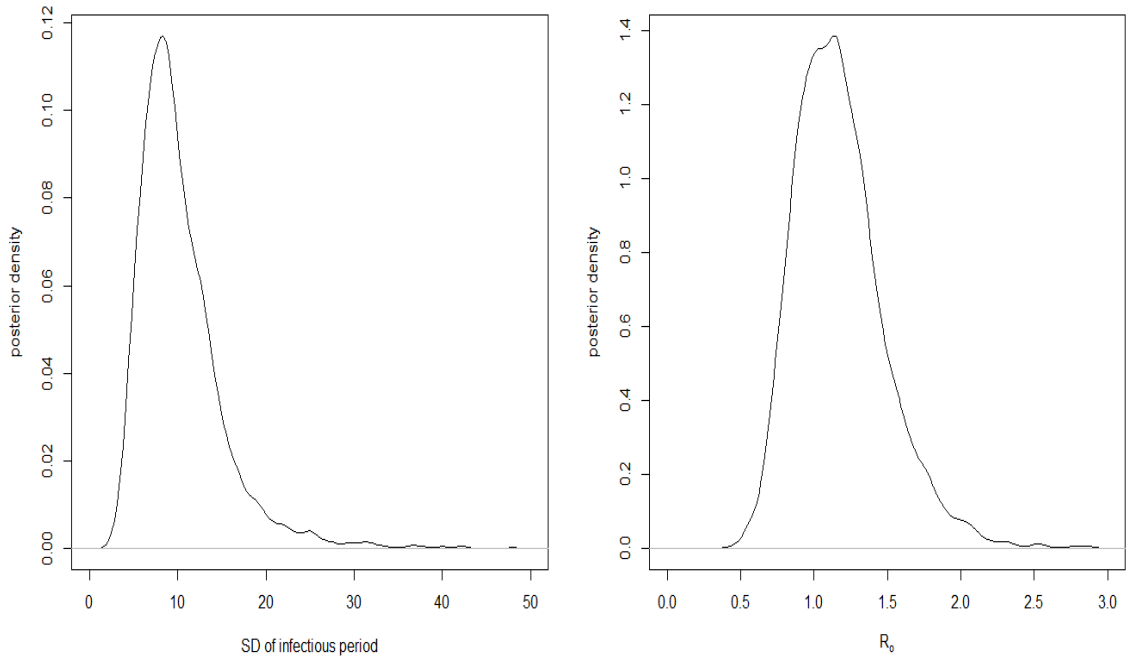
(a) (b)

Figure 3.1: Posterior density of the contact rate β and the scale parameter λ.



(a) (b)

Figure 3.2: Posterior density of the shape parameter ν and the mean infectious period.



(a) (b)

Figure 3.3: Posterior density of the standard deviation of the infectious period and the basic reproduction number R_0 respectively.

3.6.2 Example of SIR model with non-complete epidemic

More frequently, one is interested in making inference about an epidemic in progress. Therefore it is reasonable to make inference on a non-complete epidemic where there are some infected individuals that are not yet detected. In this case $n_I \geq n_R$ where n_I and n_R respectively denote the number of infected and removed individuals. Since the exact number of infections is unknown, inference from such data using traditional MCMC methods including Gibbs sampling and Metropolis-Hastings is impossible. The MCMC method that can address this issue is RJMCMC described in 3.5.2.5.

Gibson and Renshaw (1998), O'Neill and Roberts (1999) and Streftaris and Gibson (2004) have all adapted reversible-jump methods to the epidemic situation. They proposed changes to the infection times by inserting, deleting, or moving the infection time of an individual depending on its current state. We illustrate the effectiveness of this algorithm (see Section 3.5.2.5) in the following simulation.

We simulate a non-complete epidemic from a general epidemic model (Threshold-Weibull model) using the Sellke construction described in Section 2.2.2. We consider a population of size $N = 100$ and the contact rate $\beta = 0.003$. We assume that the infectious period follows a Weibull distribution with shape parameter $\nu = 1.1$, and

scale $\lambda = 0.08$. We observed the epidemic until time 60 where 83 individuals got removed.

The likelihood is given by the Equation which is similar to (3.6.1), with an additional term to account for the contribution of the infected individuals that are not removed before the end of the observation period. That is

$$\begin{aligned} L(\beta, \nu, \lambda, \boldsymbol{s}; \boldsymbol{r}) = & \prod_{j \in \mathcal{I}_{-1}} \{\beta I(s_j^-)\} exp\left\{ - \int_{s_1}^{t_{obs}} \beta S(t) I(t) dt \right\} \\ & \times \prod_{j \in \mathcal{R}} [\nu\lambda(r_j - s_j)^{\nu-1} \exp\{-\lambda(r_j - s_j)^\nu\} \\ & \times \prod_{j \in \mathcal{I} \cap \overline{\mathcal{R}}} \exp\{-\lambda(t_{obs} - s_j)^\nu\} \end{aligned} \tag{3.6.11}$$

As previously, we suppose that β, ν and λ follow *a priori* Gamma distributions $\Gamma(a, b)$, $\Gamma(c, d)$ and $\Gamma(m, p)$ respectively. We obtain the joint posterior distribution as

$$\pi(\beta, \nu, \lambda, \boldsymbol{s}|\boldsymbol{r}) \propto L(\beta, \nu, \lambda, \boldsymbol{s}; \boldsymbol{r}) \times \beta^{a-1}\nu^{c-1}\lambda^{m-1} \exp(-b\beta - d\nu - p\lambda) \tag{3.6.12}$$

The full conditional distribution of the contact rate and the scale parameter of the Weibull distribution are given respectively as:

$$\beta|\boldsymbol{r}, \boldsymbol{s}, a, b \sim \Gamma\left(a + n_I - 1, b + \int_{s_1}^{t_{obs}} S(t) I(t) dt\right) \tag{3.6.13}$$

$$\lambda|\nu, m, p, \boldsymbol{r}, \boldsymbol{s} \sim \Gamma\left(n_R + m, \sum_{j \in \mathcal{R}} (r_j - s_j)^\nu + \sum_{j \in \overline{\mathcal{R}} \cap \mathcal{I}} (t_{obs} - s_j)^\nu + \phi\right) \tag{3.6.14}$$

The full conditional distribution of the shape parameter is given as:

$$\begin{aligned} \pi(\nu|\lambda, c, d, \boldsymbol{r}, \boldsymbol{s}) \propto & \nu^{n_R + c - 1} \prod_{j \in \mathcal{R}} (r_j - s_j)^{\nu-1} \\ & \times \exp\left[-\lambda \left\{ \sum_{j \in \mathcal{R}} (r_j - s_j)^\nu + \sum_{j \in \mathcal{I} \cap \overline{\mathcal{R}}} (t_{obs} - s_j)^\nu \right\} - d\nu \right] \end{aligned} \tag{3.6.15}$$

Again, since this is not in a closed form, we update ν employing a Gamma proposal in a Metropolis-Hastings step as described in the case of a complete epidemic.

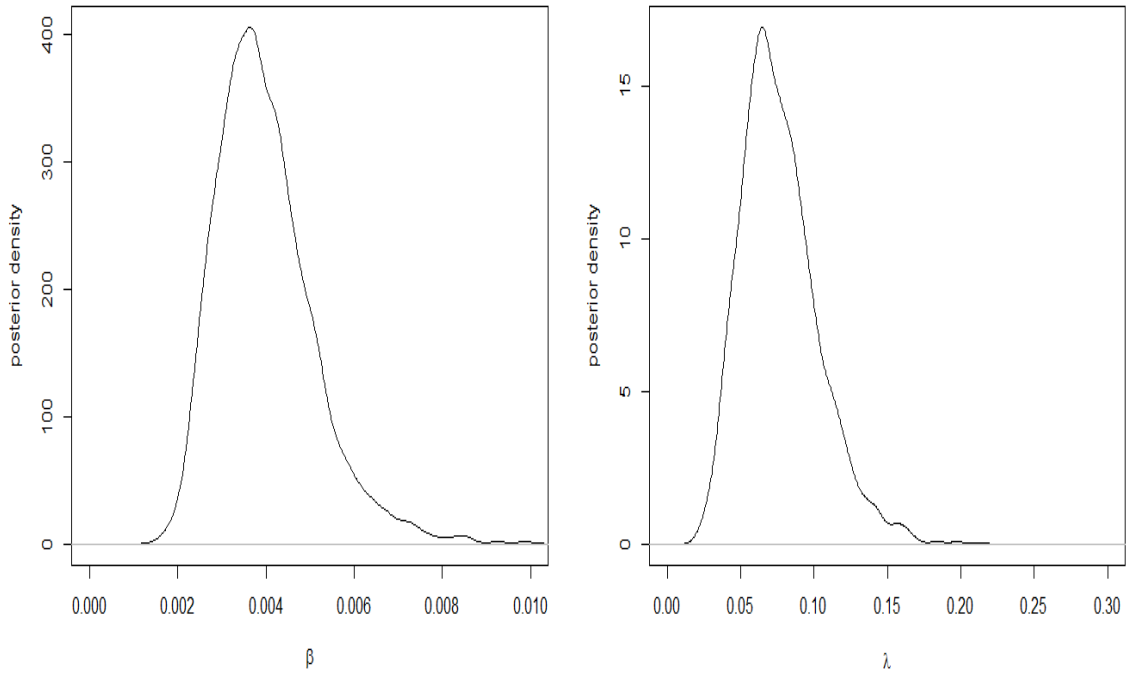
We also consider an informative prior on the parameters setting $a = 3$, $b = 1000$, $c = 11$, $d = 10$, $m = 8$, $p = 100$.

We report the outputs of the MCMC analysis in Table 3.3 and Figures 3.4 and **??**.

Clearly, in both posterior summary and posterior distribution, the true values of the parameters lie within the range of values supported by the posterior distributions.

	Mean	SD	MC Error	0.025	Median	0.95
β	0.003792087	0.0006321461	1.210681e-05	0.002717592	0.003746766	0.005181308
λ	0.102457424	0.0298863501	4.797205e-04	0.052489454	0.099438206	0.169175884
ν	1.257495123	0.1634417421	2.912114e-03	0.967980038	1.247155771	1.618339730

Table 3.3: Summary Statistic for the model parameters.



(a) (b)

Figure 3.4: Posterior density of the parameter β and λ respectively using Reversible Jump MCMC technique on a non-complete epidemic simulated from a General epidemic model (Threshold-Weibull model) on a population of size $N = 100$, $\lambda = 0.08$, $\nu = 1.1$ and $\beta = 0.003$.

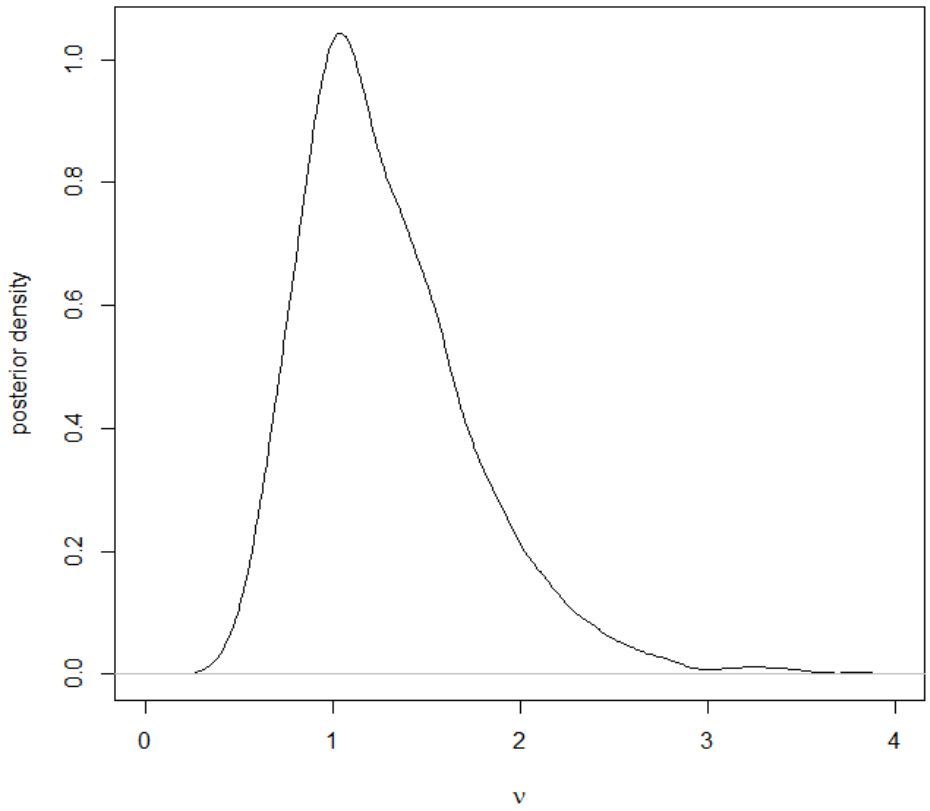


Figure 3.5: Posterior density of the shape parameter ν using Reversible Jump MCMC technique on a non-complete epidemic simulated from a General epidemic model (Threshold-Weibull model) on a population of size $N = 100$, $\lambda = 0.08$, $\nu = 1.1$ and $\beta = 0.003$.

Nonetheless, this algorithm could be modified to predict the dynamics of the process by considering all removals that could happen after the observation time as hidden events. This is done by extending the augmented data to include future events such as infections and removals that could happen beyond the observation time. It is worth noting that this approach could serve as a way of checking the efficacy of the previous algorithm since $\pi(\boldsymbol{\theta}|\mathbf{y})$ does not change if an additional information x is imputed beyond the time of last removal. In this case, the likelihood is given by Equation 3.6.11 with t_{obs} replaced by T, the predicted or assessment time.

3.6.2.1 Details of the algorithm

The diagram in Figure 3.6 shows the transition between different states. Note that the S = susceptible At each iteration of the MCMC we randomly choose an individual j

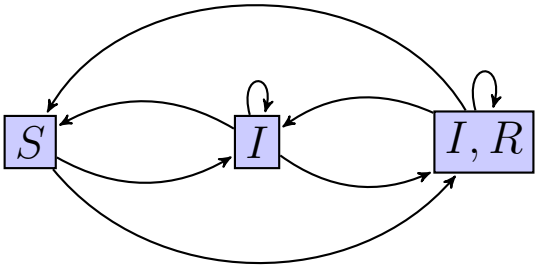


Figure 3.6: State diagram for the infection times and the removal times to show the states transitions. I, R corresponds to the state of individuals removed in the interval $[t_{obs}, T]$, S represents the susceptibles individual and I the infected but not removed.

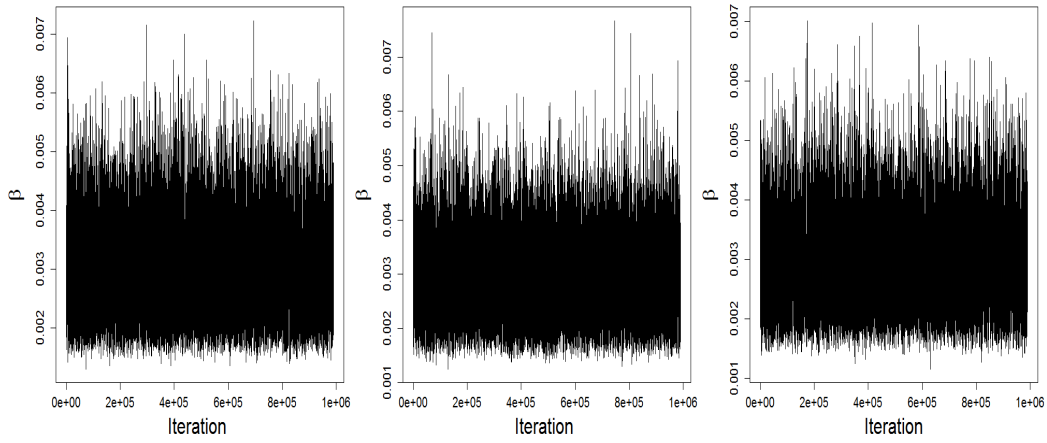
1. If j is susceptible (S), we either add an infection time $s_j \in [T_0, T]$ or both infection and removal times simultaneously where $s_j \in [T_0, T]$ and $r_j \in [\max(s_j, t_{obs}), T]$. The new vector $\boldsymbol{s}'$ is accepted in the first case with probability
$$\alpha = \min\left\{1, \frac{2(T-T_0)}{3}\frac{L(\beta, \lambda, \nu, \boldsymbol{s}'; \boldsymbol{r})}{L(\beta, \lambda, \nu, \boldsymbol{s}; \boldsymbol{r})}\right\} \tag{3.6.16}$$
while in the latter case the acceptance probability is given by Equation (3.6.16) with $\frac{2(T-T_0)}{3}$ replaced by $\frac{(T-\max(s_j,t_{obs}))(T-T_0)}{2}$.

2. If j is infected but not removed (I), we either propose with equal probability to delete the infection time s_j or move uniformly $s_j \in [T_0, T]$ or add a removal time. The new vector $\boldsymbol{s}'$ of the infection times is accepted with probability given by Equation (3.6.16) with $\frac{2(T-T_0)}{3}$ respectively replaced by $\frac{3}{2(T-T_0)}$, omitted and replaced by $\frac{3(T-\max(s_j,t_{obs}))}{4}$.

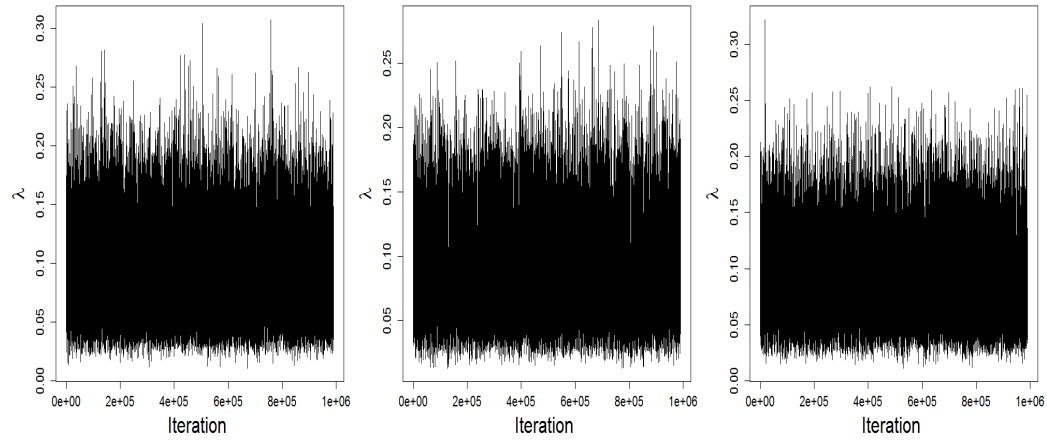
3. If j corresponds to an individual observed to be removed, we can only propose to move its infection time s_j uniformly in $[T_0, t_{obs}]$ and the acceptance probability of the new infection time is then given by Equation (3.6.16) with $\frac{2(T-T_0)}{3}$ omitted.

4. Finally, if j is removed after the observation time (I, R), with equal probability we either delete its removal time r_j or delete the coupled infection and removal time (r_j, s_j), or we move its infection time s_j uniformly in $[T_0, T]$ or move its removal time r_j in $[\max(s_j, t_{obs}), T]$. The acceptance probability is obtained from Equation (3.6.16) with $\frac{2(T-T_0)}{3}$ respectively replaced by $\frac{4}{3(T-\max(s_j,t_{obs}))}$ and $\frac{1}{2(T-\max(s_j,T_{obs}))(T-T_0)}$ in the first two cases respectively; and omitted in the last two cases.

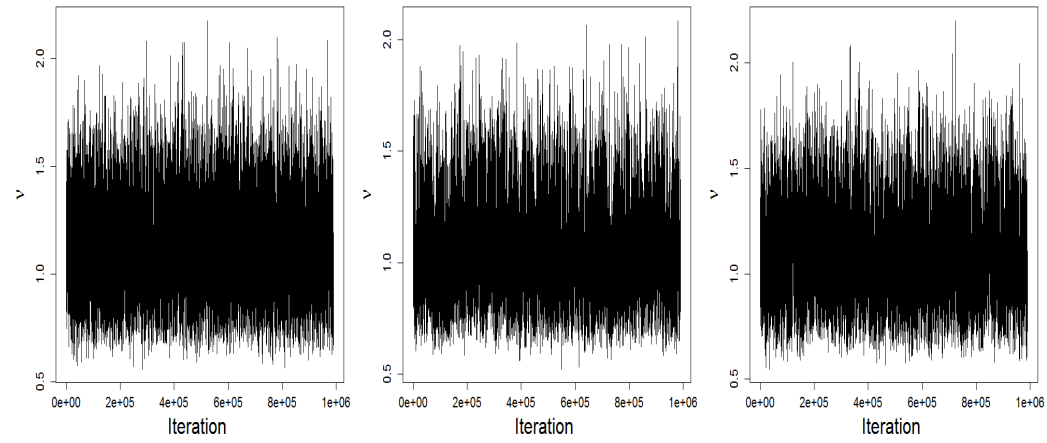
Instead of using a uniform proposal distribution, we can propose an independence sampler which makes use of the likelihood. That is $q(r_j - s_j, r'_j - s_j) \approx Weibull(\lambda, \nu)$. However, since the additional removal times must lie beyond $\max(s_j, t_{obs})$, we



(a) (b) (c)

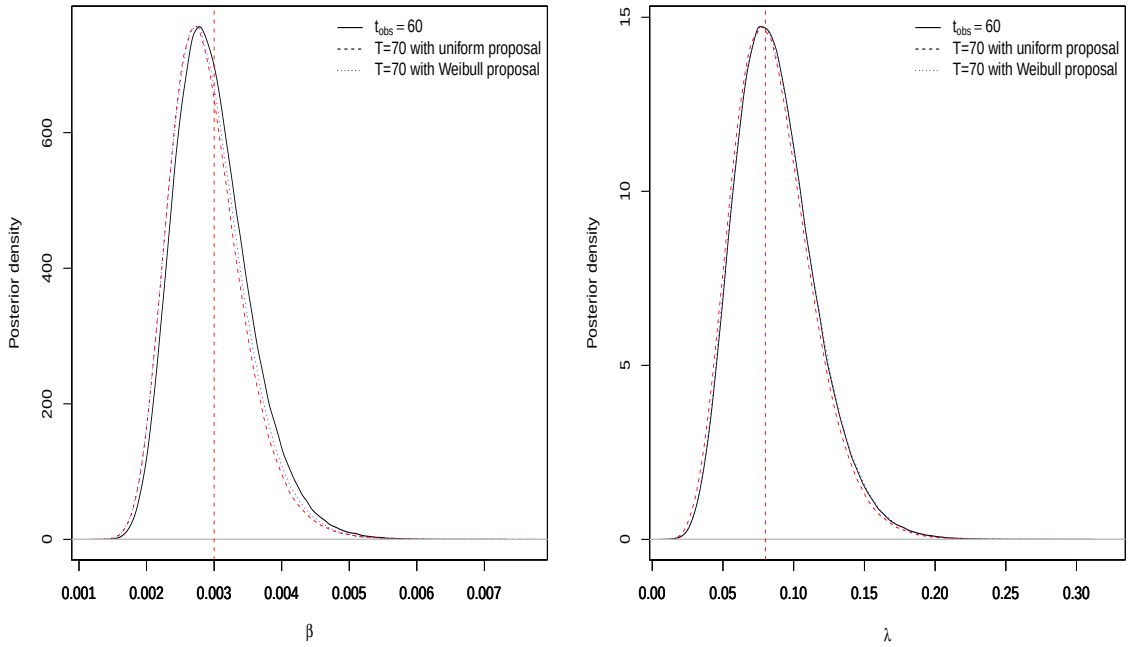


(d) (e) (f)

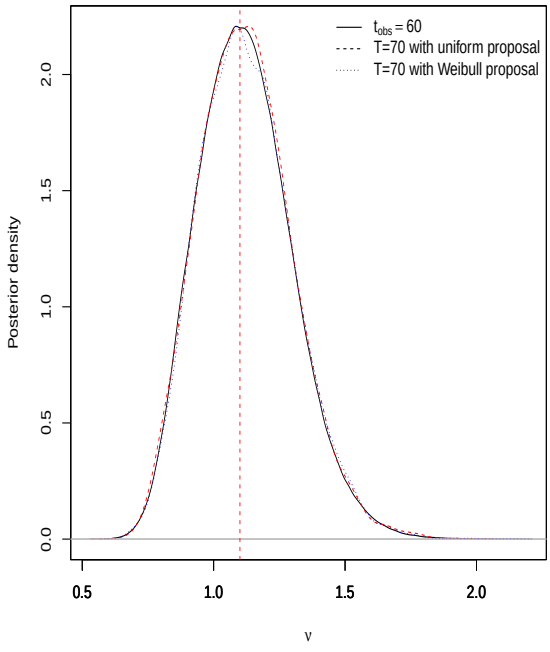


(g) (h) (i)

Figure 3.7: Trace plots of model parameters for the three algorithms. (a), (d) and (g) $t_{obs} = 60$ with uniform proposal; (b), (e) and (h) $T = 70$ with uniform proposal; (c), (f) and (i) $T = 70$ with Weibul proposal



(a) (b)



(c)

Figure 3.8: Posterior distribution of the contact rate β (a), the shape parameter λ (b) and the scale parameter ν (c) for different models: Streftaris-Gibson model (black), removal prediction using uniform proposal (red) and prediction using Weibull distribution (blue).

must truncate its distribution to this interval. In this case, the above algorithm is maintained with the only change occurring in the uniform proposal for the removal replaced by the truncated Weibull distribution. This implies that $\frac{1}{(T-\max(s_j,t_{obs}))}$ is then replaced by $\frac{f_W(r_j-s_j)}{F_W(T-s_j)-F_W(max(s_j,t_{obs})-s_j)}$ where f_W and F_W are respectively the density and the distribution function of the Weibull distribution. We name the algorithm with uniform proposal and the Weibull proposal distribution respectively as *uniform-proposal* and *Weibull-proposal*.

To illustrate the efficiency of our algorithms, we consider the simulated data of the thresholds model described previously and observe the progress till time $t_{obs} = 60$. We impute new infections up to $T = 70$. We run the MCMC routine for 10^6 steps with the first 10000 discarded using both *uniform* and *Weibull* proposal distributions. We then compare the results with the one obtained at t_{obs}. The trace plots in Figure 3.7 show that the chains mixed well and do not show any evidence of non-convergence. From Figure 3.8, we can see that the posterior distributions of the model parameters match in the three cases considered for the model parameters (see figure 3.8). Also, the chain obtained from the model with *Weibull-proposal* seems to mix better than the one with uniform (see Figures 3.7h and 3.7i). This results confirm that the algorithm performs well. Note that we will use this type of prediction later on to compute measures for control.

3.7 Conclusions

This chapter has reviewed the statistical approach mostly used to draw inference from epidemic data especially when the number of infected individual is unknown. The Bayesian techniques presented here, particularly the Reversible-Jump, play a significant role in epidemic modelling given the flexibility it offers when dealing with missing information, which is relevant when designing control strategies.

In chapter 5, we expand these techniques to the spatio-temporal models and models that present structure within the population, and incorporate the Sellke construction in order to provide an efficient control for real time epidemics. We go further in the chapter 7.2 to provide a statistical method in Bayesian framework, treating the transmission network as missing data.

Chapter 4

Coupling non-spatial epidemics using latent processes

4.1 Introduction

The main question that arises for an epidemiologist at the outbreak of an epidemic is which control strategy to adopt in order to prevent future outbreaks. In this chapter, we will consider strategies for controlling the spread of an epidemic using control strategies based on the removal of individuals diagnosed as infected. The question of epidemic control has been crucial over the history of mankind. Several authors raised this question in the context of non-spatial epidemic models and several strategies for their eradication were proposed. For instance, Bootsma and Ferguson (2007); Daley and Gani (1999); Castilho (2006) developed control model based on education campaigns to contain epidemics such as AIDS and Ebola. Anderson (1982); Smith (1964); Becker and Dietz (1995, 1996), on the other hand, consider different vaccination strategies to optimise the control of diseases in human populations. In particular, Ball et al. (1997a) utilised a technique, called an "equalising" strategy, which aims to maintain the number of susceptibles in each group to be similar as possible. Also, Daley and Gani (1999) and WHO (2003) adopted a strategy whereby hosts are screened and put in quarantine if deemed to be at risk.

Here, we assume for simplicity that there is a perfect diagnostic test which, when applied to an individual, can tell us about its current state (e.g. screening test) with 100% accuracy, although this is unrealistic in practice. In other words the test has 100% sensitivity and specificity. Once infected individuals are identified, they can be removed (cured or isolated depending on the disease). Intuitively, the epidemic spread could be retarded due to the fact that susceptibles that would have been infected by those individuals, removed after testing, are "rescued", at least for a certain period of time. Mathematically, the infection pressure exerted on a susceptible individual is reduced, causing them to approach their Sellke threshold (see Section 2.2.2) at a

slower rate.

The benefits of using the latent processes, in this case the Sellke thresholds, is that epidemics can be coupled *via* a common set of thresholds (Andersson and Britton, 2000). The coupling method has found many important applications in several fields of probability theory such as Poisson approximation, renewal processes and Markov processes (Lindvall, 1992). More recently, the technique has been applied successfully to compare retrospectively alternative strategies on historical epidemic data (Cook et al., 2008). With the correct choice of coupling mechanism, the random variables constructed (the outcome of the controls) are highly dependent, leading to high correlation between the outcomes of the controls, reducing the variance of the difference. This surely gives a better estimation of the expected difference between control strategies. The sample size needed to estimate differences in strategies is reduced compared to independent sampling as a consequence of the high correlation between the outcomes of controls. This is analogous to the well-known paired t-test where two population means are compared on the same set of experimental units.

In the remainder of this chapter, we will first describe a general framework for epidemic controls using Sellke thresholds. Conditioned on the individual thresholds and the model parameters, the focus will then be on comparing different control strategies using different epidemic models including the general stochastic epidemic model (SIR), and a non-typical, multi-type SIS model. In our initial investigations we will assume known parameters. When optimising control strategies we will make use of the technique of Simulated Annealing. This concept has been used in the past in epidemiology by Demon et al. (2011) to identify the sample size that maximises the probability of detecting an invasive pathogen. We then show that, the outcomes of controls are highly correlated when we couple realisations using the Sellke thresholds, which reduces the variability in the estimate of the expected differences as suggested above.

4.2 Coupling epidemics subject to alternative control strategies

The description given in this section concerns the SI model (see Equation (2.1.1)). As mentioned in Section 2.2.2, the Sellke construction is one approach to formulating an underlying process for stochastic epidemic models. We denote by $\underline{x}$ the complete set of data (time, nature and affected individuals) for every event describing an epidemic process generated using a standard stochastic model parametrised by $\boldsymbol{\theta}$, assumed known in the rest for this chapter. Then there exists a one-to-one function h such that $\underline{x} = h(\boldsymbol{Q}, \boldsymbol{\theta})$ where $\boldsymbol{Q}$ is the set of Sellke thresholds (Anderson, 1982). Recall

that this non-centred parameterisation is essentially a functional model in the sense of Dawid and Stone (1982). Hence, given $(\boldsymbol{Q}, \boldsymbol{\theta})$ the epidemic trajectory is uniquely defined.

One of the important properties of the Sellke thresholds is that their sample path is not affected by any control strategy which affects only the infection pressure, therefore the epidemic trajectory is uniquely defined for any control strategy. A graphical representation of this property is shown in figure 4.1. The outcome of any control strategy $\boldsymbol{d}$ given $(\boldsymbol{Q}, \boldsymbol{\theta})$ is then uniquely determined as $\underline{x} = h(\boldsymbol{Q}, \boldsymbol{\theta}, \boldsymbol{d})$. The effect of

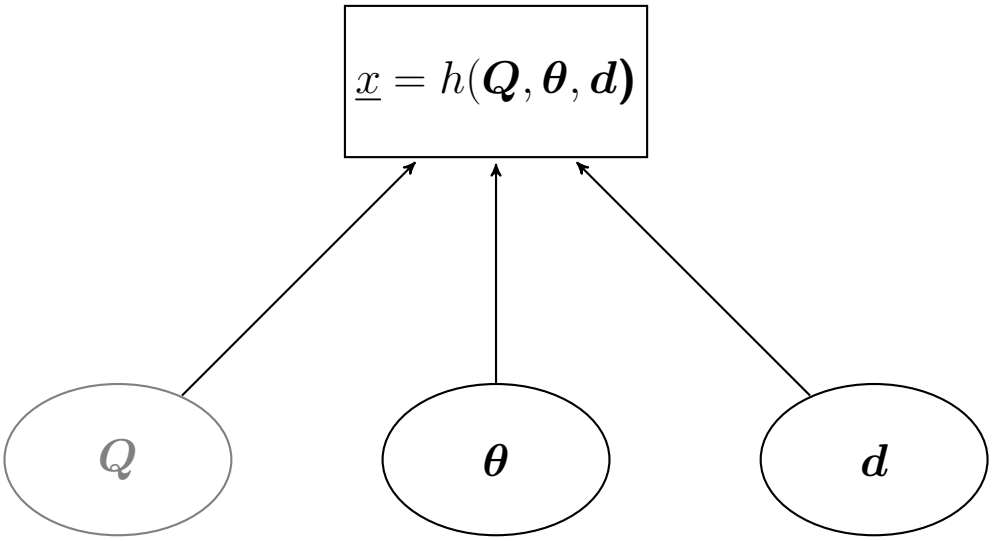


Figure 4.1: The graphical model of the non-centered parameterisation.

different controls can then be compared on a common set of $(\boldsymbol{Q}, \boldsymbol{\theta})$. Specifically, let $\boldsymbol{Q} = (Q_1, \ldots, Q_N)$ define the set of Sellke thresholds of individuals in a population of size N and $\boldsymbol{\theta}$ the vector of model parameters (the contact rate, the infectious period parameter(s), the primary infection rate etc). If we construct two epidemics with different control strategies $\boldsymbol{d}$ and $\boldsymbol{d}'$, only the infection pressure processes $A(t)$ and $A'(t)$ are different for the two processes. We then say that $\boldsymbol{d}$ outperforms $\boldsymbol{d}'$ if the total number of infections $n(\underline{x}(t)) < n(\underline{x}'(t)$ where $\underline{x}(t)$ and $\underline{x}'(t)$ are respectively the complete trajectory of the epidemic process when applying respectively control strategies $\boldsymbol{d}$ and $\boldsymbol{d}'$ on $(\boldsymbol{Q}, \boldsymbol{\theta})$.

4.3 General Stochastic SIR model with control

Throughout this section,

$$\underline{\boldsymbol{Q}} = \begin{pmatrix} Q_{1,1} & Q_{1,2} & \ldots & Q_{1,m} \\ Q_{2,1} & Q_{2,2} & \ldots & Q_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{N,1} & Q_{N,2} & \ldots & Q_{N,m} \end{pmatrix} \text{ and } \underline{\boldsymbol{I}} = \begin{pmatrix} I_{1,1} & I_{1,2} & \ldots & I_{1,m} \\ I_{2,1} & I_{2,2} & \ldots & I_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ I_{N,1} & I_{N,2} & \ldots & I_{N,m} \end{pmatrix} \text{ will denote}$$

respectively m realisations of the set of Sellke thresholds and the corresponding infectious period of the respective individuals in a population of size N. Therefore, $(\boldsymbol{Q}^i, \boldsymbol{I}^i, \boldsymbol{\theta})$ where $\boldsymbol{Q}^i = (Q_{1i}, Q_{2i}, \ldots, Q_{Ni})$ and $\boldsymbol{I}^i = (I_{1i}, I_{2i}, \ldots, I_{Ni})$, denotes a single replicate and $(\underline{\boldsymbol{Q}}, \underline{\boldsymbol{I}}, \boldsymbol{\theta})$ denotes a set of m replicates of the epidemic process with fixed $\boldsymbol{\theta}$. In this section, we design controls and identify the optimal one using the most commonly studied stochastic model, the general stochastic epidemic SIR model described in Section 1.3.4. We consider controls based on deploying diagnostic tests at specified times and immediately removing detected infections. This is analogous to the cost-effective model for the quarantine policy developed by Gani et al. (1997) to reduce the HIV rate in prison. In this chapter, hosts in the population are assumed to mix homogeneously.

4.3.1 Testing protocol

Intuitively the tests should be administered so that individuals detected for testing are most likely to be infectious at the time of the test. Here, we propose to test individuals cyclically. More precisely, if t corresponds to a time when a test carried out on individual i reveals that it is still susceptible, then it remains in the population and can be tested again, say at time $t' > t$. Meanwhile it has to wait until the remaining individuals are tested. This is sensible given that the individual that has waited the longest has the greatest chance of being infected, given that it has received the greatest accumulated challenge since its last test. Table 4.1 shows the representation of how the tests are applied.

i		i+1		i+2		i+3
i+1		⋮		⋮		⋮
⋮	$\Longrightarrow$	N	$\Longrightarrow$	N	$\Longrightarrow$	N
N	susceptible	1	susceptible	1	infective	1
1		⋮		⋮		⋮
⋮		i-1		i		i-1
i-1		i		i+1		i

Table 4.1: Schematic of the test procedure. Labels in gray correspond to the individual tested. If the individual is infected then he is removed but takes the last position in the list when he still susceptible.

4.3.2 Evaluation of the function $h(\theta, Q, I, d)$

In reality, to embed a control strategy (or testing regime in this case) into an epidemic process generated through Sellke construction, we first generate realisations of the Sellke thresholds and the infectious period for each host. Then, starting from $t = 0$, we

compute a potential infection times for all hosts and identify the next infection time. The next event is then chosen to be either a test, an infection, or a removal depending on which event type occurs first. The disease status is then changed according to the event and the process continues until no further event can occur or until a defined stopping criterion is reached. The algorithm is obtained by embedding the process for the diagnostic tests into that described in Section 2.2.2.

Calculation of $\underline{x} = h(\theta, Q, I, d)$ for SIR model with primary and secondary infection.

Notation

- $d = (\tau_1, ..., \tau_{N'})$ say, the (ordered) times of the tests to be applied.
- $I = (I_1, ..., I_N)$ the infectious periods for the N individuals in the population (all initially susceptible).
- $\boldsymbol{Q} = (Q_1, ..., Q_N)$, the Sellke thresholds for the individuals.
- $\boldsymbol{\theta} = (\delta, \beta)$, the primary and secondary infection rates.

Calculation

- Set $t = 0$, number of infections $n_I = 0$, infection rate $\lambda(t) = \alpha + \beta n_I$. Set initial state of each individual, $s_i = 0, i = 1, ..., N$, and time till next transition $r_i = Q_i/\lambda(t)$. Set final time T.
- While $t < T$ do the following {

  Identify next event: Let $t^* = \min\{r_1, ..., r_N, \tau_1\}$ and $j = \arg\min\{r_1, ..., r_N, \tau_1\}$.

  If $t^* = r_j$ (event is an infection or a removal), implement event.

  - case $s_j = 0$ (infection event), set $s_j = 1$, $n_I = n_I + 1$, $r_j = I_j$.
  - case $s_j = 1$ (removal event), $n_I = n_I - 1$, delete row j from tableau and move lower rows up, set $N = N - 1$.

  If $t^* = \tau_1$ {

  (event is test) delete τ_1 and shift remaining times along.

  If $s_1 = 1$, delete row 1 from tableau and move lower rows up, setting $N = N - 1$.

  If $s_1 = 0$, move row 1 to row N and move the others up

}

Finally update remaining Sellke thresholds and waiting times in the tableau, and test times.

- $R' = \lambda(t)$, $t = t + t^*$, $\lambda(t) = \alpha + \beta n_I$.
- If $s_i = 0$, $Q_i = Q_i - R't^*$, $r_i = Q_i/\lambda(t)$.
  If $s_i = 1$, $r_i = r_i - t^*$,
- Update remaining time till scheduled tests: $\tau_i = \tau_i - t^*$. }

4.3.3 Design construction

We assume that the tests are deployed in non-intersecting time windows $[a_j, b_j]$, $j = 1, \ldots, l$. It is assumed that there are n_j test in $[a_j, b_j]$ deployed at times separated by δt_j so that $b_j = a_j + (n_j - 1)\delta t_j$, $j = 1, \ldots, l$. We specify a control strategy $\boldsymbol{d}$ as $\boldsymbol{d} = \{(a_1, n_1, \delta t_1), \ldots, (a_l, n_l, \delta t_l)\}$.

4.3.4 Simulation study using a single realisation of Sellke thresholds

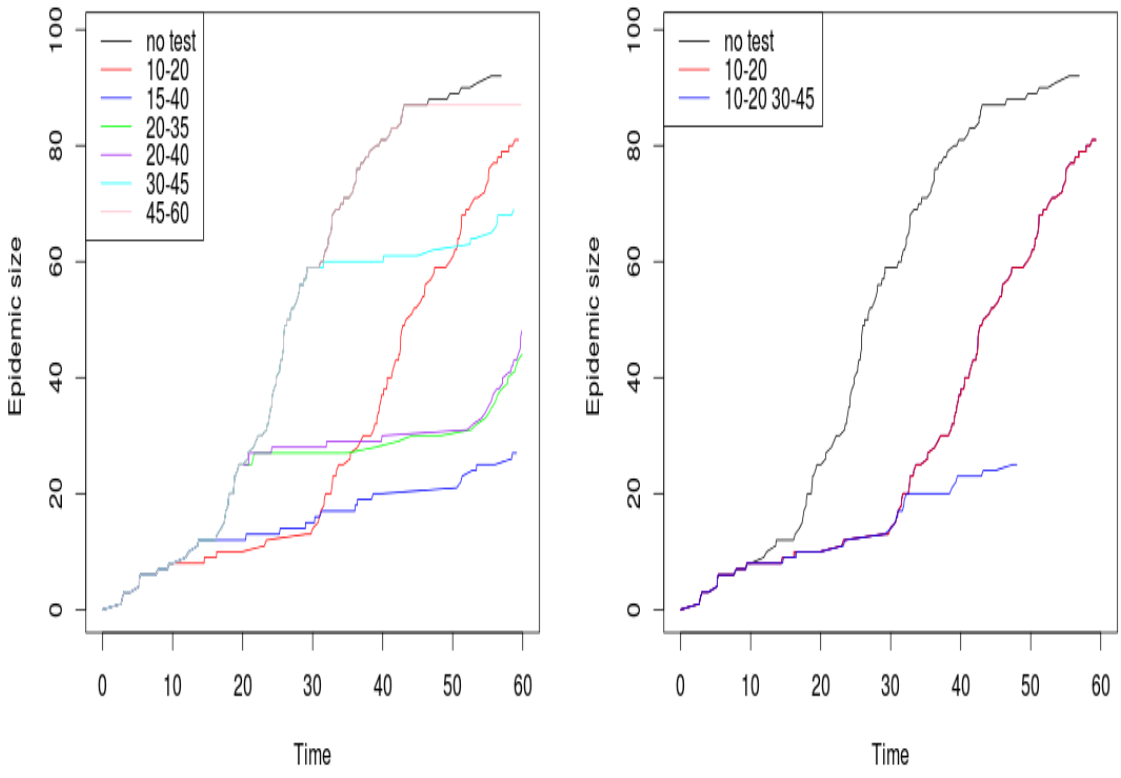
To illustrate our approach to designing control strategies based on Sellke thresholds using diagnostic tests, we implement the method described above. We consider a population of size $N = 100$ with initial number of susceptibles, infectives and removals respectively $S(0) = 100$, $I(0) = 0$ and $R(0) = 0$. We assume the presence of a primary infection source with a constant pressure over the observation time. Thus for the simulation, we fixed the primary infection rate to $\delta = 0.002$, the contact rate $\beta = 0.003$, and the parameters of the Weibull infectious period $\nu = 1.1$, $\lambda = 9.94$. We then assign to each individual a single realisation of a Sellke threshold and an infectious period which is Weibull distributed as described in Section 4.3.2. We observe the epidemic evolution up to a time $T = 60$. In addition, we assume that there are $N' = 150$ available tests. Recall that an individual is automatically removed if a test reveals it to be infectious but remains in the process otherwise. In the latter case, such an individual could be tested again once it has again reached the top of the queue as in Table 4.1.

Initially, we consider various control strategies $d_0, d_1, \ldots, d_6$ where the tests are applied in a single interval. Table 4.2 shows the values used for the controls. Figure 4.2a depicts the trajectory of epidemics generated by a common set of Sellke thresholds, subjected to control strategies that differ in term of timing as shown in Table 4.2. As

	Control strategy
d_0	No test
d_1	$\{(10, 150, 0.067)\}$
d_2	$\{(15, 150, 0.1)\}$
d_3	$\{(20, 150, 0.1)\}$
d_4	$\{(20, 150, 0.134)\}$
d_5	$\{(30, 150, 0.1)\}$
d_6	$\{(45, 150, 0.1)\}$

Table 4.2: Tests deployed in single intervals.

expected, the trajectories coincide with the test free trajectory until the deployment of the tests, after which they decrease in comparison to the test-free trajectory. This result is reasonable since the total pressure on susceptibles decreases as soon as an infective is removed.



(a) (b)

Figure 4.2: Trajectories of an SIR epidemic process with Weibull infectious period where $S(0) = 100$, $I(0) = 0$, $R(0) = 0$, $\beta = 0.003$, $\nu = 1.1$, $\lambda = 9.94$, $T_{max} = 60$, primary infection rate $\delta = 0.002$, when applying different test regimes (including no test) and $N' = 150$ diagnostic tests in : (a) single intervals of time (b) single and multiple intervals.

These results are generated assuming all tests are deployed in a single interval. A more general approach involves splitting the available tests over more than one interval. We compare both cases by repeating the above simulation (coupling epidemics) using the same Sellke thresholds and the same infectious period as previously but now splitting the tests over two intervals. For the illustrative purpose, we consider the control strategy $d^* = \{(10, 100, 0.1), (35, 50, 0.3)\}$. Figure 4.2b indicates how important is the reduction in the latter case when comparing with the single interval case for the design d_1. On the Figure 4.2b, it can be seen that the reduction is about 10 when applying the control using the design d_1 whilst this reduction is 65 when the tests are split over two intervals using the strategy d^*. As a result, splitting the tests may be more cost effective than concentrating all in a single interval.

4.3.5 Optimal control using Simulated Annealing (SA)

When applying tests to reduce the final size of an epidemic at some finite time, the reduction can be thought of as a function of the control strategy $\boldsymbol{d}$ (time interval, number of tests, etc.), with the optimal strategy giving the greatest expected reduction. However, the space of possible control strategies may be complicated. Hence an appropriate optimization method such as Simulated Annealing could be useful for finding an optimal control. We will discuss this method in the next section.

4.3.5.1 Motivation

Denote by $\boldsymbol{d} \in \mathcal{D}$ a control design. We define the impact function $u(\underline{x}(T))$ characterising the number of infections, where $\underline{x}(T)$ denotes the trajectory of the epidemic up to T. The function u is obtained numerically by applying the method described in Section4.3.2. The choice of $\boldsymbol{d}$ that gives the best reduction is of concern here. Therefore we define the optimal strategy as the one that minimises the expectation of the impact function. We can state the design problem as

$$\boldsymbol{d}^* = \arg\min_{\boldsymbol{d}\in\mathcal{D}}(U(\boldsymbol{d})) \tag{4.3.1}$$

where,

$$U(\boldsymbol{d}) = E_{\boldsymbol{d}}\left(u(\underline{x}(T))\right) = \int u(\underline{x}'(T))\pi_{\boldsymbol{d}}(\underline{x}'(T))d\underline{x}'(T) \tag{4.3.2}$$

Note that $\pi_{\boldsymbol{d}}(\underline{x}(T))$ is the distribution of the process trajectory when a control $\boldsymbol{d}$ is applied. We omit $\boldsymbol{d}$ to signify no control is applied. In addition, we write $\underline{x}(\boldsymbol{Q}, \boldsymbol{I}, \boldsymbol{\theta}, \boldsymbol{d}, T)$ to denote a draw from $\pi_{\boldsymbol{d}}(\underline{x}(T))$ which is the trajectory of the process generated by $(\boldsymbol{Q}, \boldsymbol{I}, \boldsymbol{\theta})$ and observed until an "assessment time" T ($\boldsymbol{d}$ is applied before T, depending on the intervals).

The integral in $U(\boldsymbol{d})$ could then be approximated using the Monte Carlo approximation in Section 3.5.1 as follows:

$$\hat{U}(\boldsymbol{d}) = \frac{1}{p}\sum_{i=1}^{p} u(\underline{x}(\boldsymbol{\theta}^i, \boldsymbol{Q}^i, \boldsymbol{I}^i, \boldsymbol{d}, T)) \tag{4.3.3}$$

and the optimal design $\boldsymbol{d}^*$ could then be determined. This will involve computing differences of the form $\hat{U}(\boldsymbol{d}') - \hat{U}(\boldsymbol{d})$ for different designs $\boldsymbol{d}, \boldsymbol{d}'$. In causal inference, this quantity is known as the average causal effect of $\boldsymbol{d}'$ compare to $\boldsymbol{d}$. Moreover, contrary to most of the literature on causal inference (Holland, 1986), which focusses on applying control on each individual considered as an experimental unit, we are interested in the effect of the control on the whole population. Hence, the so-called *unit* in causal inference is represented here as the whole population under study.

In the rest of this section, we denote by

$$n(\boldsymbol{\theta}, \underline{\boldsymbol{Q}}, \underline{\boldsymbol{I}}, \boldsymbol{d}, , T) = (u(\underline{x}(\boldsymbol{\theta}^1, \boldsymbol{Q}^1, \boldsymbol{I}^1, \boldsymbol{d}, T)), \ldots, u(\underline{x}(\boldsymbol{\theta}^m, \boldsymbol{Q}^m, \boldsymbol{I}^m, \boldsymbol{d}, T))) \tag{4.3.4}$$

the epidemic size obtained by applying the control strategy $\boldsymbol{d}$ on each simulated realisation observed up to time T. Note that for a single realisation i.e. $m = 1$, we maintain the notation $(\boldsymbol{Q}, \boldsymbol{I}, \boldsymbol{\theta})$ to denote a realisation of the epidemic process. We define by

$$r(\boldsymbol{\theta}, \boldsymbol{Q}, \boldsymbol{I}, \boldsymbol{d}, T) = u(\underline{x}(\boldsymbol{\theta}, \boldsymbol{Q}, \boldsymbol{I}, T)) - u(\underline{x}(\boldsymbol{\theta}, \boldsymbol{Q}, \boldsymbol{I}, \boldsymbol{d}, T)) \tag{4.3.5}$$

the reduction in the epidemic size of a control $\boldsymbol{d}$ compared to the no-control case.

4.3.5.2 Simulated Annealing

Simulated Annealing (SA) is an optimization technique originally used in the field of combinatorial optimization. Due to its simplicity and efficiency, it has been extensively applied in the field of heuristic search (Kirkpatrick et al., 1983). SA is a stochastic algorithm that was inspired by the analogy between the physical annealing process and the search for a minimum in more complex systems. Moreover it has been proved that it has the ability to escape from local optima and reach the global optimum (Kirkpatrick et al., 1983; Corana et al., 1987; Rajasekaran, 1990; Bertsimas and Tsitsiklis, 1998; Busetti, 2003). Here, the goal is to find the design that maximises the objective function f i.e. the strategy that satisfies the relation

$$\arg\max_{d} f(d)$$

To that end, the algorithm employs, essentially, a random search procedure which not only accepts changes that increase the objective function f, but also may accept changes that decrease it. In this way, moves are accepted according to the Metropolis

criterion. The probability of accepting a move is

$$\alpha = \min\left\{1, \exp\left(-\frac{\Delta f}{T_p}\right)\right\}; \tag{4.3.6}$$

where Δf is the decrease in f and T_p is the control parameter which is known as the *system temperature*. In our context, Δf represents the expected reduction in the outcome of the controls or the average causal effect analogous to Holland (1986).

4.3.6 Application of SA in epidemic control using Sellke thresholds

Commonly used in the field of heuristic search, SA is less common in epidemiology. It has been recently applied to epidemics by Demon et al. (2011) to identify the sample size that maximises the probability of detecting an invasive pathogen. Here, we wish to identify the optimal strategy that gives the highest reduction in the outcome of the epidemic, the size of infected population by a certain assessment time.

4.3.6.1 Optimal control strategy for a single realisation of the epidemic process

Assume that, the trajectory of the epidemic without control is known precisely. In this case, the process can be represented with only one realisation of the Sellke thresholds and the infectious periods. The optimal strategy using the SA is then obtained as follows:

Algorithm

Denote by $\boldsymbol{\theta}$ the model parameters. Let $(\boldsymbol{Q}, \boldsymbol{I})$ be the Sellke thresholds and the infectious periods for population members.

Algorithm 1. SA for a single realisation of the epidemic process

1: **Initialisation**

Set the cooling temperature c_1, $k = 1$ and the initial design $\boldsymbol{d}^{opt} = \boldsymbol{d}^1$

Compute the approximation in Equation (4.3.3) for $p = 1$

$\hat{U}^{opt} = \hat{U}^1 = \hat{U}(\boldsymbol{d}^1) = u(\underline{x}(\boldsymbol{\theta}, \boldsymbol{Q}, \boldsymbol{I}, \boldsymbol{d^1}, T))$.

2: Update to $\boldsymbol{d}$ (see application below).

3: **while** stopping criteria is not reached **do**

4: **for** $l = 1, \ldots, q$ **do**

5: Evaluate $\hat{U}' = \hat{U}(\boldsymbol{d}') = u(\underline{x}(\boldsymbol{\theta}, \boldsymbol{Q}, \boldsymbol{I}, \boldsymbol{d}', T))$

6: Set $(\hat{U}, \boldsymbol{d}) = \begin{cases} (\hat{U}', \boldsymbol{d}') & \text{if } \hat{U}' < \hat{U}^k \\ (\hat{U}', \boldsymbol{d}') & \text{if } \hat{U}' > \hat{U}^k \text{ with probability } \alpha \end{cases}$

where $\alpha = \min\left\{1, \exp\left(-\frac{\hat{U}' - \hat{U}^k}{c_k}\right)\right\}$

Set $(\hat{U}^{opt}, \boldsymbol{d}^{opt}) = (\hat{U}', \boldsymbol{d}')$ if $\hat{U}' < \hat{U}^{opt}$

7: **end for**

8: Set $(\hat{U}^{k+1}, \boldsymbol{d}^{k+1}) = (\hat{U}, \boldsymbol{d})$

9: Reduce the parameter c_k such that $c_{k+1} = \gamma c_k$, $\gamma \in [0, 1]$

10: $k \leftarrow k + 1$

11: **end while**

Application

To illustrate the approach described above, we consider the strategy defined in Section 4.2. We reduce the space of the control by considering only two intervals and by assuming that $\delta t_1 = \delta t_2 = \delta t$. We then use the SA algorithm 1 to update the control strategy $d = \{(T_1, n_1, \delta t), (T_2, n_2, \delta t)\}$. Here, $a_1 = T_1$ and $a_2 = T_2$. For simplicity, we fix $\delta t = 0.1$. Consequently, we only update n_1 (ultimately n_2). Specifically, a new design is proposed by performing a random walk on the current value of n_1 and the new value of n_2 is deduced.

Again, as in Section 4.3.4, we consider a population of size $N = 100$ with initial susceptibles, infectives and removals considered respectively to be $S(0) = 100$, $I(0) = 0$ and $R(0) = 0$. We assume that the epidemic is introduced by an external source with a constant pressure $\delta = 0.002$ over the observation period. We fixed the contact parameter $\beta = 0.003$ and set the parameters of the Weibull infectious period to be $\nu = 1.1$, $\lambda = 9.94$. We then assign to each individual one threshold and an infectious period. Once again we observe the epidemic evolution up to a time $T = 60$.

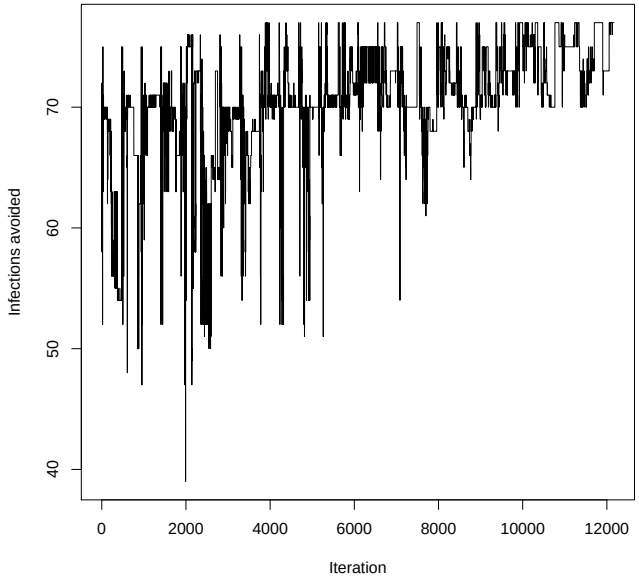


Figure 4.3: Progress of Simulated Annealing on the epidemic reduction $r(\boldsymbol{\theta}, \boldsymbol{Q}, \boldsymbol{I}, \boldsymbol{d}, T)$.

We run the algorithm for 12000 iterations taking $\gamma = 0.85$. A new value of the design parameter is proposed from a uniform distribution g centered on the current value. Since $n_1 + n_2 = N'$, we only update n_1 and derive $n_2 = N' - n_1$. Figure 4.3 shows the progress of SA algorithm on the epidemic size reduction $r(\boldsymbol{\theta}, \boldsymbol{Q}, \boldsymbol{I}, \boldsymbol{d}, T)$ for the current design against iteration number. It can be seen that initially, some large decrease in the epidemic size reductions were accepted due to the fact that the initial temperature is high. As the temperature decreases, fewer downhill moves are tolerated, as a result the later iterations maintain the system close to the optimum. The optimal control strategy resulting from this algorithm is given by $\boldsymbol{d}^* = \{(11.172, 85, 0.1), (33.487, 65, 0.1)\}$ i.e. $T_1 = 11.172$, $T_2 = 33.487$, $n_1 = 85$, $n_2 = 65$. We can see how efficient the SA is in reducing the epidemic size. This reduction can be seen as the causal effect of the optimal control on the epidemic at the population level (Holland, 1986).

4.3.6.2 Optimal design for a multi-replicate of the epidemic process

Given the model parameters, the primary infection rate, if such a source is present, the contact rate and the Weibull parameters $\delta, \beta, (\nu, \lambda)$ respectively, by simulating only one realisation of the epidemic process from these parameters, we only capture a particular scenario that one can have from those specific parameters. The inherent stochasticity in the process is then ignored. Therefore a multi-replicate experiment (large number of simulated realisations) needs to be investigated in order to have a good representation of the variability of the epidemic.

The algorithm of 1 identifies the optimal control for a single realisation of the process. For a multi-replicate of the process (multiple realisations of the Sellke thresholds with given parameter distribution), we rather find the optimal control strategy by adaptively increasing the value p in the function $\hat{U}(\boldsymbol{d})$ in Equation (4.3.3). The advantage of adopting such a strategy is that one can identify the minimum sample size required for the algorithm to find the optimal control effectively.

Method

We start by finding the optimal design of $u(\underline{x}(\boldsymbol{\theta^1}, \boldsymbol{Q^1}, \boldsymbol{I^1}, \boldsymbol{d}))$ using a single realisation of the process, say replicate 1. Whenever the optimisation is done, we introduce in a new realisation at a time. The new optimal is obtained by minimising $\sum_{i=1}^{n} u(\underline{x}(\boldsymbol{\theta}^i, \boldsymbol{Q}^i, \boldsymbol{I}^i, \boldsymbol{d}, T))$ where n is the current number of replicates included in the sum. For efficiency, we specify that the initial design for the optimisation with n replicates should be the optimum obtained using $n-1$ replicates.

The algorithm is described as follows:

Algorithm

We start by generating $(\boldsymbol{\theta}^i, \boldsymbol{Q}^i, \boldsymbol{I}^i)$ for $i = 1, \ldots, p$ from the set of replicates.

Algorithm 2. Optimal design for a multi-replicate of the epidemic process

1: **Initialisation**
Set the cooling temperature c_{11}, $p = 1$ and the initial design $\boldsymbol{d}^{opt} = \boldsymbol{d}^1$
Compute the approximation in Equation (4.3.3) $\hat{U}^{opt} = \hat{U}^1 = \hat{U}(\boldsymbol{d}^1) = u(\underline{x}(\boldsymbol{\theta}^1, \boldsymbol{Q}^1, \boldsymbol{d}^1))$.
2: Update to $\boldsymbol{d}$ (see below).
3: **while** stopping criteria is not reached **do**
4: **for** $l = 1, \ldots, m$ **do**
5: **for** $k = 1, \ldots, q$ **do**
6: Use the same p samples and evaluate
$\hat{U}' = \hat{U}(\boldsymbol{d}') = \frac{1}{p}\sum_{i=1}^{p} u(\boldsymbol{\theta}^i, \boldsymbol{Q}^i, \boldsymbol{d}')$
7: Set $(\hat{U}, \boldsymbol{d}) = \begin{cases} (\hat{U}', \boldsymbol{d}') & \text{if } \hat{U}' < \hat{U}^k \\ (\hat{U}', \boldsymbol{d}') & \text{if } \hat{U}' > \hat{U}^k \text{ with probability } \alpha \end{cases}$
where $\alpha = \min\left\{1, \exp\left(-\frac{\hat{U}' - \hat{U}^k}{c_{pk}}\right)\right\}$
Set $(\hat{U}^{opt}, \boldsymbol{d}^{opt}) = (\hat{U}', \boldsymbol{d}')$ if $\hat{U}' < \hat{U}^{opt}$
8: **end for**
9: Reduce the parameter c_{pk} such that $c_{p(k+1)} = \gamma c_{pk}$, $\gamma \in [0, 1]$
10: **end for**
11: $p \leftarrow p + 1$, $\boldsymbol{d} \leftarrow \boldsymbol{d}^p = \boldsymbol{d}^{opt}$, $U \leftarrow U(\boldsymbol{d}^p)$
12: Set c_{p1}
13: **end while**

Application

To illustrate this procedure, we consider the same parameters for the epidemic model used previously. Recall that we consider a population of size $N = 100$ and set the primary infection rate $\delta = 0.002$, the contact rate $\beta = 0.003$ and the Weibull infectious period parameters as $\nu = 1.1$, $\lambda = 9.94$. We assume that there are $N' = 150$ tests. Again, we reduce the design space by partitioning the tests in two intervals where δt is kept fixed and takes the value 0.1. Specifically, we consider the design of the form $\boldsymbol{d} = \{(T_1, n_1, \delta t), (T_2, n_2, \delta t)\}$ where the parameters T_1, T_2, n_1 and n_2 are of interest.

Similar to the single replicate case, we propose a new value of each parameter in the design by adopting a random walk using an uniform proposal distribution g. The algorithm is run for 425000 iterations and the convergence is assessed. The progress of the cumulative reduction $\sum_{i=1}^{n} r(\boldsymbol{\theta}^i, \boldsymbol{Q}^i, \boldsymbol{I}^i, d, T)$, $n = 1, \ldots, 50$ for various control strategies $\boldsymbol{d}$ proposed in the course of the algorithm is assessed and shown in Figure

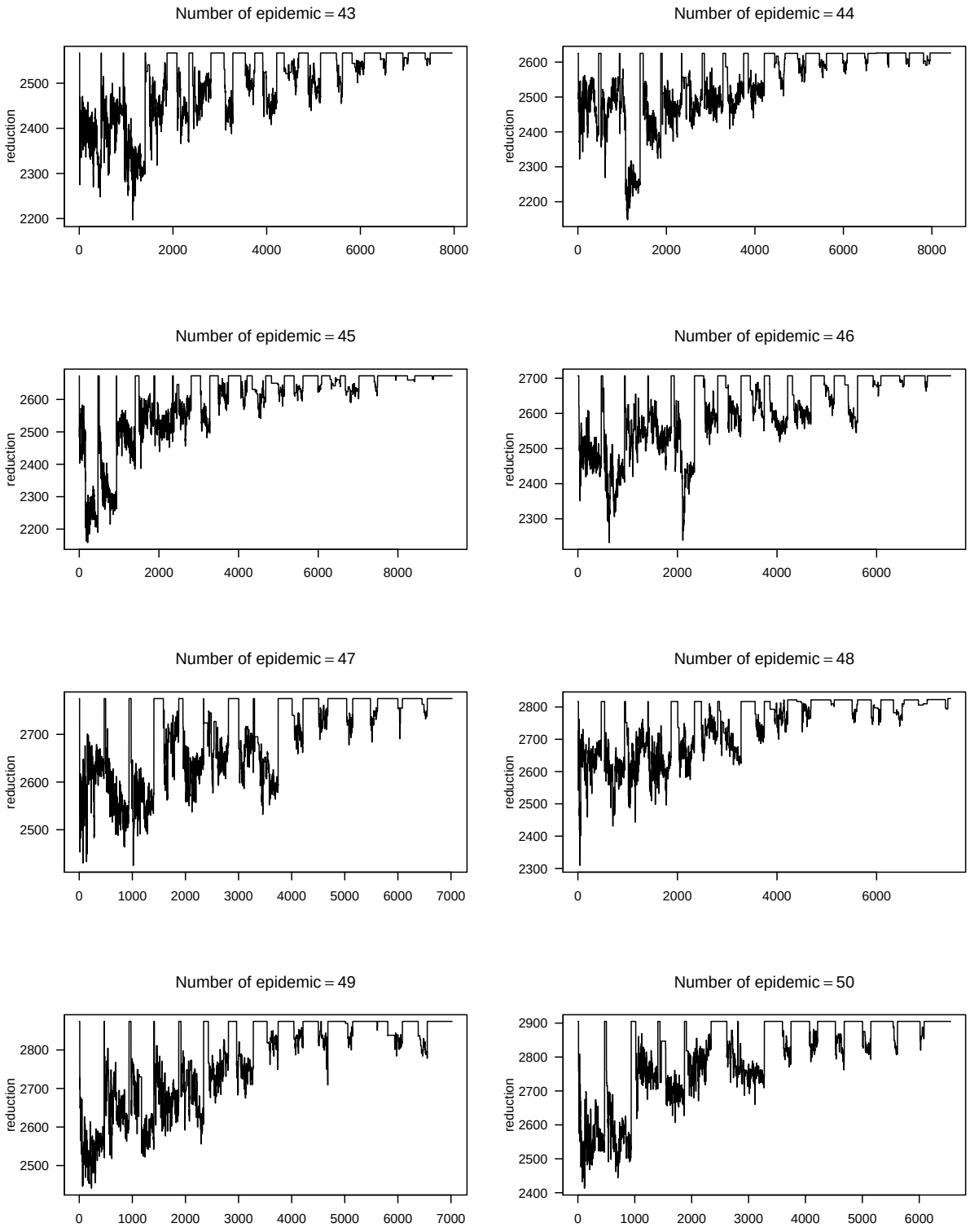


Figure 4.4: The progress of the cumulative reduction $\sum_{i=1}^{n} u(\theta^i, Q^i, I^i, d_0) - u(\theta^i, Q^i, I^i, d)$ for $n = 43, 44, \ldots, 50$ replicates of the epidemic process where d is any design proposed and accepted during the Simulated annealing algorithm.

4.4. We only present the result for $43 - 50$ replicates. We can see that from the 43^{th} replicate, there is no clear difference in the reduction obtained from the initial

strategy and the optimal one; in other words the algorithm does not find an improved design. We can conclude that the algorithm converges towards an optimum in the neighbourhood of $\boldsymbol{d}^* = \{(9.216, 83, 0.1), (32.172, 67, 0.1)\}$.

Next, we fixed $T_2 = 32.172$ and we vary the parameters n_1 and T_1. Figure 4.5 shows the contour plot of n_1 against T_1. It can be seen from Figure 4.6 that the domain explored by the algorithm represents the high density regions. This shows indeed the effectiveness of the algorithm in term of convergence towards the optimal strategy.

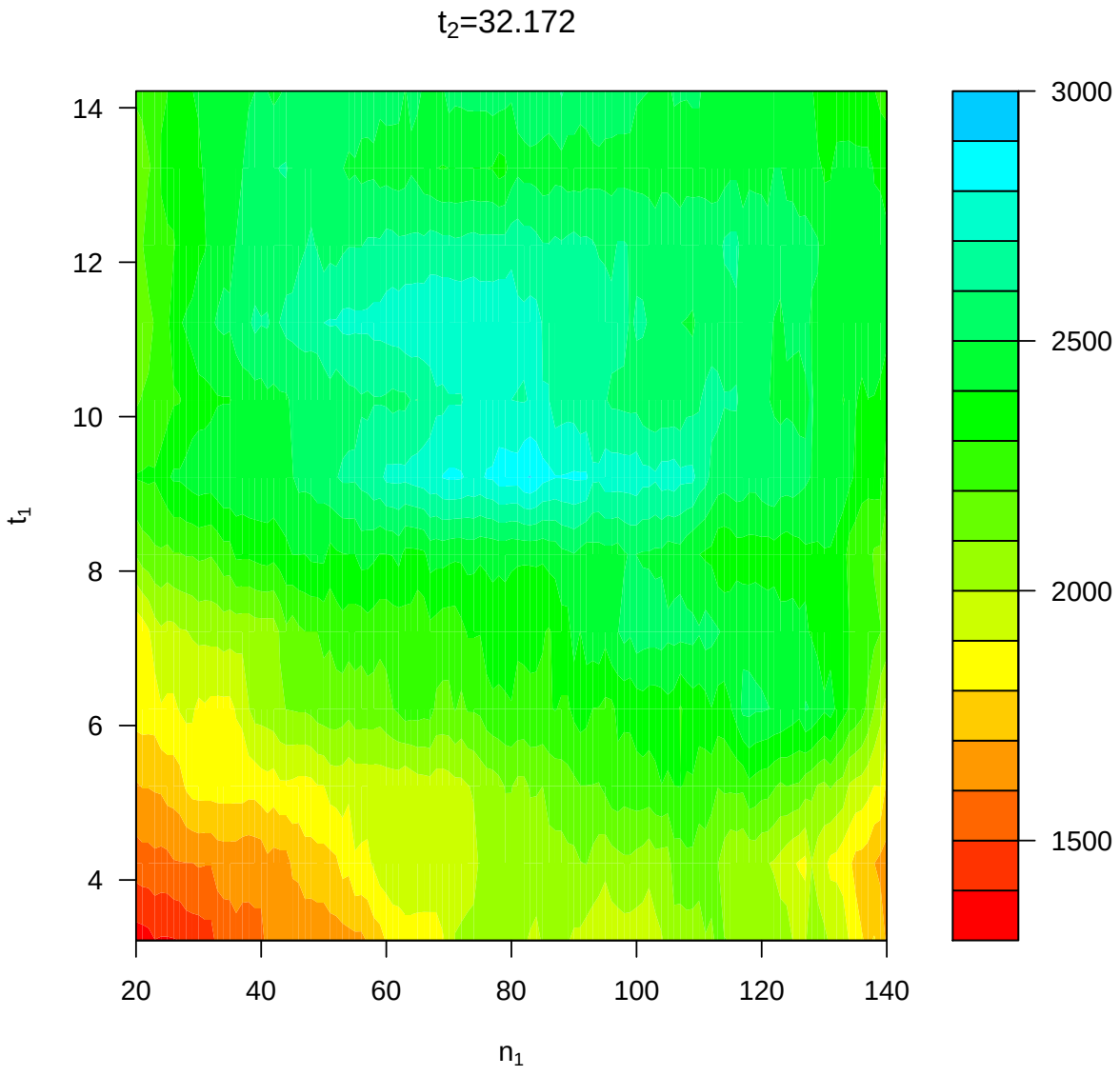


Figure 4.5: Contour plot of $\hat{U}$ over (n_1, t_1) when fixed $t_2 = 32.172$.

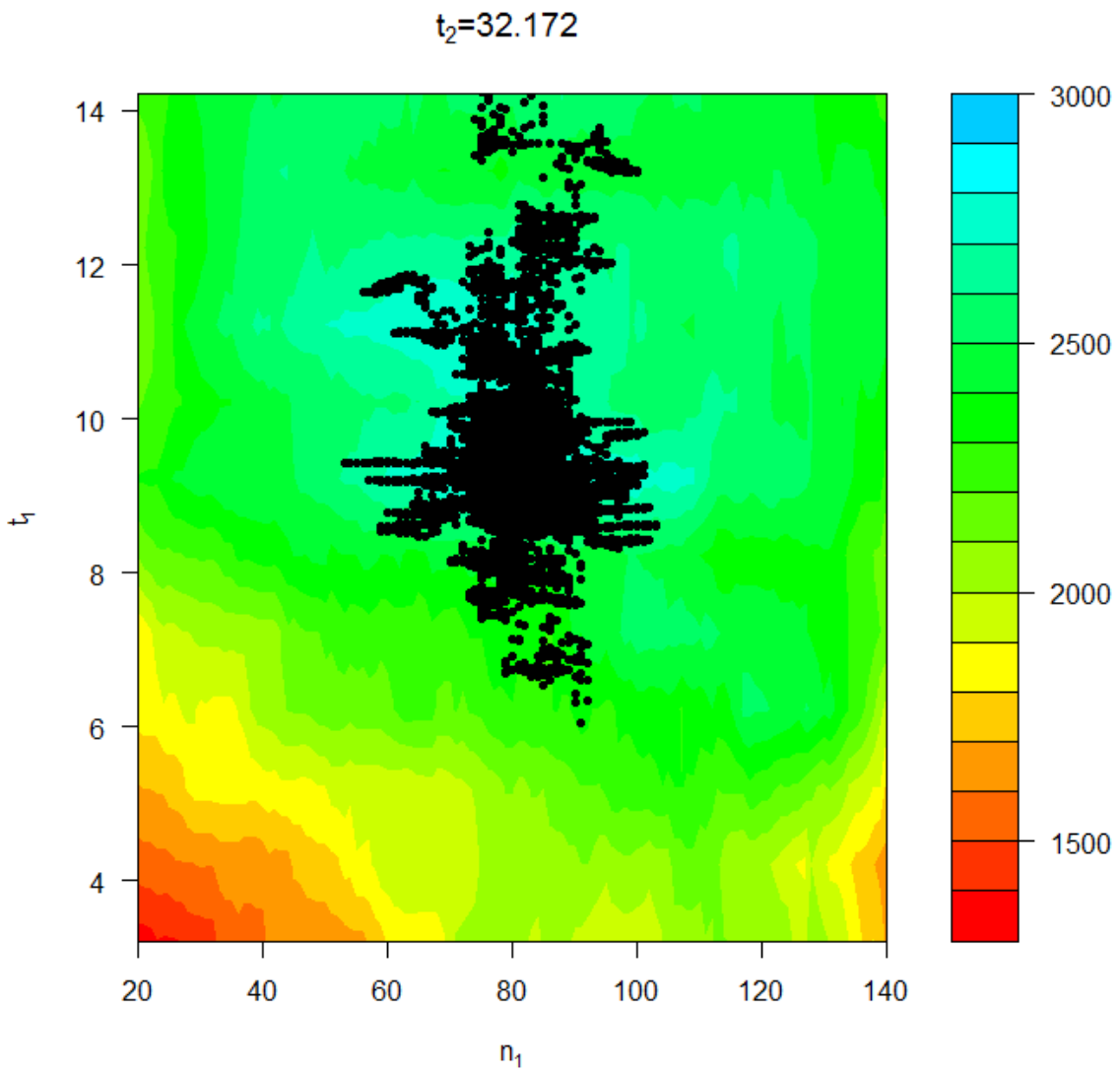


Figure 4.6: Contour plot of $\hat{U}$ over (n_1, t_1) when fixed $t_2 = 32.172$ showing the area exploired during the latter part of the algorithm $(31875 - 42500)$.

Finally we investigate the sample correlation $Corr(n\left(\boldsymbol{\theta}, \underline{Q}, \underline{I}, \boldsymbol{d}^*, T\right), n(\boldsymbol{\theta}, \underline{Q}, \underline{I}, \boldsymbol{d}, T)\big)$ for various control strategies $\boldsymbol{d}$ with T_2 fixed to the optimum (see Figure 4.7). It can be seen that there is a strong positive correlation between the epidemic size obtained from the p replicates of the process subject to the optimal control $\boldsymbol{d}^*$ $(n\left(\boldsymbol{\theta}, \underline{Q}, \underline{I}, \boldsymbol{d}^*, T\right))$, and the one of various control strategies $\boldsymbol{d}$ $(n\left(\boldsymbol{\theta}, \underline{Q}, \underline{I}, \boldsymbol{d}, T\right))$. This observation implies that the variance of the difference between the reduction obtained by the optimal control and any sub-optimal control $Var(n\left(\boldsymbol{\theta}, \underline{Q}, \underline{I}, \boldsymbol{d}^*, T\right) - n(\boldsymbol{\theta}, \underline{Q}, \underline{I}, \boldsymbol{d}, T))$ is then reduced compared to the independent sampling case. Subsequently, this improves estimation of the expected difference between these strategies. This is analogous to the paired t-test where two treatments are applied to the same set of experimental units.

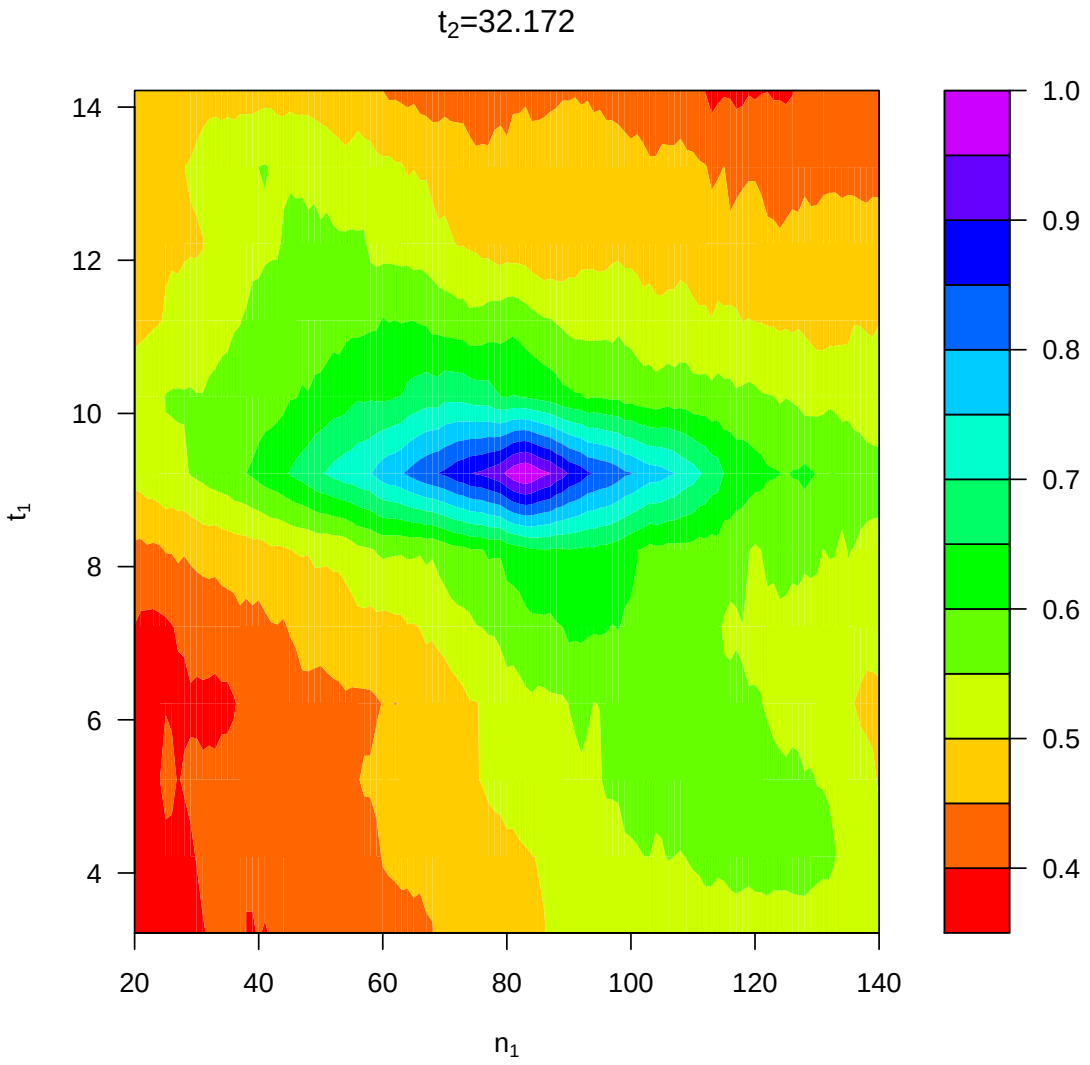


Figure 4.7: Figure showing the correlation between different strategies with the optimal strategy

4.3.7 The multi-type SIS models

4.3.7.1 Model description

Traditionally, intervention during an outbreak is considered to provide the most reliable evidence in epidemiological research. In this section, we give an example to illustrate how targeted control using latent processes (Sellke thresholds) might offer benefits when considering a non-typical SIS model.

We consider a population of fixed size N partitioned into k sub-groups where, initially, each group contains exactly N_i susceptibles $i = 1, \ldots, k$. It follows that $N = \sum_{i=1}^{k} N_i$. Moreover, each individual may be in one of the two states, either susceptible (S) or infected (I). We assume that the epidemic is essentially caused by

an external source to the population. In other words there is no risk of transmission between individuals. In addition, the model assumes that the epidemic is non-fatal to the individuals. As soon as a susceptible is infected, it moves to the infective class. On being diagnosed infected, the infected individual moves back to the S class, either due to treatment or replacement. We again assume that our test has perfect sensitivity and specificity so that no false positives or false negatives can arise. Model of the sojourn time in the I state for each individual is determined *via* the testing process.

Recall that each individual possesses a level of tolerance to the disease known as Sellke thresholds which is exponentially distributed ($Q_j \sim Exp(1)$, $j = 1, 2, \ldots, N$). In this model an individual is assigned a new threshold wherever it recovers from the disease. In other words once an individual is diagnosed positive, a new threshold is then assigned as an independent draw from $Exp(1)$. Infection occurs when the total pressure exerted on an individual up to a certain time reaches its threshold $\int_0^{t_j} \alpha\beta_{g(i)}dt = Q_j$. The process then follows the following probability transition: In an infinitesimal interval $[t, t + dt]$ the transition from state S to the state I is given by

$$P[i \text{ infected in } (t, t + dt)] = \alpha\beta_{g(i)}dt + o(dt) \tag{4.3.7}$$

where α represents the rate of the external source of the epidemic (primary infection rate), $\beta_{g(i)}$ stands for the i^{th} group susceptibility.

4.3.7.2 Modelling Control strategies using Sellke thresholds

As mentioned previously, the transition from the class I to S occurs through a diagnostic test. Whenever individuals are tested positive (I), they immediately reset to the S class. Numerically, this is done by assigning a new Sellke threshold. However, in practice, it is almost impossible to test all individuals in the population (e.g. in a large population) at the time due to the limitation of the available resources. It is therefore important to find a prioritisation scheme in order to reduce the impact of the outbreak.

Different approaches have been proposed in the literature for infectious epidemics. For example Becker and Dietz (1995, 1996) discussed four particular strategies. They are:

1. Households chosen at random and all their members are immunised.
2. Random vaccination of individuals.
3. Preferential selection of large households for immunisation.
4. Strategies for households of equal size (imunisation of a fixed fraction of members in every household).

These strategies are compared by identifying the optimal proportion of the population needed to be vaccinated in order to reduce below 1 the basic reproduction number. Here we wish to minimize the time individuals remain infected for.

A- **Proposed strategy**

We propose to prioritise individuals with respect to the expectation of the time they have been infected. Assume that an individual was last tested at time T_1. If the time of the next test is T, the expected infection time for which the individual has been infected, τ, is given by:

$$\tau = \int_{T_1}^{T} (T-t) \frac{\alpha\beta_i \exp(-\alpha\beta_i t)}{\exp(-\alpha\beta_i T_1) - \exp(-\alpha\beta_i T)} dt. \tag{4.3.8}$$

where t is the time of the infection period, and $T-t$ the length of time that the individual is infected for, up to T. It follows that

$$\tau = T - \frac{1}{(\exp(-\alpha\beta_i T_1) - \exp(-\alpha\beta_i T))\,\alpha\beta_i} \left(F(T;2,\alpha\beta_i) - F(T_1;2,\alpha\beta_i)\right), \tag{4.3.9}$$

where $F(T;2,\alpha\beta_i)$ represents the CDF of the Gamma distribution with shape 2 and rate $\alpha\beta_i$. Doing some algebraic manipulation, Equation (4.3.9) leads to:

$$\tau = T - \frac{1}{\alpha\beta_i} + \frac{T\exp(-\alpha\beta_i T) - T_1 \exp(-\alpha\beta_i T_1)}{\exp(-\alpha\beta_i T_1) - \exp(-\alpha\beta_i T)}. \tag{4.3.10}$$

Allocating resources (tests in this case) using this strategy is sensible since we expect to sample individuals that have been infected for longer period.

Alternative strategies: We assess the performance of three other strategies. In particular, we consider

B- **Probability or high risk based prioritisation of individuals**

This sampling method involves sampling the N' individuals that have the largest probability of being infected. Recall that under this process, each individual in the population independently has a probability

$$P(t) = 1 - \exp(-\alpha\beta_i t) \tag{4.3.11}$$

of being infected since the time of the last test, where t is the time elapsed since the last test.

C- **Simple random scheme**

Simple random sampling involves randomly selecting N' individual to test.

D- **Ranking based prioritisation**

For ranking based prioritisation, individuals are randomly label from 1 to N and then the tests are carried out according to that ordering; once an individual is tested, it reverts to the bottom of the queue (see Table 4.1).

Method:

To generate a realised epidemic with this type of control using Sellke's construction, we proceed as follows:

Notation

- $d = (\tau_1, ..., \tau_{N'})$ say, the (ordered) times of the tests to be applied.
- $\underline{Q} = \begin{pmatrix} Q_{1,1} & Q_{1,2} & \dots & Q_{1,m} \\ Q_{2,1} & Q_{2,2} & \dots & Q_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{N,1} & Q_{N,2} & \dots & Q_{N,m} \end{pmatrix}$ where each row j corresponds to the individual j's sequence of Sellke thresholds.
- $\boldsymbol{\theta} = (\alpha, \beta)$, the primary rate and the vector of host's group susceptibility.
- E to store infected hosts identified by the tests.

Calculation

- Set $t = 0$, infection rate of each individual $R_i = \alpha\beta_{g(i)}$. Set initial state of each individual, $s_i = 0, i = 1, ..., N$, $Q_i = Q_{i,1}$ and time till next transition $r_i = Q_i/R_i$. Set final time T.
- While $t < T$ do the following {

  Identify next event: Let $t^* = \min\{r_1, ..., r_N, \tau_1\}$ and $j = \arg\min\{r_1, ..., r_N, \tau_1\}$.

  If $t^* = r_j$ (event is an infection), implement event.

  - case $s_j = 0$ (infection event), set $s_j = 1$.

  If $t^* = \tau_1$ {

  (event is test) delete τ_1 and shift remaining times along.

  If $s_1 = 1$, move row 1 to row N and lower rows up and set $s_i = 0$. Store individual in E

If $s_1 = 0$, move row 1 to row N and move the others up.

}

Finally update remaining Sellke thresholds and waiting times in the tableau, and test times.

- $t = t + t^*$.
- If $s_i = 0$, $i \notin E$, $Q_i = Q_i - R_i t^*$, $r_i = Q_i / R_i$.
  If $s_i = 0$, $i \in E$, assign the next value in $\underline{Q}$ to Q_i.
  If $s_i = 1$, $r_i = r_i - t^*$,
- Update remaining time till scheduled tests: $\tau_i = \tau_i - t^*$.
- Initialise E. }

To compare the strategies, we consider the evolution of the cumulative time individuals are infected for; more formally,

$$\Omega(t) = \sum_{j=1}^{N} \sum_{i=1}^{n_j} (T_{(j_i)} - t_j^i) \tag{4.3.12}$$

where $\left\{t_j^i, \forall i = 1, \ldots, n_j, \forall j = 1, \ldots, N\right\}$ and $\left\{T_{(j_i)}, \forall i = 1, \ldots, n_j, \forall j = 1, \ldots, N\right\}$ are respectively the set of times at which individuals are infected during the observation period and their corresponding time at which they are tested positive. It is worth noting that, these times are readily available from the above algorithm.

4.3.7.3 Simulation study

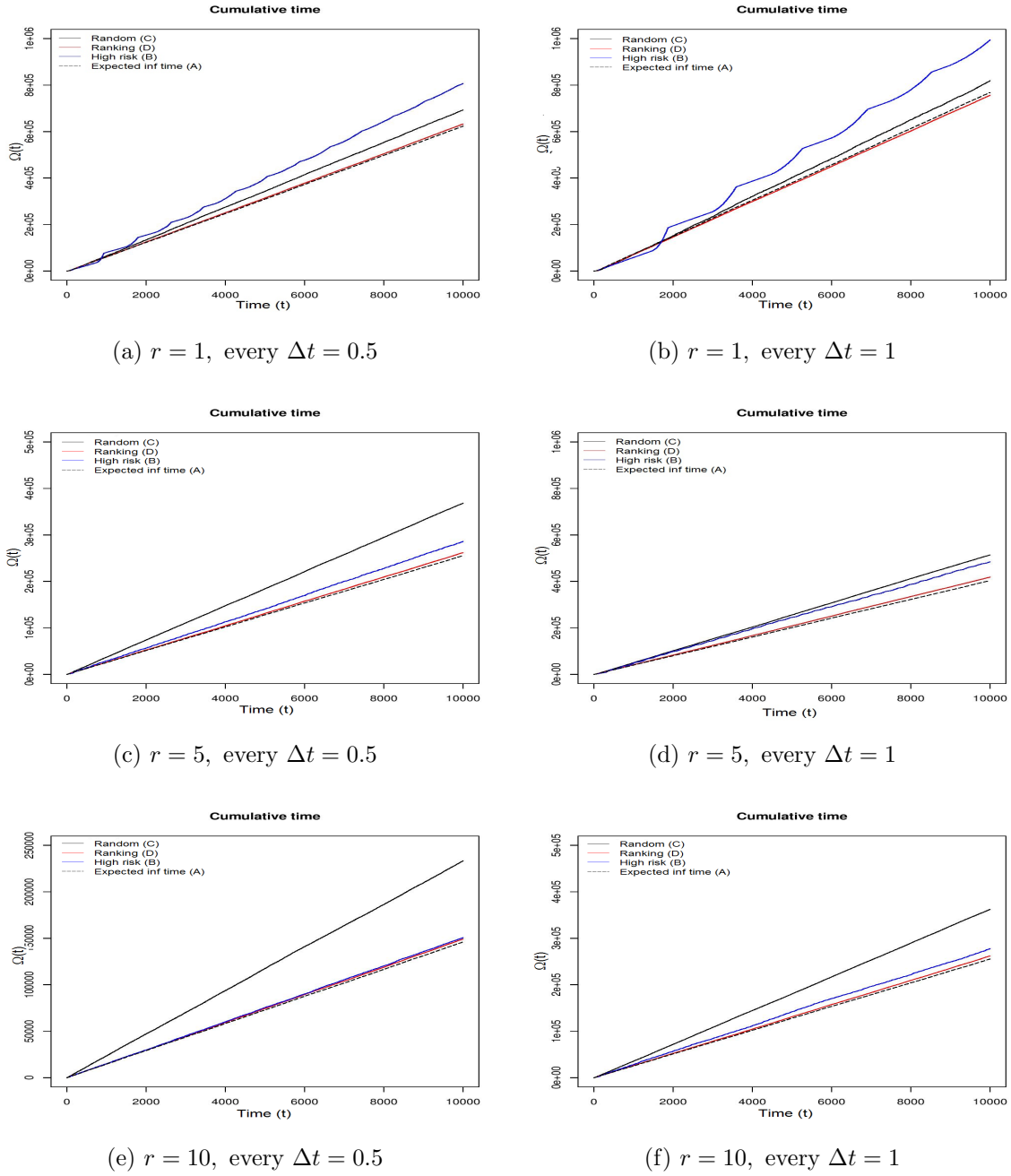
To illustrate the choice of the expected infection time as the measure of ordering the individuals for the tests, we consider an outbreak with fixed parameters occurring in a population of size $N = 125$ subdivided into two groups ($k = 2$) with initial size of susceptibles $N_1 = 75$ and $N_2 = 50$ in each group respectively. Since the susceptibility varies with respect to the group, we set $\beta_1 = 0.07$ and $\beta_2 = 0.0035$ for group 1 and 2 respectively and we set the rate of external source of infection to be $\alpha = 1$. We assume that $T = 10000$ and that the tests are applied at $T_1, T_2, \ldots, T_{N'}$ where $T_i = t_0 + i\Delta t$, $\forall i = 1, \ldots, N'$ and $T_{N'} \leq T$. We simulate m replicates of Sellke threshold for each individual where m corresponds to the maximum number of possible tests one can carry out on an individual. At every T_i, we assume that it is possible to sample r individuals for testing.

Here we use the values $r = 1, 5$ and $\Delta t = 0.5, 1$. In practice, an epidemic has to take off before we take any action, therefore we start deploying the controls from time

$t_0 = 10$ where few individuals present some signs of infection.

4.3.7.4 Results

Figure 4.8 shows the cumulative time Ω (see Equation 4.3.12) using the strategies A, B, C and D defined previously. It can be seen that as time goes on, the strategy A gives



(a) $r = 1$, every $\Delta t = 0.5$

(b) $r = 1$, every $\Delta t = 1$

(c) $r = 5$, every $\Delta t = 0.5$

(d) $r = 5$, every $\Delta t = 1$

(e) $r = 10$, every $\Delta t = 0.5$

(f) $r = 10$, every $\Delta t = 1$

Figure 4.8: Plot showing the evolution of the cumulative time infection during the process over time, with four different measures used for different test regime ($r = 1$ and $r = 5$ each $\Delta t = 0.5$ and $\Delta t = 5$).

.

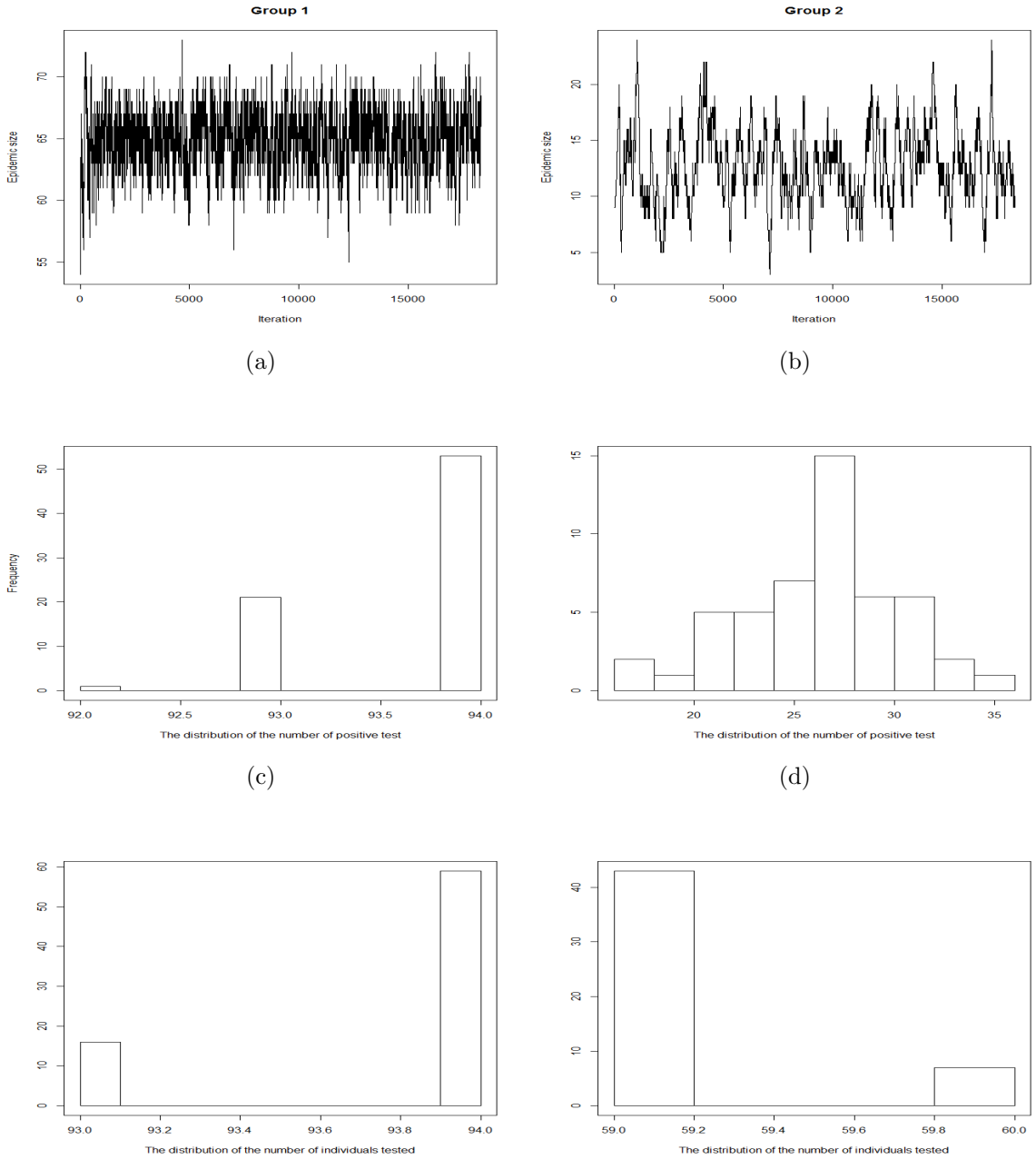


Figure 4.9: Plot showing the epidemic size, the number of times indivuduals are tested positive and the number of times an individual is tested during the epidemic process with parameters as mentioned above, in first, second and third rows respctively where $r = 1$ and $\Delta t = 1$.

.

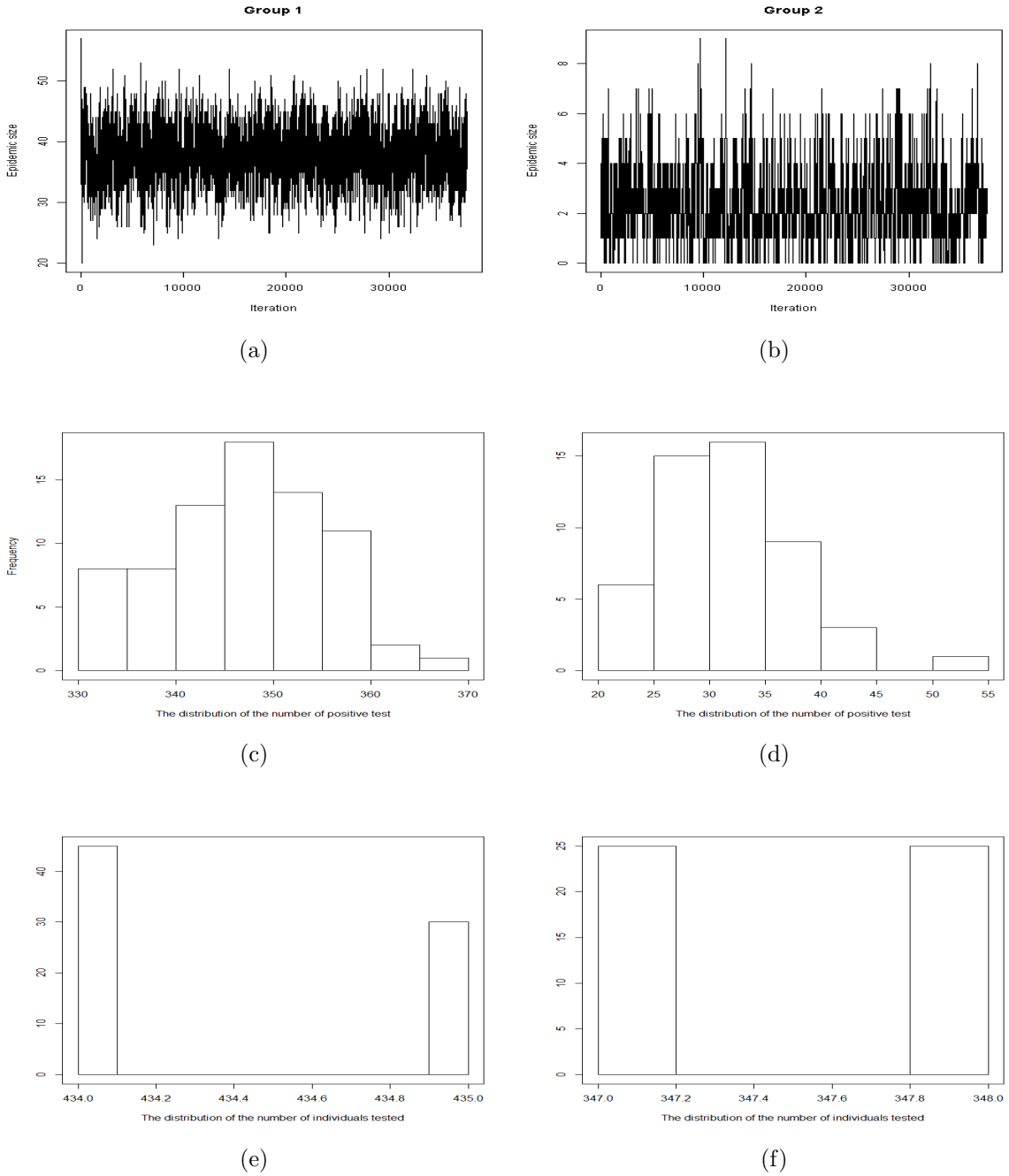


Figure 4.10: Plot showing the epidemic size, the number of times individuals are tested positive and the number of times an individual is tested during the epidemic process with parameters as mentionned above, in first, second and third rows respctively where $r = 5$ and $\Delta t = 1$.

.

better reduction in terms of the time individuals are infected for (shown by the dashed line), even though there is little difference with the strategy D. Hence, the strategy based on prioritising individuals to test using their expected time since infection is beneficial in comparison to the other approaches, when the objective function is Ω (see Equation 4.3.12). However, as we increase the number of individuals r to sample

for the test, the effectiveness of strategy B approaches that of strategy A as shown in Figure 4.8. Figures 4.9 and 4.10 show the epidemic size, the distribution of the positive tests and the distribution of the number of tests in each group only for the case $\Delta t = 1$ where $r = 1$ and $r = 5$ respectively using the optimal control (expectation infection time based). As expected, Figures 4.9a, 4.9b, 4.10a and 4.10b show that the epidemic never dies out; instead, the process seems to have reached its equilibrium when the final time considered is $T = 10000$ and the number of infectives fluctuates around 65 for the group 1 and 13 for the group 2 in the case $r = 1$.

Unsurprisingly, we note that there are more individuals tested in group 1 compare to group 2; as shown in Figures 4.10f and 4.10e, 4.9f and 4.9e. This is due to the fact that the group of individuals highly susceptible to the disease is more likely to be infected more often after their recovery. This is emphasized in Figures 4.10d and 4.10c, 4.9d and 4.9c where the positive tests are more frequent in group 1. Hence for a better control, more resources must be deployed to the more susceptible group.

4.4 Conclusions

This chapter has introduced the idea of using latent process (Sellke thresholds) to compare different control strategies for epidemics in the presence of diagnostic tests when there is no data observed and the model parameters are assumed to be known. The approach leads to a coupling of realisations under different controls to compare competing control strategies in term of their expected outcome (epidemic size or total time hosts remain infected for). It also gives evidence that under this coupling, the outcome of different controls are highly positively correlated which, from a statistical perspective reduces the variance of the estimates of the expected difference. We showed that given a set of Sellke thresholds, an optimal control strategy could be determined using a general numerical optimisation method. Although Simulated Annealing is relatively new to epidemic modelling (Demon et al., 2011), it is widely used in geostatistics (van Groenigen, 1999; Marchant and Lark, 2007).

The current study has only examined controls on coupling epidemics with the assumption that the model parameters are known beforehand. No inference of parameters from data is carried out. These considerations are not realistic for real-life epidemics since most often controls are conditioned on observed data, i.e. the dynamics of the disease in the targeted population are estimated. In this case, the distribution of the model parameters is determined using a statistical approach, frequently the Bayesian inference described in chapter 3. We will discuss these aspects in the next chapter.

Chapter 5

Controlling spatio-temporal epidemics using latent processes in the Bayesian framework

5.1 Introduction

5.1.1 Motivation

Highly infectious diseases such as citrus canker (Gottwald et al., 2002; Gottwald and Irey, 2007; Neri et al., 2014), Huanglongbing (Bové, 2006; Gottwald, 2010; Parry et al., 2014), Chalara dieback (DEFRA, 2013), Food and Mouth disease (Ferguson et al., 2001; Tildesley et al., 2009) and classical swine fever traditionally have been and continue to be a source of major global and regional threat in agricultural and animal systems, causing important economical loss (Schubert et al., 2001; Ferguson et al., 2001; Gottwald et al., 2002; Thompson et al., 2004; DEFRA, 2013). In recent years there has been an increasing interest in controlling these severe diseases (Schubert et al., 2001; USDA/APHIS et al., 2006; Parnell et al., 2009; Bassanezi et al., 2011; DEFRA, 2013; Cunniffe et al., 2014). Efforts to control an outbreak of such diseases often involve pre-emptive culling of a large number of healthy hosts around a detected infected host. This kind of control has been always surrounded by controversy given the socio-economic impact it can have on farm owners (Schubert et al., 2001; Ferguson et al., 2001; Graham et al., 2004). This highlights the need of strategies which optimize the distribution of the available resources and minimize the number of susceptible hosts, that must be removed.

Here we assume that we observe some epidemic data $\mathbf{y}$, typically host removal times or snapshots of the system at certain observation times in a time interval $[0, t_{obs}]$. The latter form of data is the result of a sequence of surveys identifying the set of symptomatic hosts at times $\tau_1 = t_0, \tau_2, \ldots, \tau_n = t_{obs}$. Given that interventions will

typically occur at a future time $t_c > t_{obs}$ and that the object of the control is to minimise some aspect of the epidemic (which may be the total number of infections-removals prior to some future or an assessment time), the fundamental question that arises during an outbreak of this kind is the following: *What is the optimal control at t_c so that the specified impact, of the epidemic is minimised given the observation* $\mathbf{y}$? (see Figure 5.1).

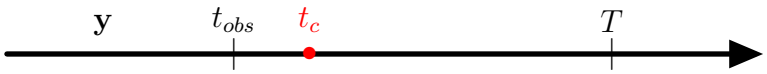


Figure 5.1: Graphical representation of the observation-control-impact system.

5.1.2 Methodology

To answer this fundamental question, we consider a range of measures that can be used to characterise each host at a future time and to prioritise those hosts which shall be targetted for control. A similar idea is used in Boender et al. (2007) where the map is constructed based on the basic reproduction number. However, a control based on such a map may reduce the epidemic impact but will involve culling healthy hosts, especially in high-density regions. In Tildesley et al. (2009) and Kao (2003), a similar approach is used to prioritise farms with high probability of infection. However, the combination of both approaches has been proved to outperform either of these two approaches (te Beest et al. (2011)). Our approach follows a similar idea, though in contrast to the model in te Beest et al. (2011), our model accounts for cryptic infections. We define the so-called *threat* for each host which takes into account the predicted risk of infection (probability of being infected at a given time) and the infection hazard posed to the other susceptibles if it becomes a focus of a new local epidemic. To reflect the constraints on deploying resources which might limit capability to inspect an arbitrary subset of hosts, we will consider scenarios where a host population is partitioned into sub-regions and controls are deployed in a subset of these. We will again use the notion of coupling epidemics using Sellke thresholds to reduce the variance of estimates of differences in control efficacy.

Finding the optimal design involves comparing the predicted effect of different control strategies on the trajectory of the epidemic. In practice, however, it is impossible to observe the effect of two control strategies on the same population unit during an outbreak. This is known as the Fundamental Problem of Causal inference (Holland, 1986). However, techniques such as randomised, cohort and case-control studies have been used to tackle this issue in medical sciences. Also techniques based on comparing controls on independent realisations obtained using the Gillespie algorithm (Gillespie, 1977) have been used by Keeling et al. (2001); Ripley et al. (2003)

and Cunniffe et al. (2015). These aforementioned techniques may increase the bias in the expected effect between controls compared to the case where experimental units are paired. They may also involve the use of a large number of realisations. Here, we follow the same mechanism as in Cook et al. (2008) where epidemics are coupled by matching latent processes assumed to be unaffected by control, to circumvent this problem. Not only does this mechanism allow us to compare directly the effect of different controls but it also provides a reduction in the variance of estimates of the difference between controls. Following Cook et al. (2008), we use the Sellke thresholds as the latent processes whose sampling properties are known given the model specification. We illustrate the methodology on various data sets, including a simulated outbreak in an uniformly distributed locations, and using citrus locations from urban Miami (Gottwald et al., 2002; Cook et al., 2008; Neri et al., 2014; Cunniffe et al., 2015).

5.2 Modelling approaches

As mentioned in the previous section, consideration is given to identifying a more general modelling approach that could be appropriate for controlling spatio-temporal epidemics, particularly arboreal epidemics. The goal is to be able to design efficient controls to reduce future impact of the epidemic taking account of the nature of the available data and the size of the disposal resources.

5.2.1 Structure of the control

Epidemics such as citrus canker, FMD, Classical and African Swine Fever, etc, are often observed in a period of time $[t_0, t_{obs}]$ before any action is taken. We denote by $\mathbf{y}$ the data observed during that period which typically consist of a sequence of 'snapshots' of the symptomatic set of hosts at discrete times. We assume the epidemic proceeds according to some model with parameter vector $\boldsymbol{\theta}$. We define the trajectory of the epidemic up to time t to be $\underline{x}(t)$ so that $\underline{x}(t)$ specifies the time and nature of every transition occurring during $[t_0, t]$. The intervention time when the control is applied is denoted by $t_c > t_{obs}$ and we denote by $T_a \geq t_c$ the 'outcome or assessment time' at which the effectiveness is assessed (e.g. in term of the numbers of infections up to T_a). Similar to Chapter 4, we define the impact function $u(\underline{x}(T))$ as the total number of infections by T.

We denote by $\pi_0(\underline{x}(T)|\mathbf{y})$ and $\pi_{\boldsymbol{d}}(\underline{x}(T)|\mathbf{y})$ the posterior distribution of the trajectory of the epidemic subject to no control and control $\boldsymbol{d}$ respectively. The best choice of control $\boldsymbol{d}$, given the data $\mathbf{y}$ is the one that minimises the expected impact

conditional on what is observed. That is,

$$U(\boldsymbol{d}) = E_{\boldsymbol{d}}\left(u(\underline{x}(T))|\mathbf{y}\right) = \int u(\underline{x}'(T))\pi_{\boldsymbol{d}}\left(\underline{x}'(T)|\mathbf{y}\right) d\underline{x}'(T) \tag{5.2.1}$$

The most common approach used in simulation-based optimal design utilises Monte Carlo Simulation (see Chapter 3). Typically, we draw sample $(\underline{x}(T), \boldsymbol{\theta})$ from $\pi_d\left(\underline{x}(T), \boldsymbol{\theta}|\mathbf{y}\right)$ to generate a sample from $\pi_{\boldsymbol{d}}\left(\underline{x}(T)|\mathbf{y}\right)$. Therefore, for different strategies, we use these samples to evaluate the impacts and evaluate expectations. The expectation of the effect of these control strategies could then be compared to identify the optimal one. However, the posterior distribution $\pi_{\boldsymbol{d}}\left(\underline{x}(T), \boldsymbol{\theta}|\mathbf{y}\right)$ is not known in our context. One approach would be to develop a fully Bayesian approach as in Müller (1999) where the design $\boldsymbol{d}$ is estimated through an MCMC scheme using a Metropolis-Hastings technique for updating $\boldsymbol{d}$. However, by doing so we are comparing the effect of strategies on independent epidemics, leading to higher variance between the outcomes.

5.2.2 Representation of the epidemic process

We will again make use of a functional-model representation of the process introduced in 4.3.1. Recall that there is a process $\boldsymbol{Q}$ whose sampling distribution is independent of the model parameter vector $\boldsymbol{\theta}$ such that the outcome of the epidemic is obtained deterministically as

$$\underline{x} = h(\boldsymbol{Q}, \boldsymbol{\theta}) \tag{5.2.2}$$

where h is a 1-1 function. Our concern is to find optimal controls by comparing different strategies using this construction rather than following the idea developed by Cunniffe et al. (2015) where controls are compared on simulated replicates using the Gillespie algorithm (Gillespie, 1976). We again use the Sellke thresholds as the process $\boldsymbol{Q}$ as in 2.2.2.

Using the above parameterisation, we proceed as follows to solve the optimisation problem defined in Equation (5.2.1):

1. Draw m samples from $\pi_0\left(\underline{x}(T), \boldsymbol{\theta}|\mathbf{y}\right)$ to generate samples from $\pi_0\left(\boldsymbol{Q}, \boldsymbol{\theta}|\mathbf{y}\right)$ by inverting the Equation 5.2.2.

2. We then apply a control $\boldsymbol{d}$ to all realisations $(\boldsymbol{Q}, \boldsymbol{\theta}) \sim \pi_0\left(\boldsymbol{Q}, \boldsymbol{\theta}|\mathbf{y}\right)$, to generate samples from $\pi_{\boldsymbol{d}}\left(\underline{x}(T)|\mathbf{y}\right)$.

3. We approximate the expected value defined in Equation 5.2.1 as

$$\hat{U}(\boldsymbol{d}) = \frac{1}{m}\sum_{i=1}^{m} u(\underline{x}^i(\theta^i, Q^i, \boldsymbol{d}, T)) \tag{5.2.3}$$

where $\underline{x}^i(\boldsymbol{\theta}^i, \boldsymbol{Q}^i, \boldsymbol{d}, T)$ is a sample from $\pi_{\boldsymbol{d}}(\underline{x}(T)|\mathbf{y})$ corresponding to the process $(\boldsymbol{\theta}^i, \boldsymbol{Q}^i)$ to which the design $\boldsymbol{d}$ is applied and observed until time T.

The expected effect of various control strategies can then be compared. For different control strategies $\boldsymbol{d}$ and $\boldsymbol{d}'$, the expected difference in the effect of $\boldsymbol{d}$ and $\boldsymbol{d}'$ can be computed as $\hat{U}(\boldsymbol{d}) - \hat{U}(\boldsymbol{d}')$. This quantity is known in causal inference as the average causal effect (Holland, 1986).

5.2.3 Epidemiological model

We consider a spatially-explicit, stochastic, individual-based, compartmental SI (Neri et al., 2014) model for the spread of an infectious disease throughout a discrete population where members are located as points in a 2-dimensional continuous region. Hosts are then identified by their coordinates (e.g. latitude/longitude). We assume that at time t, any host in the population is either Susceptible (S) or Infected (I). Hosts in the latter category are either cryptic or symptomatic. The stochastic process that governs the transition from state S to state I in the infinitessimal time increment $[t, t + dt]$ is given by the following equation:

$$P\left(i \text{ infected in } [t, t + dt]\right) = \lambda_i(t)dt + o(dt), \tag{5.2.4}$$

$$\lambda_i(t) = \beta \sum_{j \in I(t)} K(d_{ji}, \alpha) + \epsilon \tag{5.2.5}$$

where $\lambda_i(t)$ is the hazard, or the force of infection on host i at time t, β is the contact parameter and ϵ the primary infection rate, this being the rate at which any individual contracts the disease from external sources, assumed fixed over the observation period. In addition, $K(d_{ji}, \alpha)$ is a non-negative function characterizing the challenge posed by the host j to i as a function of the inter-host distance d_{ji}, and known as the dispersal kernel with parameter α (the dispersal parameter). It is worth noting that with fixed α, the function K in most typical cases decreases with d_{ij}. We denote by $S(t)$ and $I(t)$ respectively the set of susceptibles and infectives at time t.

Moreover, we assume without loss of generality that individuals remain asymptomatic for a fixed period of time Δ which is known, then ultimately become symptomatic. A random sojourn time in the cryptic compartment could alternatively be modelled by assigning an appropriate distribution, for instance a Gamma or Weibull distribution (Parry et al., 2014).

It is worth noting that any symptomatic host is easily identifiable as being infective, while cryptic hosts are only identifiable through some diagnostics tests. This makes control difficult. Moreover, both symptomatic and cryptic hosts present a threat to other individual in the population. The described model is indeed relevant to many plant diseases, especially to arboreal pathogens such as citrus canker.

5.2.4 Control space construction

We make some assumptions regarding the type of controls we use in this chapter. We assume that although a host, cryptic at the time of a survey contributing to $\mathbf{y}$, will not be recorded as infected in that survey, all infections are observable during the control phase, thanks to the availability of a diagnostic test. We consider a system where the host population is partitioned into N_r fixed sub-regions. Specifying the control here consists of identifying the number n_r of sub-regions to visit and the maximum number of hosts V_i, $i = 1, \ldots, n_r$ to remove in each region where $\sum_{i=1}^{n_r} V_i \leq N'$. The precise regions to be considered will be decided using the prioritisation scheme of the next section.

5.2.5 Prioritisation scheme

We now define measures used as criteria for host prioritisation. For each host, we construct different metrics that are used to prioritise hosts for consideration under a given control strategy.

We define by $E(G^j(\underline{x}(t))|\mathbf{y})$, the posterior expectation of a measure at time $t \in [t_c, T]$ for host j. A measure of this kind is used to select hosts within regions when deploying controls. This concept has been often used in the literature to target priority sites (Boender et al., 2007; Tildesley et al., 2009; Kao, 2003; te Beest et al., 2011; DEFRA, 2013). Typically, the candidate host with the highest expected measure is prioritised.

Here, for each host we define the measures $G_R^j(\underline{x}(t))$ and $G_C^j(\underline{x}(t))$ respectively as the infection status of j at t under trajectory $\underline{x}(t)$ and the challenge or the infection hazard posed to the remaining susceptibles if that host were infected at time t. More formerly, the risk measure is given by

$$G_R^j(\underline{x}(t)) = \mathbb{1}_{\{x_j \leq t\}} \tag{5.2.6}$$

where x_j is the infection time of host j and $\mathbb{1}$ is the indicator function. The challenge is defined as:

$$G_C^j(\underline{x}(t)) = \beta \sum_{i \neq j} K(d_{ij}, \alpha) \mathbb{1}_{\{x_i > t\}} \tag{5.2.7}$$

Hence, we denote by the posterior expected risk and the challenge respectively as

$$\mathcal{R}_j(t) = E\left(G_R^j(\underline{x}(t))|\mathbf{y}\right) \tag{5.2.8}$$

and

$$\mathcal{C}_j(t) = E\left(G_C^j(\underline{x}(t))|\mathbf{y}\right) \tag{5.2.9}$$

5.2.5.1 Threat measure

In (DEFRA, 2013), it has been argued that considering such measures separately for prioritisation may not be cost-effective. For example removing a host with high risk might be less cost-effective when its challenge is low. The argument is that such a host is probably surrounded by other infectious hosts. They concluded that a combination of the two measures will provide the best prioritisation scheme. Following a similar idea, we define the *threat* posed by each host at time t as the posterior expectation of the product of the risk and challenge given the observed data $\mathbf{y}$ as :

$$\mathcal{T}_j(t) = E\left(G_R^j(\underline{x}(t))G_C^j\left(\underline{x}(t)\right)|\mathbf{y}\right) \tag{5.2.10}$$

5.2.5.2 Alternative measure

We propose an additional metric for comparison purposes as being the product of posterior expectations:

$$\mathcal{U}_j(t) = \mathcal{R}_j(t)\mathcal{C}_j(t) \tag{5.2.11}$$

Intuitively, the two metrics in Equations 5.2.10 and 5.2.11 are not the same given that we might anticipate a negative correlation between G_R^j and G_C^j. The reason is that, as the likelihood of infection increases, so does the probability that infection close to j have occurred, reducing the number of susceptibles in the vicinity. The challenge is then decreased.

5.3 Data and Inference

We suppose that the data $\mathbf{y}$ consist of a sequence of snapshots observed at particular times in $[t_0, t_{obs}]$ and we wish to predict the trajectory of the epidemic up to some future time T. As this requires the imputation of additional quantities, notably the infection times, it is well suited to Bayesian data-augmentation approaches (Neri et al., 2014; Parry et al., 2014; Lau et al., 2015).

Recall that in Bayesian framework, any inference is obtained from the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y}) \propto L(\boldsymbol{\theta})\pi(\boldsymbol{\theta})$, where $L(\boldsymbol{\theta})$ is the likelihood of the model parameters and $\pi(\boldsymbol{\theta})$ reflects any *a priori* knowledge of the distribution of the parameters. Therefore, we conduct our analysis in this framework using Bayesian computational method namely Markov Chain Monte Carlo (MCMC) along with data augmentation to draw samples from the posterior distribution $\pi(\boldsymbol{\theta}, \underline{x}(T)|\mathbf{y})$.

5.3.1 'Complete data' likelihood

We fist define the likelihood in the 'complete data' setting where the observations are $\underline{x}(T)$ and comprise the precise times of infection, and identify the infected individuals, for all infections occurring in the interval $[t_0, T]$ for some $T > t_0$. We assume t_0 to be the time at which the primary source of infection becomes active. Let $\mathcal{I}$ denote the set of infected hosts, $\tilde{\mathcal{I}}$ its complement and, for $i \in \mathcal{I}$, let t_i denote its infection time. Then a complete-data likelihood can be constructed as:

$$L(\boldsymbol{\theta}; \underline{x}(T)) = \prod_{i \in \mathcal{I}} \lambda_i(t_i^-) \exp\left(- \int_{t_0}^{t_i} \lambda_i(u) du \right) \times \prod_{\boldsymbol{s} \in \bar{\mathcal{I}}} \exp\left(- \int_{t_0}^{T} \lambda_s(u) du \right)$$

where $\lambda_i(t) = \lambda_i(t, \underline{x}(T), \theta)$ denotes the infections challenge presented to i at time t, under the realisation $\underline{x}(T)$, for $t \in [t_0, T]$.

Four computational purposes it is helpful to write this likelihood using the approach of Britton and O'Neill (2002) and Neal and Roberts (2005). First assign the 'notional' infection time for all $j \in \tilde{\mathcal{I}}$ to be $t_j = T$. Then we can write

$$L(\boldsymbol{\theta}; \underline{\boldsymbol{x}}(\boldsymbol{T})) = \prod_{i \in \mathcal{I}} \lambda_i(s_i^-) \exp\left(- \sum_{i \in \mathcal{I}} \sum_{i \in \mathcal{I} \cup \tilde{\mathcal{I}}} A_{ij}(t_j - t_i) \mathbb{1}_{t_j > t_i} + \epsilon \sum_{j=1}^{N} t_j \right) \tag{5.3.1}$$

where $A_{ij} = \beta K(d_{ij}, \alpha)$ is the infective pressure of host i on j and $\mathbb{1}$ the indicator function.

5.3.2 Parameter estimation with MCMC method

Parameter estimation should be straightforward if the observation $\mathbf{y}$ is complete in the sense of 5.3.1, in other words if all infection times are observed precisely. However, epidemiological data are partially observed. Therefore the problem of parameter estimation with missing data must be addressed using MCMC along with data augmentation methods (see Section 3.5). Recall that it entails treating the unobserved events as nuisance parameters and, given the observed data, the joint posterior density of the model and these nuisance parameters is then investigated. Inference on model parameters is achieved from consideration of $\pi_0(\boldsymbol{\theta}|\mathbf{y})$ with inference on individual components being made from the respective marginal density.

Our algorithm will generate samples from the joint posterior $\pi_0(\boldsymbol{\theta}, \underline{x}(T)|\mathbf{y})$ where T can be chosen in a number of ways. First note that the data $\mathbf{y}$, being a sequence of snapshots of symptomatic sets of hosts can be interpreted as specifying a period for the infection of each symptomatic host of the form $[\tau_{j-1} - \Delta, \tau_j - \Delta]$ where τ_j is the

time at which the host was first observed as symptomatic. It follows that a suitable algorithm could be designed by setting $T = t_{obs} - \Delta$, as the data in effect distinguish hosts infected before $t_{obs} - \Delta$ from those infected after $t_{obs} - \Delta$. However, given the need to impute infections beyond $t_{obs} - \Delta$ to investigate the posterior distribution of the prioritisation measures we implement a more general algorithm with $T > t_{obs} - \Delta$.

We describe the MCMC algorithm used to simulate from $\pi(\boldsymbol{\theta}|\mathbf{y})$ in 5.3.2.1 and 5.3.2.2 where the approach to updating components of $\boldsymbol{\theta}$ and of $\underline{x}(T)$ are discussed separately.

5.3.2.1 Updating the model parameters, θ

We update parameters using a single-component Metropolis algorithm, since the posterior distribution of each one of the parameters is not in a closed form, by proposing and accepting or rejecting changes to the current values of α, β and ϵ respectively. A new value of each parameter condition on all the others is proposed from a normal distribution with mean the current value. More precisely, the new parameter θ_k', $k = 1, 2, 3$ is proposed as follows:

$$\theta_k' = \theta_k + u, \text{ where } u \sim N(0, \sigma_k^2). \tag{5.3.2}$$

If $\theta_k' > 0$, it is then accepted with probability

$$\rho = \min\left\{1, \frac{\pi(\boldsymbol{\theta}', \underline{x}(T)|\mathbf{y})}{\pi(\boldsymbol{\theta}, \underline{x}(T)|y)}\right\} = \min\left\{1, \frac{\pi(\boldsymbol{\theta}')L(\boldsymbol{\theta}'; \underline{x}(T))}{\pi(\boldsymbol{\theta})L(\boldsymbol{\theta}; \underline{x}(T))}\right\} \tag{5.3.3}$$

where $\boldsymbol{\theta}'$ is the vector parameter $\boldsymbol{\theta}$ with θ_k replaced by θ_k'. Note that σ_k is a positive parameter which is tuned for each parameter separately to ensure the chain mixes well.

5.3.2.2 Updating $\underline{\mathbf{x}}$(T)

To update $\underline{x}(T)$ we need to take account of the fact that when $T > t_{obs}$, the number of infection events in $\underline{x}(T)$ is not specified by $\mathbf{y}$. This is overcome using a simple reversible-jump algorithm introduced in Section 3.5.2.5.

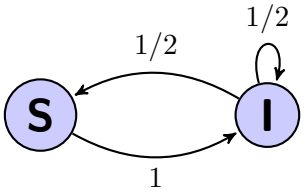


Figure 5.2: State diagram for the infection times to show the state transitions.

Given the current state of an individual, we propose changes to the infection times by adding, moving or deleting its infection time (see Figure 5.2). This is done as follows:

Updating $\underline{x}(T)$ using Reversible-Jump MCMC ($T > t_{obs} - \Delta$)

i) Choose an individual j in the population.

ii) If j is symptomatic at t_{obs} so that $t_j < t_{obs} - \Delta$, propose a new infection time

$$t' \sim U\left(\tau_{f(j)-1} - \Delta, \tau_{f(j)} - \Delta\right) \tag{5.3.4}$$

where $f(i)$ indexes the assessment time at which j was first symptomatic. Accept t' with probability

$$\rho = \min\left\{1, \frac{\pi(\boldsymbol{\theta}, \underline{x}'(T)|\mathbf{y})}{\pi(\boldsymbol{\theta}, \underline{x}(T)|\mathbf{y})}\right\} = \min\left\{1, \frac{L(\boldsymbol{\theta}; \underline{x}'(T))}{L(\boldsymbol{\theta}, \underline{x}(T))}\right\} \tag{5.3.5}$$

where $\underline{x}'(T)$ denotes the trajectory with t_j replaced by t'.

iii) If j is infected at T (but not symptomatic at t_{obs})

(a) with probability $p = 1/2$ move its infection time $t' \sim U[t_{obs} - \Delta, T]$. Accept t' with probability given by Equation 5.3.5.

(b) Delete its infection time with probability $p = 1/2$. The acceptance probability is then given by

$$\rho = \min\left\{1, \frac{2}{T - t_{obs} + \Delta} \frac{L(\boldsymbol{\theta}; \underline{x}'(T))}{L(\boldsymbol{\theta}; \underline{x}(T)}\right\}. \tag{5.3.6}$$

iv) If j is not infected at T, propose an infection time $t' \sim U[t_{obs} - \Delta, T]$. Accept t' with probability

$$\rho = \min\left\{1, \frac{T - t_{obs} + \Delta}{2} \frac{L(\boldsymbol{\theta}; \underline{x}'(T))}{L(\boldsymbol{\theta}; \underline{x}(T)}\right\}. \tag{5.3.7}$$

5.3.3 Imputation of the Sellke thresholds

Given a sample $(\boldsymbol{\theta}, \underline{x}(t))$ from $\pi_0(\boldsymbol{\theta}, \underline{x}(T)|\mathbf{y})$ we can impute the Sellke thresholds $\boldsymbol{Q}$ as follows:

$$Q_j = \begin{cases} \int_0^{t_j} \left(\beta \sum_{i \in I(u)} K(d_{ij}, \alpha) + \epsilon\right) du & \text{if } j \text{ is infected at } t_j < T \\ \int_0^{T} \left(\beta \sum_{i \in I(u)} K(d_{ij}, \alpha) + \epsilon\right) du + \zeta & \text{if } j \text{ is susceptible at } T \end{cases} \tag{5.3.8}$$

where $\zeta \sim Exp(1)$. Given a random draw $(\underline{x}(T), \boldsymbol{\theta}) \sim \pi_0(\underline{x}(T), \boldsymbol{\theta}|\mathbf{y})$, it is straightforward to use the construction in Equation 5.3.8 to impute the corresponding Sellke thresholds $\boldsymbol{Q}$. This results in a draw from the joint posterior distribution of the parameter and the thresholds, $\pi_0(\boldsymbol{\theta}, \boldsymbol{Q}|\mathbf{y})$.

The joint distribution $\pi_0(\boldsymbol{\theta}, \boldsymbol{Q}|\mathbf{y})$ is the key object of interest in this chapter since the underlying or latent stochastic process for the stochastic epidemic models are specified with a draw from $\pi_0(\boldsymbol{\theta}, \boldsymbol{Q}|\mathbf{y})$. Specifically, a random sample from the posterior distribution $(\boldsymbol{\theta}, \boldsymbol{Q}) \sim \pi_0(\boldsymbol{\theta}, \boldsymbol{Q}|\mathbf{y})$ is used as a population of 'pre-epidemics' on which subsequent analyses to compare controls can be based.

5.3.4 Imputation of the prioritisation measures

The imputation of the prioritisation measures $G^j(\underline{x}(t))$ given $(\underline{x}(t), \boldsymbol{\theta})$ is straightforward using Equations (5.2.8), (5.2.9), (5.2.10) and (5.2.11). In particular, for each draw $(\underline{x}(t)^k, \boldsymbol{\theta}^{(k)}) \sim \pi_0(\underline{x}(t), \boldsymbol{\theta}|\mathbf{y})$, the vector $\underline{G}_{\cdot}^{(k)}(\underline{x}(t)) = (G_{\cdot}^{1(k)}(\underline{x}(t)), \ldots, G_{\cdot}^{N(k)}(\underline{x}(t)))$ is computed, with $\underline{G}_{\cdot}^{(k)}$ being either the risk or the challenge. This will provide a sample from the joint posterior distribution $\pi_0\left(\underline{G}_R^j(\underline{x}(t)), \underline{G}_C^j(\underline{x}(t))|\mathbf{y}\right)$ for $1, \ldots, N$.

The measures defined in Equations (5.2.8), (5.2.9), (5.2.10) and (5.2.11) could then be approximated using the Monte Carlo approximation (see Chapter 3) respectively by:

$$\mathcal{R}_j^*(t) = \frac{1}{m} \sum_{k=1}^{m} G_R^{j(k)}(\underline{x}(t)) \tag{5.3.9}$$

$$\mathcal{C}_j^*(t) = \frac{1}{m} \sum_{k=1}^{m} G_C^{j(k)}(\underline{x}(t)) \tag{5.3.10}$$

$$\mathcal{T}_j^*(t) = \frac{1}{m} \sum_{k=1}^{m} \left(G_R^{j(k)}(\underline{x}(t)) G_C^{j(k)}(\underline{x}(t))\right) \tag{5.3.11}$$

$$\mathcal{U}_j^*(t) = \left(\frac{1}{m} \sum_{k=1}^{m} G_R^{j(k)}(\underline{x}(t))\right) \times \left(\frac{1}{m} \sum_{k=1}^{m} G_C^{j(k)}(\underline{x}(t))\right) \tag{5.3.12}$$

where m is the number of draws generated from $\pi_0(\underline{x}(t), \boldsymbol{\theta}|\mathbf{y})$.

5.4 Optimised control strategies

Having defined the measures for prioritisation, different controls strategies could be compared on the epidemic realisations using the Sellke construction. More importantly, given the available resources one might be able to identify the design that provides an optimal reduction in the outcome of the epidemic. With that goal in mind, an approach is followed whereby the outcome of the epidemic is represented deterministically and an optimisation method is applied to identify the optimal strategy.

Müller and Parmigiani (1995) introduced a general algorithm for curve-fitting based optimisation to find an optimal design of Equation (5.2.3). Their algorithm could be adapted here as follows:

1. Select a design $\boldsymbol{d}_i \in D$, $i = 1, \ldots, m$
2. Draw $(\boldsymbol{\theta}^i, \boldsymbol{Q}^i)$ from the joint posterior distribution $\pi_0(\boldsymbol{\theta}, \boldsymbol{Q}|\mathbf{y})$ for $i = 1, \ldots, m$ to generate $\pi_{\boldsymbol{d}_i}(\underline{x}(t)|\mathbf{y})$ and evaluate each $\hat{U}(\boldsymbol{d}_i)$.
3. Fit a smoothing curve as 'hypersurface' $\tilde{U}(\boldsymbol{d})$ to the scatterplot $(\boldsymbol{d}_i, -\hat{U}(\boldsymbol{d}_i))$
4. The optimal design could then be obtained deterministically for example the mode of the smoothed curve.

They remarked that this approach could be applied only to a low-dimensional design problem. In our framework however, optimised control of an epidemic may involve high-dimensional designs. Therefore, fitting a smoothing curve on the resulting scatterplot may be difficult.

We therefore propose to adapt the Simulated Annealing algorithm introduced in Chapter 4 (Bertsimas and Tsitsiklis, 1998; Busetti, 2003) to epidemics (Demon et al., 2011).

5.4.1 Control strategies and selection of regions

We assume that the host population is partitioned into N_r regions. We let M_i and H_i denote respectively the number of hosts in region i and the set of hosts in region i (so that $M_i = |H_i|$).

We now describe several strategies for selecting regions and hosts based on the measures defined in (5.3.9)-(5.3.12) as calculated for $t > t_{obs} - \Delta$. For each region i we define a measure, calculated from the samples $(\boldsymbol{\theta}, \underline{x}(T))$, and given by (in the case of $\mathcal{T}^*$)

$$w_i(t) = \frac{1}{M_i} \sum_{k \in \mathcal{H}_i} \mathcal{T}_k^*(t),\ i = 1, \ldots, N_r \tag{5.4.1}$$

with w_i analogously defined for $\mathcal{R}^*$, $\mathcal{C}^*$, $\mathcal{U}^*$.

We consider regions in descending order of $w_i(t)$. The first step in defining a control is to identify the first n_r regions from this order as those which will be considered under the control. The second step involves specifying the maximum number of hosts that can be removed in each region in the form of a vector $\boldsymbol{V} \in \mathbb{N}^{n_r}$. A control $(n_r, \boldsymbol{V})$ must satisfy a constraint $\sum_{j=1}^{n_r} V_j \leq N'$.

Finally, we must specify which V_i host should be considered in the i^{th} region selected. This is done by choosing the V_i hosts with the highest value of $\mathcal{T}^*$. We consider three strategies for identifying the values in $\boldsymbol{V} \in \mathbb{N}^{n_r}$.

1. **Strategy 1 (Equal allocation)**

$$V_i = \max\{k | k n_r \leq N'\}$$

2. **Strategy 2 (Probability-based allocation)**

   $(V_1, \ldots, V_{n_r})$ is selected to be a draw from a multinomial distribution with parameters N' and $\boldsymbol{P}$ where

$$P_i = \frac{w_i(t)}{\sum_{j=1}^{n_r} w_j(t)}.$$

   Note that Strategy 2 has the property that it involves a random step in the selection of $\boldsymbol{V}$. Moreover, it would select with high probability a region with few hosts but which had a high average value of the prioritisation measures. Therefore we consider a further strategy:

3. **Strategy 3 (Adaptation of Neyman allocation)**

$$\text{Here } V_i = \frac{N_i^2 w_i}{\sum_{j=1}^{N_r} N_j^2 w_j} N' \tag{5.4.2}$$

   The rationale of allocating resources in this manner is that they are distributed across regions proportionally to both the size of the regions and considering hosts measures. This is an analogy to the Neyman allocation in stratified sampling (Neyman, 1934) where the goal is to sample more heavily from a stratum (region) where the population size of the stratum is large and the variability within the stratum is large. By doing so, we hope to find a sampling method that could be easy to use in practice.

5.4.1.1 Optimisation strategy using Simulated Annealing

We begin by drawing $(\boldsymbol{\theta}^i, \boldsymbol{Q}^i)$ from the joint posterior distribution $\pi_0(\boldsymbol{\theta}, \boldsymbol{Q}|\mathbf{y})$, for $i = 1, \ldots, m$. We then modify the Algorithm 1 of Section 4.3.6.1 as follows:

Algorithm 1. Simulated Annealing for optimising the outcome of the epidemic considering fixed sample size m

1: **Initialisation**

Set the cooling temperature c_1, $k = 1$ and the initial design $\boldsymbol{d}^{opt} = \boldsymbol{d}^1$

Compute the approximation in Equation (5.2.3)

$\hat{U}^{opt} = \hat{U}^1 = \hat{U}(\boldsymbol{d}^1) = \frac{1}{m}\sum_{i=1}^{m} u(\underline{x}(\boldsymbol{\theta}^i, \boldsymbol{Q}^i, \boldsymbol{d}^{\mathbf{1}}))$.

2: Update to $\boldsymbol{d}$

3: **while** stopping criteria is not reached **do**

4: **for** $l = 1, \ldots, p$ **do**

5: **Update** n_r

$n_r' \leftarrow n_r + u$ where $u \sim U[-a, a]$ (only integer values).

6: **Update** $V_1, \ldots, V_{n_r'}$ using one of the following

i. **Strategy 1**

ii. **Strategy 2**

iii. **Strategy 3**

7: Use the same m samples and evaluate $\hat{U}' = \hat{U}(\boldsymbol{d}') = \frac{1}{m}\sum_{i=1}^{m} u(\underline{x}(\boldsymbol{\theta}^i, \boldsymbol{Q}^i, \boldsymbol{d}'))$

8: Set $(\hat{U}, \boldsymbol{d}) = \begin{cases} (\hat{U}', \boldsymbol{d}') & \text{if } \hat{U}' < \hat{U}^k \\ (\hat{U}', \boldsymbol{d}') & \text{if } \hat{U}' > \hat{U}^k \text{ with probability } \alpha \end{cases}$

where $\alpha = \min\left\{1, \exp\left(-\frac{\hat{U}' - \hat{U}^k}{c_k}\right)\right\}$

Set $(\hat{U}^{opt}, \boldsymbol{d}^{opt}) = (\hat{U}', \boldsymbol{d}')$ if $\hat{U}' < \hat{U}^{opt}$

9: **end for**

10: Set $(\hat{U}^{k+1}, \boldsymbol{d}^{k+1}) = (\hat{U}, \boldsymbol{d})$

11: Reduce the parameter c_k such that $c_{k+1} = \gamma c_k$, $\gamma \in [0, 1]$

12: $k \leftarrow k + 1$

13: **end while**

It has been mathematically proven that this algorithm converges to the global optimum (Kirkpatrick et al., 1983). Clearly, we are presented with a problem of choosing the Monte Carlo sample size m. However, it may be determined either empirically, using some prior information if available (e.g. an expert opinion) or simply considering a large enough m. The latter will involve more computations, therefore a more efficient method is proposed which adaptively increases the value of m. We modify the Algorithm 1 so that it gradually increases the size of the sample

used in the impact estimator.

Algorithm 2. Simulated Annealing with incremented sample size

1: **Initialisation**
Set the cooling temperature c_{11}, $m = 1$ and the initial design $\boldsymbol{d}^{opt} = \boldsymbol{d}^1$
Compute the approximation in Equation (5.2.3)
$\hat{U}^{opt} = \hat{U}^1 = \hat{U}(\boldsymbol{d}^1) = u(\underline{x}(\boldsymbol{\theta}^1, \boldsymbol{Q}^1, \boldsymbol{d}^1))$.
2: Update to $\boldsymbol{d}$
3: **while** stopping criteria is not reached **do**
4: **for** $l = 1, \ldots, p$ **do**
5: **for** $k = 1, \ldots, q$ **do**
6: **Update** n_r
$n_r' \leftarrow n_r + u$ where $u \sim U[-a, a]$ (only integer values).
7: **Update** $V_1, \ldots, V_{n_r'}$ using one of the following

i. **Strategy 1**

ii. **Strategy 2**

iii. **Strategy 3**

8: Use the same m samples and evaluate
$\hat{U}' = \hat{U}(\boldsymbol{d}') = \frac{1}{m} \sum_{i=1}^{m} u(\underline{x}(\boldsymbol{\theta}^i, \boldsymbol{Q}^i, \boldsymbol{d}'))$
9: Set $(\hat{U}, \boldsymbol{d}) = \begin{cases} (\hat{U}', \boldsymbol{d}') & \text{if } \hat{U}' < \hat{U}^k \\ (\hat{U}', \boldsymbol{d}') & \text{if } \hat{U}' > \hat{U}^k \text{ with probability } \alpha \end{cases}$
where $\alpha = \min\left\{1, \exp\left(-\frac{\hat{U}' - \hat{U}^k}{c_{mk}}\right)\right\}$
Set $(\hat{U}^{opt}, \boldsymbol{d}^{opt}) = (\hat{U}', \boldsymbol{d}')$ if $\hat{U}' < \hat{U}^{opt}$
10: **end for**
11: Reduce the parameter c_{mk} such that $c_{m(k+1)} = \gamma c_{mk}$, $\gamma \in [0, 1]$
12: **end for**
13: $m \leftarrow m + 1$, $\boldsymbol{d} \leftarrow \boldsymbol{d}^m = \boldsymbol{d}^{opt}$, $U \leftarrow U(\boldsymbol{d}^m)$
14: Set c_{m1}
15: **end while**

The algorithm defines a sequence $\boldsymbol{d}^1, \boldsymbol{d}^2, \ldots, \boldsymbol{d}^n$ which converges to an optimal design (see Figure 5.4), with n is the number of realisation used. Here, the global optimum is the lowest achievable expectation of the number of infections across the epidemic realisations and the design is hereafter referred as the optimal design (Demon et al., 2011). In practice, to reduce the computation time, one can define a stopping

criterion. Here, the algorithm stops if the design remains the same after a certain number s of successive realisations, where s is determined empirically. We mainly choose $s = 10$.

5.4.1.2 Optimal strategy (Algorithm 3)

In theory, the optimal estimate of the approximation in the Equation (5.2.3) would be the one obtained when each $(\boldsymbol{\theta}^i, \boldsymbol{Q}^i)$ has its own optimum i.e. the sequence of designs that minimises:

$$\hat{U}(\boldsymbol{d}^1, \ldots, \boldsymbol{d}^m) = \frac{1}{m} \sum_{i=1}^{M} u(\underline{x}(\boldsymbol{\theta}^i, \boldsymbol{Q}^i, \boldsymbol{d}^i, T)) \tag{5.4.3}$$

Each optimum can be obtained by optimising the resulting impact function i.e. the problem is then reduced to finding the design $\boldsymbol{d}^i$ that minimises the impact $u(\underline{x}(\boldsymbol{\theta}^i, \boldsymbol{Q}^i, \boldsymbol{d}, T))$. By doing so, we hope to check whether the optimal design obtained using our proposed method is reasonable. To that end, we fix $m = 1$ in the Algorithm 1 for each realisation $(\boldsymbol{\theta}^i, \boldsymbol{Q}^i)$.

5.5 Computational Issues and Parallel Computing

It can be easily understood that given all the missing information, the inference problems we tackle are high-dimensional. In addition, the Simulated Annealing algorithm involves computing the cumulative sum of outcomes of epidemics at every new proposed design resulting in computationally intensive algorithms. Although, solutions to the former problem are proposed by Deardon et al. (2010), some improvements need to be made. Parallel computing methods seem ideal for speeding up the computations. However, their applicability within both a dynamic MCMC and stochastic optimisation such as Simulated Annealing is not straightforward. This is because new state is always dependent on the previous one, thus we need to keep track of states that are already visited. Nevertheless, domain-decomposition parallelisation of the sums in the likelihood in the Equation 5.3.1 and the parallelisation of the cumulative sum of the outcome of the epidemics was achieved using a shared-memory architecture with an implementation of the OpenMP standard (Dagum and Menon, 1998). Although this technique is new to epidemic modelling, it has been successfully applied to speed up the computation of the likelihood by Jewell et al. (2009). Therefore, we use such method in order to be able to reduce the computation time.

5.6 Application to simulated data

Efficiency of the algorithms described here is assessed using simulated data and on simulations of outbreaks with citrus locations to mimic a much studied plant epidemic citrus canker (Gottwald et al., 2002; Cook et al., 2008; Neri et al., 2014; Cunniffe et al., 2015).

We begin by introducing some notation that will be used throughout the rest of this thesis. We denote by T_a the assessment time. We omit the design $\boldsymbol{d}$ to indicate that there is no control applied. Therefore the notation $u\left(\underline{x}(\boldsymbol{Q}, \boldsymbol{\theta}, T_a)\right)$ will denote the epidemic size at time T_a if no intervention is deployed. Given a sample $\{(\boldsymbol{\theta}^i, \boldsymbol{Q}^i) | i = 1, \ldots, m\}$ from $\pi_0(\boldsymbol{\theta}, \boldsymbol{Q}|\mathbf{y})$ and a control $\boldsymbol{d}$ we will be interested in the samples

$$\left\{u(\underline{x}(\boldsymbol{Q}^i, \boldsymbol{\theta}^i, \boldsymbol{d}, T_a) \,|\, i = 1, \ldots, m\right\} \tag{5.6.1}$$

and

$$\left\{ \mathcal{P}(\boldsymbol{Q}^i, \boldsymbol{\theta}^i, d, T_a) = \frac{u\left(\underline{x}(\boldsymbol{Q}^i, \boldsymbol{\theta}^i, T_a)\right) - u\left(\underline{x}(\boldsymbol{Q}^i, \boldsymbol{\theta}^i, d, T_a)\right)}{u\left(\underline{x}(\boldsymbol{Q}^i, \boldsymbol{\theta}^i, T_a)\right) - u\left(\underline{x}(\boldsymbol{Q}^i, \boldsymbol{\theta}^i, t_{obs})\right)} \,\middle|\, i = 1, \ldots, m \right\}, \tag{5.6.2}$$

these being samples of the total number of infections with control strategy $\boldsymbol{d}$ and the proportion of infections in $[t_{obs}, T_a]$ avoided using $\boldsymbol{d}$ in comparison to the no-control case. The sample means of the latter quantity will be used to compare the effect of competing control strategies. In addition, we denote by $d_{\{\mathcal{M},k,l\}}$ the control strategy obtained using the measure $\mathcal{M}$ (see Equations 5.3.9-5.3.12), when implementing the algorithm k, considering the strategy l and $d^*_{\{\mathcal{M},k,l\}}$ the corresponding optimal strategy. So for example, $d^*_{\{\mathcal{R}^*,2,1\}}$ will denote the optimal control strategy obtained using the Simulated Annealing Algorithm 2 with strategy 1 (equal allocation) using the risk measure $\mathcal{R}^*$. Finally, $UR(\mathcal{M})$ denotes the 'unconstrained regions' control strategy using the measure $\mathcal{M}$, this being control which consists of selecting the first N' in the ordering for examination.

5.6.1 Simulated data using a uniformly distributed location

We test the methodology on a spatio-temporal epidemic simulated in a population of size $N = 1000$, where their locations are sampled independently from a uniform distribution over a $0.75km \times 0.75km$ square region subdivided into 361 sub-regions (see Figure 5.3). The observation are made between $t = 0$ (time corresponding to the introduction of the external source of infection) and $t_{obs} = 460$; and the data consist of a sequence of snapshots of symptomatic set of hosts taken at $30-$days intervals. The entire population is assumed susceptible at time $t = 0$ and the process is governed by the Equation (5.2.4). We use $\alpha = 0.08km$, $\beta = 7.10^{-6}\text{days}^{-1}km^2$ and $\epsilon = 5.10^{-5}\text{days}^{-1}$ for the simulation and consider an exponential kernel $K(d, \alpha) =$

$\frac{1}{2\pi d\alpha}\exp(-d/\alpha)$. Moreover, we constrain the time it takes for symptoms to appear following an infection to $\Delta = 100$ days. As discussed earlier, the data $\mathbf{y}$ specify a period of time where each symptomatic host is infected and their corresponding locations. By the end of the observation, there are 128 symptomatic hosts detected while 153 are undetected (cryptic). Figure 5.3 illustrates the epidemic progress.

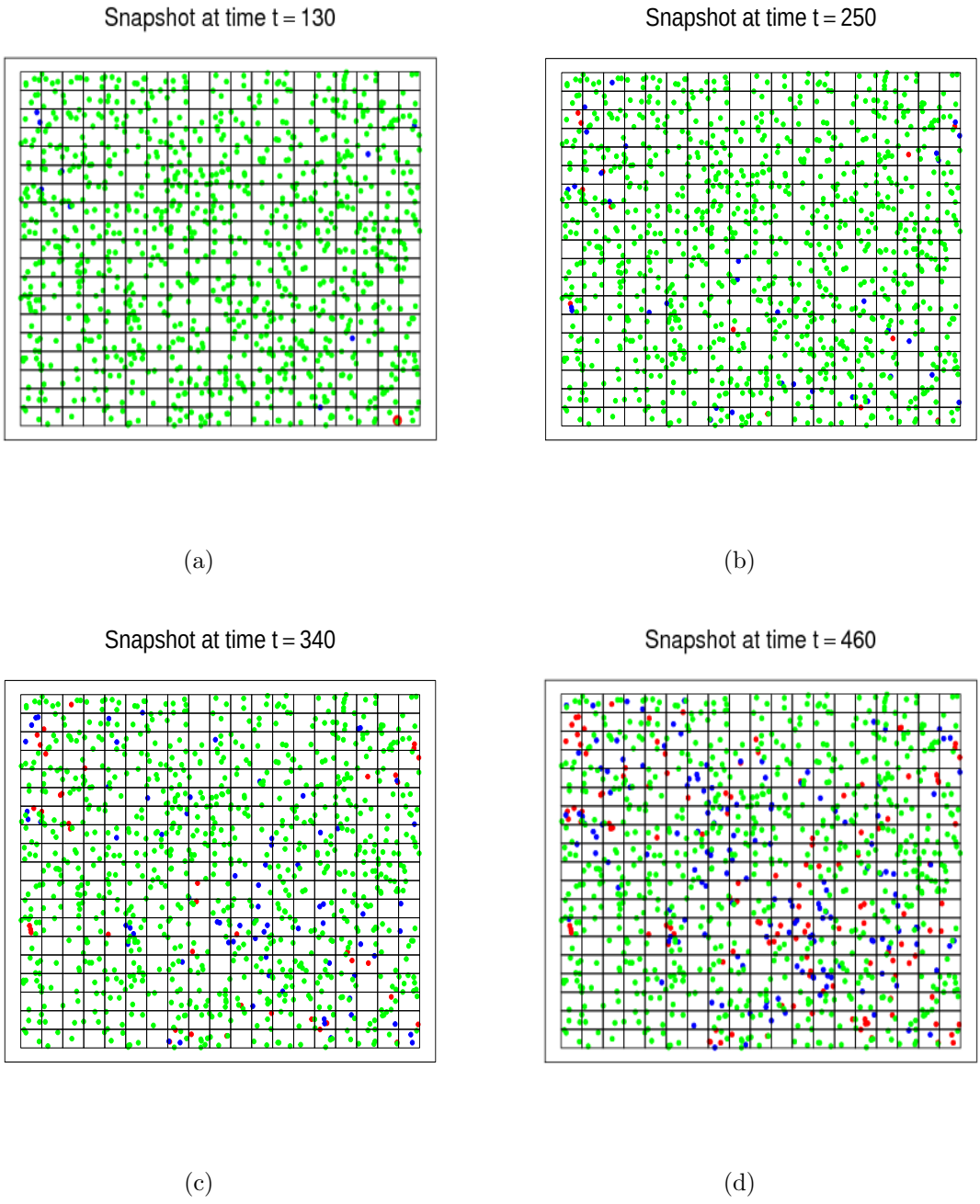


Figure 5.3: Sample of the disease progress maps made at 30 days intervals from $t = 130$ up to $t = 460$, on a population of size $N = 1000$ from simulated data. Symptomatic hosts, cryptic infections and susceptible hosts at the time of the snapshot are denoted by red, blue and green dots respectively.

We use the MCMC routines described in Section 5.3.2 to sample from the posterior

distribution $\pi_0(\boldsymbol{\theta}, \underline{x}(T)|\mathbf{y})$. Non-informative uniform priors $U[0, 1000]$ are used for all parameters.

1. The first estimation is carried out by imputing infections times up to $T = t_{obs} - \Delta$.

2. Next, the cryptic hosts are included by imputing infections times up to $T = t_{obs}$. Recall that this will correspond to imputing new infections in $[t_{obs} - \Delta, t_{obs}]$.

3. Finally, we impute infection times beyond the observation period up to $T = T_a$ (assessment time).

Note that 2) and 3) will require the use of RJMCMC techniques and the assessment time is fixed to be $T_a = 500$.

For that end, we run the algorithms for 280000 iterations for the case $T = 360$ and 10^6 iterations for $T = 460$ and $T = 500$ discarding the firsts 10000 iterations to ensure that convergence to stationary distribution is reached. The trace plots in Figure 5.5 show that the chains are mixing well and no sign of non-convergence. The posterior distributions of the parameters along with the epidemic size at $T = 460$ and $T = 500$ are shown in Figure 5.6. We can observe from this figure that the true parameter values (dash lines) are consistent with their respective posterior distributions. The posterior distributions at different T shown on Figure 5.6 suggest that estimated $\pi_0(\boldsymbol{\theta}|\mathbf{y})$ is the same regardless of which algorithm is used and how far beyond t_{obs} we impute infection times. This provides a check on the validity of the implementation of the MCMC algorithms.

Figure 5.7 shows different measures described in Equations (5.3.9-5.3.12) constructed at $T = 500$ using 100000 realisations from the joint posterior distribution $\pi_0(\boldsymbol{\theta}, \underline{x}(T)|\mathbf{y})$. These measures are used to test our control strategies described in 5.4.1. We focus the discussion on Strategy 1 (equal allocation), though similar features are obtained with the strategies 2 and 3 discussed in the next section. We assume that we are allowed to sample a maximum of $N' = 500$ hosts for removal and that the intervention occurs on day $t_c = 461$, the day following the last observation, as a quick response to the disease is more cost-effective (Cunniffe et al., 2015) though any given time beyond the last time the system is observed could be used. Recall that symptomatic are automatically removed if sampled whereas cryptic hosts are tested and removed if positive. Here, we consider the design of the form $d_{\{\mathcal{M},2,1\}}$ for $\mathcal{M} \in \{\mathcal{C}^*, \mathcal{R}^*, \mathcal{U}^*, \mathcal{T}^*\}$ (see Equations 5.3.9-5.3.12) and we use samples $\{(\boldsymbol{Q}^i, \boldsymbol{\theta}^i)|i = 1, \ldots, 1000\}$ from the posterior distribution $\pi_0(\boldsymbol{Q}, \boldsymbol{\theta}|\mathbf{y})$ for identifying optimal controls.

After setting an appropriate initial design, $\boldsymbol{d}_0 = (125; 4)$ for $\mathcal{M}$, we run the SA Algorithm 2 for $\{(\boldsymbol{\theta}^i, \boldsymbol{Q}^i),\ i = 1, \ldots, 20\}$ realisations. Figure 5.4 shows

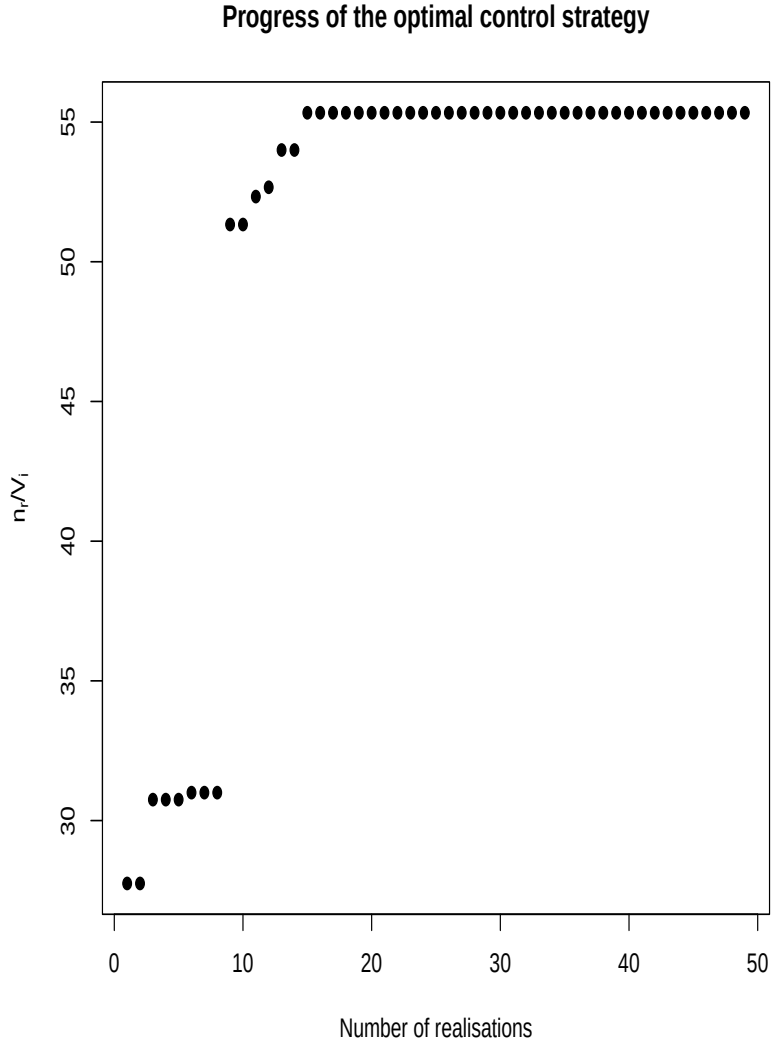


Figure 5.4: Ratio of the number of regions n_r and the number of hosts V_i to visit as a function of the number of realisation of the epidemic process for a typical run of the Algorithm 2 using the citrus locations.

the progress of a typical run of the algorithm for 50 realisations. The algorithm converges towards the optimal control strategies (second column of the Table 5.1). The summary statistics of the estimated posterior distribution of the proportion of infections avoided $\mathcal{P}\left(\underline{x}(\boldsymbol{Q}, \boldsymbol{\theta}, \boldsymbol{d}, T_a)\right)$ when applying different controls $\boldsymbol{d} \in \left\{d^*_{\{\mathcal{C}^*,2,1\}}, d^*_{\{\mathcal{R}^*,2,1\}}, d^*_{\{\mathcal{U}^*,2,1\}}, d^*_{\{\mathcal{T}^*,2,1\}}, d^*_{\{\mathcal{T}^*,3,1\}}\right\}$ on the $\{(\boldsymbol{Q}^i, \boldsymbol{\theta}^i)|i = 1, \ldots, 1000\}$ is shown in Table 5.1. The estimated posterior means of the proportion of infection avoided with control strategies $d^*_{\{\mathcal{C}^*,2,1\}}$ and $d^*_{\{\mathcal{R}^*,2,1\}}$ show that such interventions are less cost-effective in reducing the disease spread compared to the other measures; although the control using $\mathcal{R}^*$ performs better than $\mathcal{C}^*$. This is due to the fact that regions with high challenge coincide with regions where the majority of susceptible hosts are distant from infections. Therefore any removal of infectious hosts has little

effect in the amount of infective pressure on the remaining susceptibles. However, interventions using $d^*_{\{\mathcal{U}^*,2,1\}}$ and $d^*_{\{\mathcal{T}^*,2,1\}}$ differ to a very small extent, as seen from their corresponding 95% credible interval in Table 5.1 and by the distributions of infections avoided (see Figure 5.8). We found that regions selected for control in these two cases coincide and this explains why these two controls differ so little. Moreover, it appears from Figure 5.8 that there is no clear difference between the outcome of both control strategies $d^*_{\{\mathcal{T}^*,2,1\}}$ and the optimal $d^*_{\{\mathcal{T}^*,3,1\}}$. This suggests that the algorithm performs well and the optimal strategy obtained from our proposed method approaches the one of the optimal.

As expected, the control strategy $UR(\mathcal{T}^*)$ which involves applying the control on the first N' in the ordering across the entire set of 1000 realisations $(\boldsymbol{Q}^i, \boldsymbol{\theta}^i)$ gives the best reduction in the epidemic size (curve in cyan in Figure 5.8). This is an obvious candidate for disease control in a system where either the pathogen cannot be effectively controlled by chemical means (e.g. cocoa swollen-shoot virus (Dzhini-Obiatey et al., 2006), sweet potato stunt (Gibson and Aritua, 2002)) or the resources for the control allow surveying every single host in the population, which is unrealistic in a large population.

Finally, we show the pairwise joint posterior distributions of the outcome of the controls on Figure 5.9. Specifically, the lower panel shows the joint posterior distribution

$$\{(u\left(\underline{x}(\boldsymbol{Q}^i, \boldsymbol{\theta}^i, \boldsymbol{d}, T_a)\right), u\left(\underline{x}(\boldsymbol{Q}^i, \boldsymbol{\theta}^i, \boldsymbol{d}', T_a)\right) | i = 1, \ldots, 1000\}, \tag{5.6.3}$$

where $\boldsymbol{d}$, $\boldsymbol{d}' \in \left\{\text{no control}, d^*_{\{\mathcal{C}^*,2,1\}}, d^*_{\{\mathcal{R}^*,2,1\}}, d^*_{\{\mathcal{U}^*,2,1\}}, d^*_{\{\mathcal{T}^*,2,1\}}, UR(\mathcal{T}^*)\right\}$. We can notice a strong positive correlation between the outcomes of paired-controls (Figure 5.9). As a result, the variance between the outcome of different controls is reduced compared to an independent sampling allowing the difference in expected outcome to be estimated more efficiently for a given sample size. More precisely,

$$Var\left(n\left(\underline{x}(\underline{Q}, \underline{\theta}, \boldsymbol{d}, T_a)\right) - n\left(\underline{x}(\underline{Q}, \underline{\theta}, \boldsymbol{d}', T_a)\right)\right) < Var\left(n\left(\underline{x}(\underline{Q}, \underline{\theta}, \boldsymbol{d}, T_a)\right)\right) + Var\left(n\left(\underline{x}(\underline{Q}, \underline{\theta}, \boldsymbol{d}', T_a)\right)\right) \tag{5.6.4}$$

where

$$n\left(\underline{x}(\underline{Q}, \underline{\theta}, \boldsymbol{d}, T_a)\right) = \{u\left(\underline{x}(\boldsymbol{Q}^i, \boldsymbol{\theta}^i, \boldsymbol{d}, T_a)\right) | i = 1, \ldots, 1000\}$$

and $\boldsymbol{d}$, $\boldsymbol{d}' \in \left\{\text{no control}, d^*_{\{\mathcal{C}^*,2,1\}}, d^*_{\{\mathcal{R}^*,2,1\}}, d^*_{\{\mathcal{U}^*,2,1\}}, d^*_{\{\mathcal{T}^*,2,1\}}, UR(\mathcal{T}^*)\right\}$.

Therefore, for two control strategies $\boldsymbol{d}$ and $\boldsymbol{d}'$, the expected difference in their outcomes defined as $\hat{U}(\boldsymbol{d}') - \hat{U}(\boldsymbol{d}')$ (see Equation (5.2.3)) can be estimated more accurately using the Sellke construction. In addition, an independent sampling for example will require the same amount of samples for each controls strategies, leading to more sampling effort. Thus, coupling realisations under different control strategies

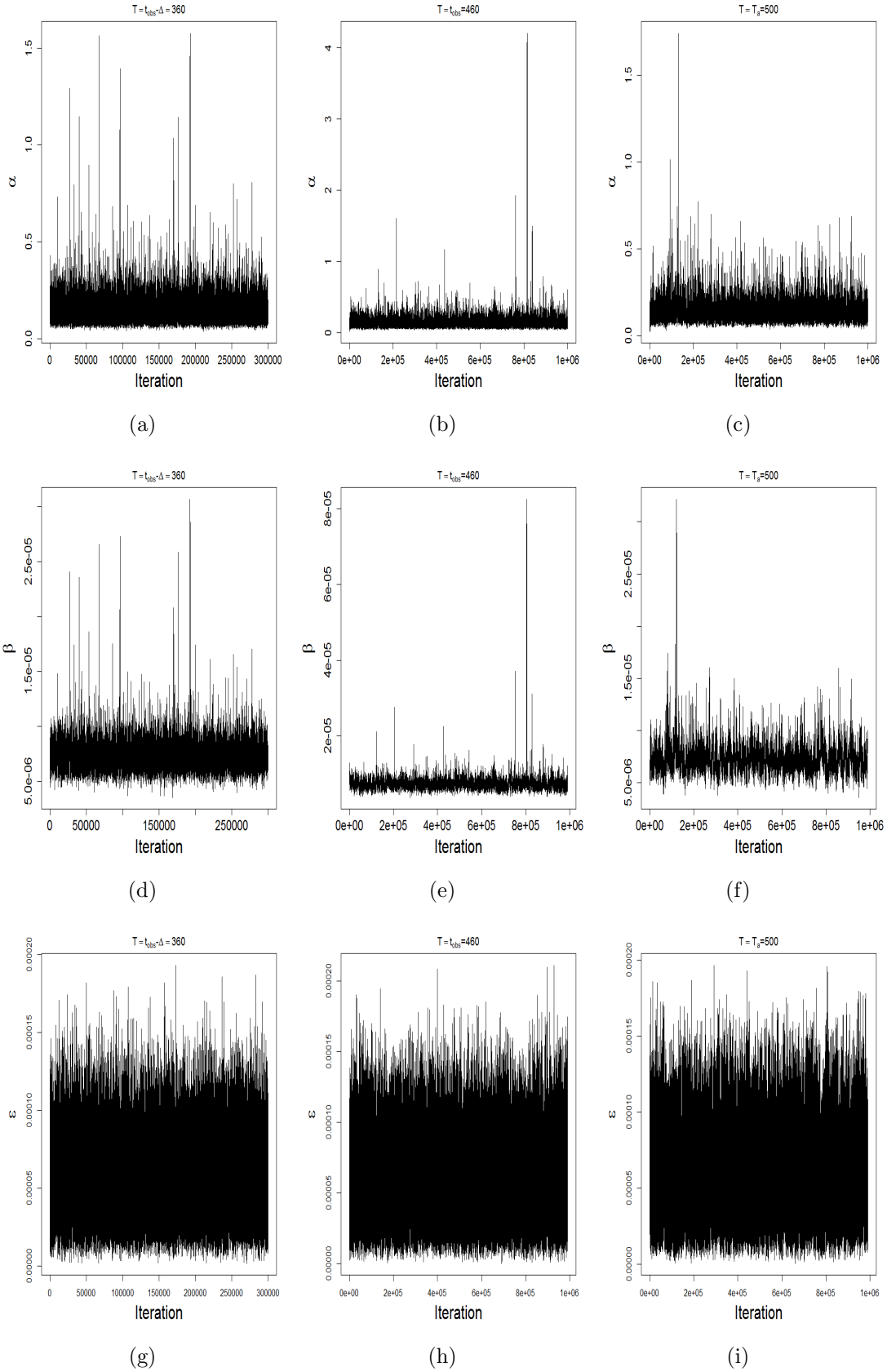
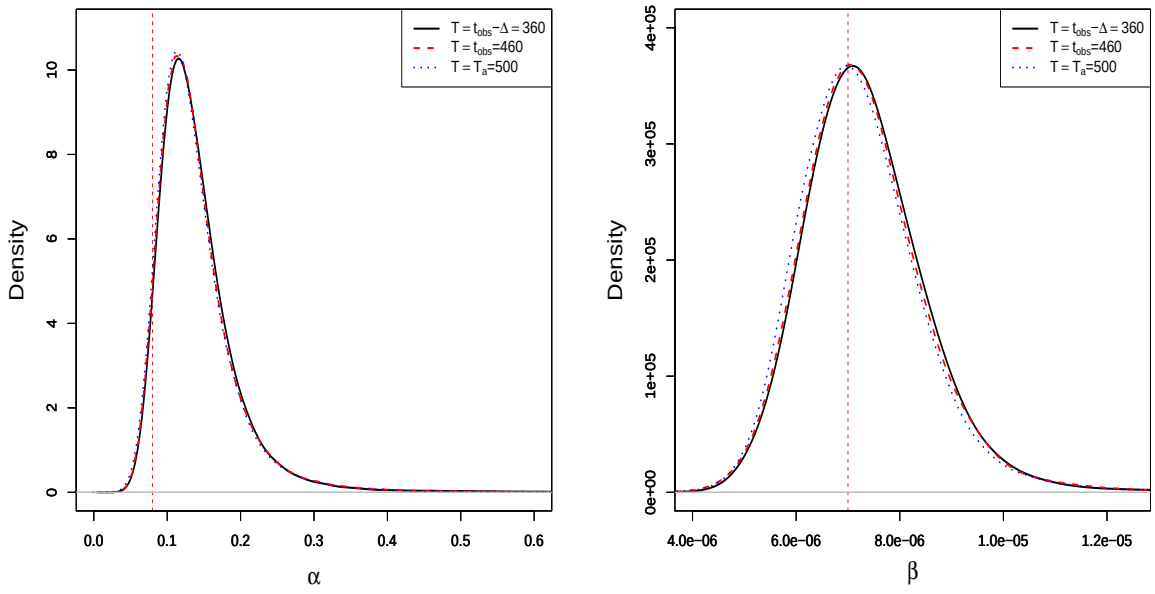
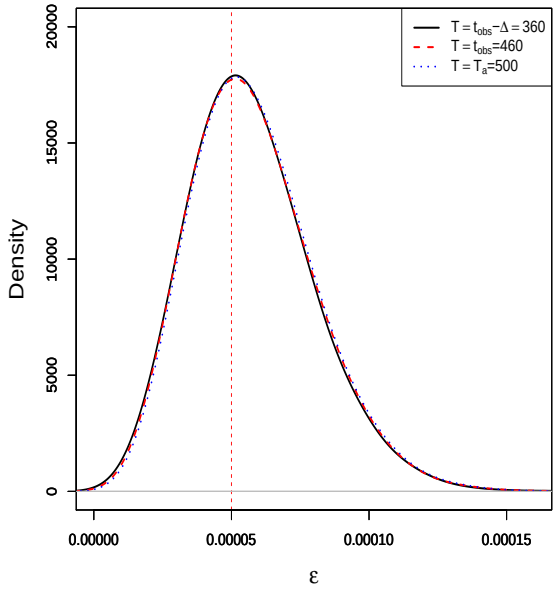


Figure 5.5: Sample trace plots for α, β and ϵ after a burn-in of 10000 iterations with no augmentation period ($T = 360$) (a), (d) and (g) and with different augmentation periods: $T = 460$ (b), (e) and (h) and $T = 500$ (c), (f) and (i).

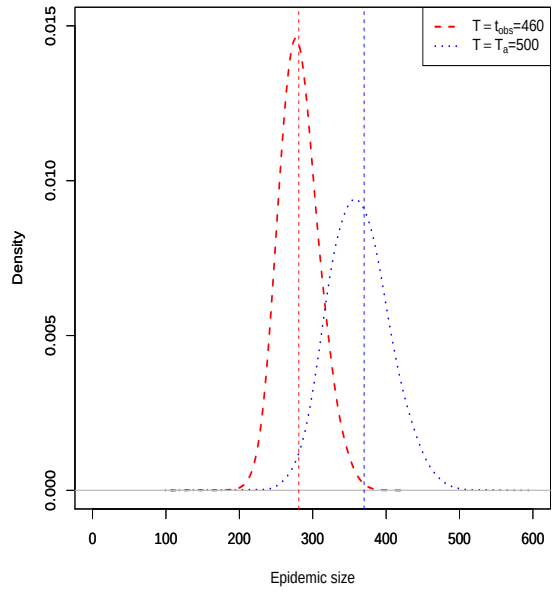


(a)

(b)

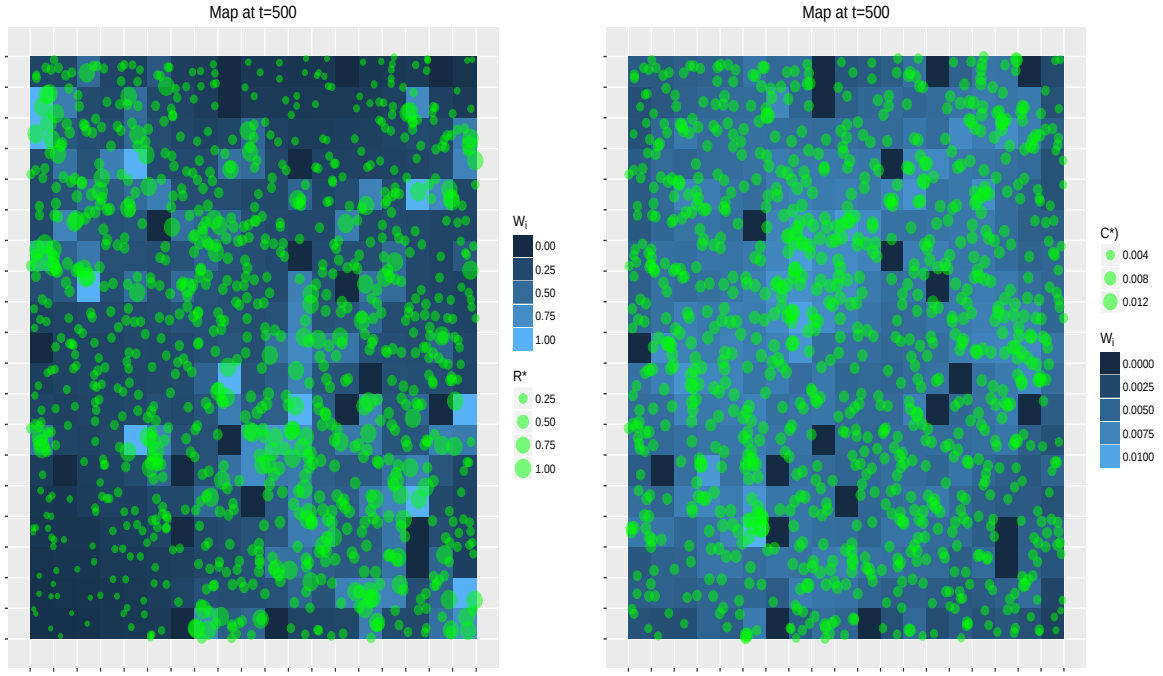


(c)



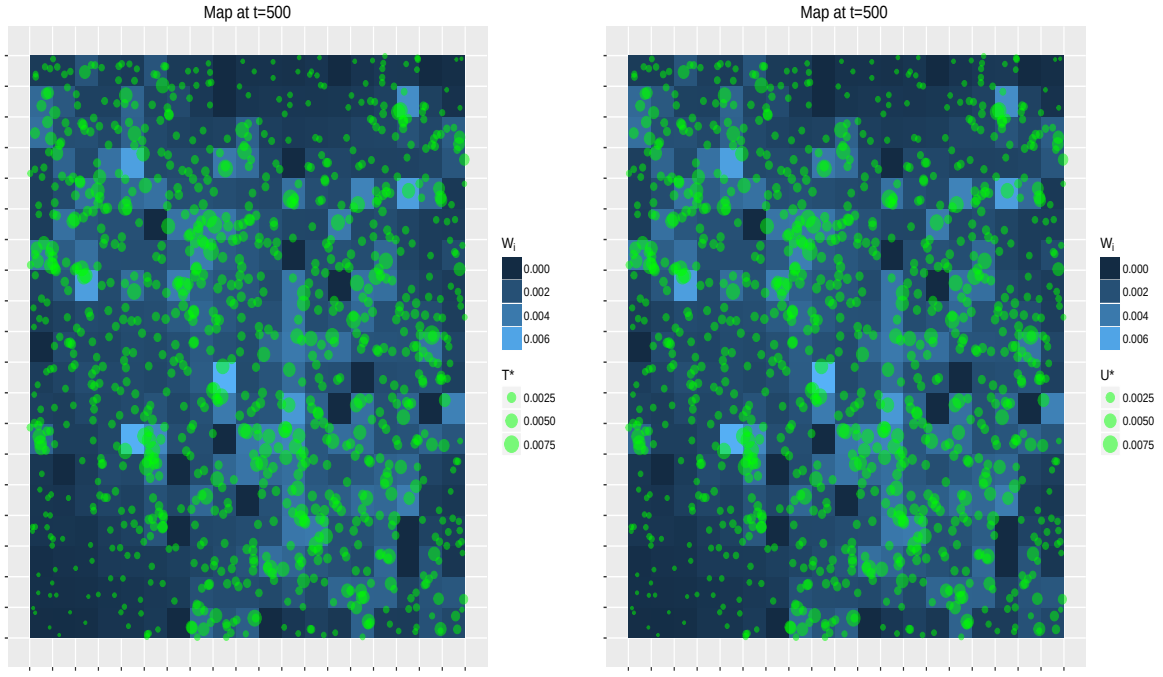
(d)

Figure 5.6: The posterior distributions for Bayesian MCMC estimation of the model parameters including the dispersal rate α (a), the secondary infection rate β (b) and the primary infection rate ϵ (c) conditioning on the data $\mathbf{y}$ while imputing the infections up to $T = 360$ (black full curve), $T = 460$ (red dash curve); $T = 500$ (blue curve). (d) The posterior distribution of the epidemic size at both 460 and 500 days. Dashed lines correspond to the actual value used for the simulation.



(a) $\mathcal{R}^*$ measure

(b) $\mathcal{C}^*$ measure



(c) $\mathcal{U}^*$ measure

(d) $\mathcal{T}^*$ measure

Figure 5.7: Plots showing different maps including $\mathcal{R}^*$ (a), $\mathcal{C}^*$ (b), $\mathcal{T}^*$ (c) and $\mathcal{U}^*$ (d) constructed at $T_a = 500$ using 100000 (after a burn-in period of 10000) samples from $\pi_0(\boldsymbol{\theta}, \underline{x}(T_a)|\mathbf{y})$.

reduces the amount of simulation required to compare control strategies in term of their expected outcomes compared to an independent sampling.

Prioritisation	Optimal design $(n_r; V_i)$	Mean	SD	0.025	Median	0.95
$d^*_{\{\mathcal{C}^*,2,1\}}$	(100;5)	0.458	0.07	0.333	0.459	0.6
$d^*_{\{\mathcal{R}^*,2,1\}}$	(125;4)	0.597	0.072	0.457	0.597	0.742
$d^*_{\{\mathcal{U}^*,2,1\}}$	(125;4)	0.614	0.072	0.474	0.616	0.756
$d^*_{\{\mathcal{T}^*,2,1\}}$	(125;4)	0.615	0.072	0.474	0.616	0.756
$UR(\mathcal{T}^*)$	-	0.72	0.067	0.597	0.72	0.842

Table 5.1: Summary Statistic for the ratio of the number of infections avoided $\mathcal{P}\left(\underline{x}(\boldsymbol{Q}, \boldsymbol{\theta}, \boldsymbol{d}, T_a)\right)$ during different interventions on day $t_c = 461$ relative to the number for no control.

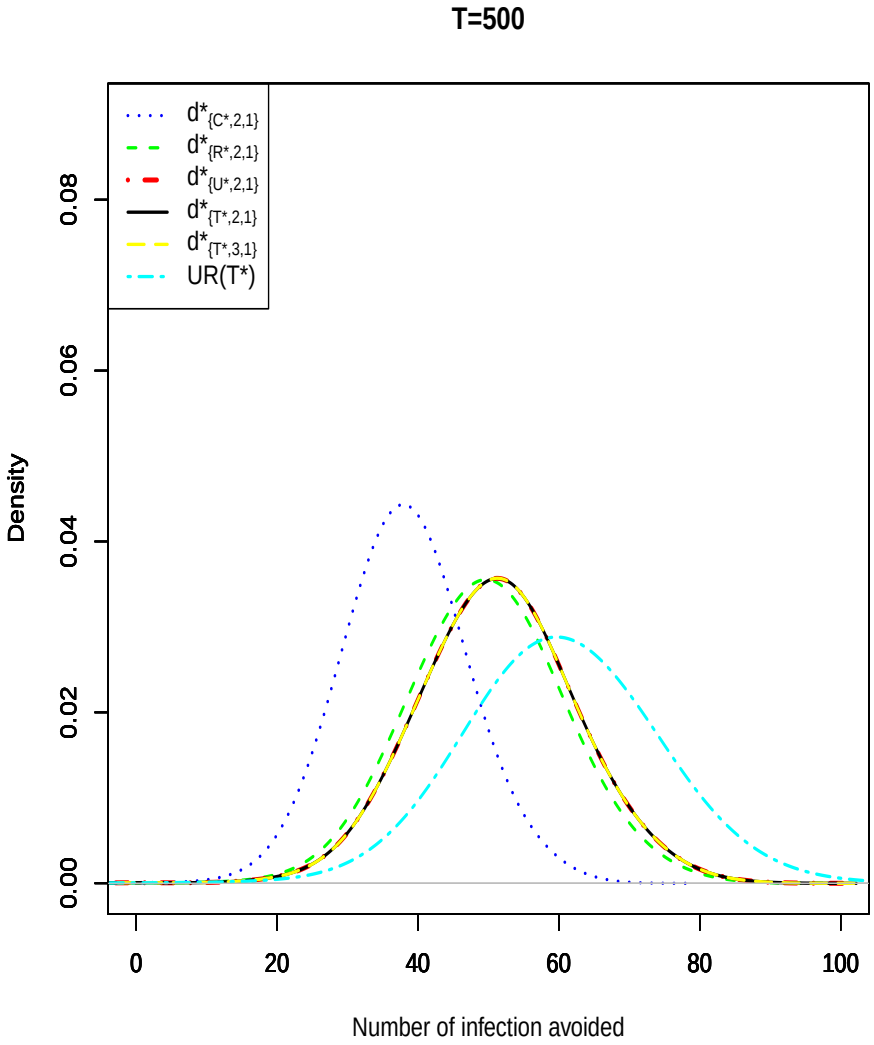


Figure 5.8: Posterior distribution of the number of infection avoided $u\left(\underline{x}(\boldsymbol{Q}^i, \boldsymbol{\theta}^i, T)\right) - u\left(\underline{x}(\boldsymbol{Q}^i, \boldsymbol{\theta}^i, \boldsymbol{d}, T)\right) \mid i = 1, \ldots, 1000$ for different $\boldsymbol{d}$ when deploying resources optimally on day $t_c = 461$, given that the maps are constructed on time $T_a = 500$ using simulated data on uniformly distributed locations.

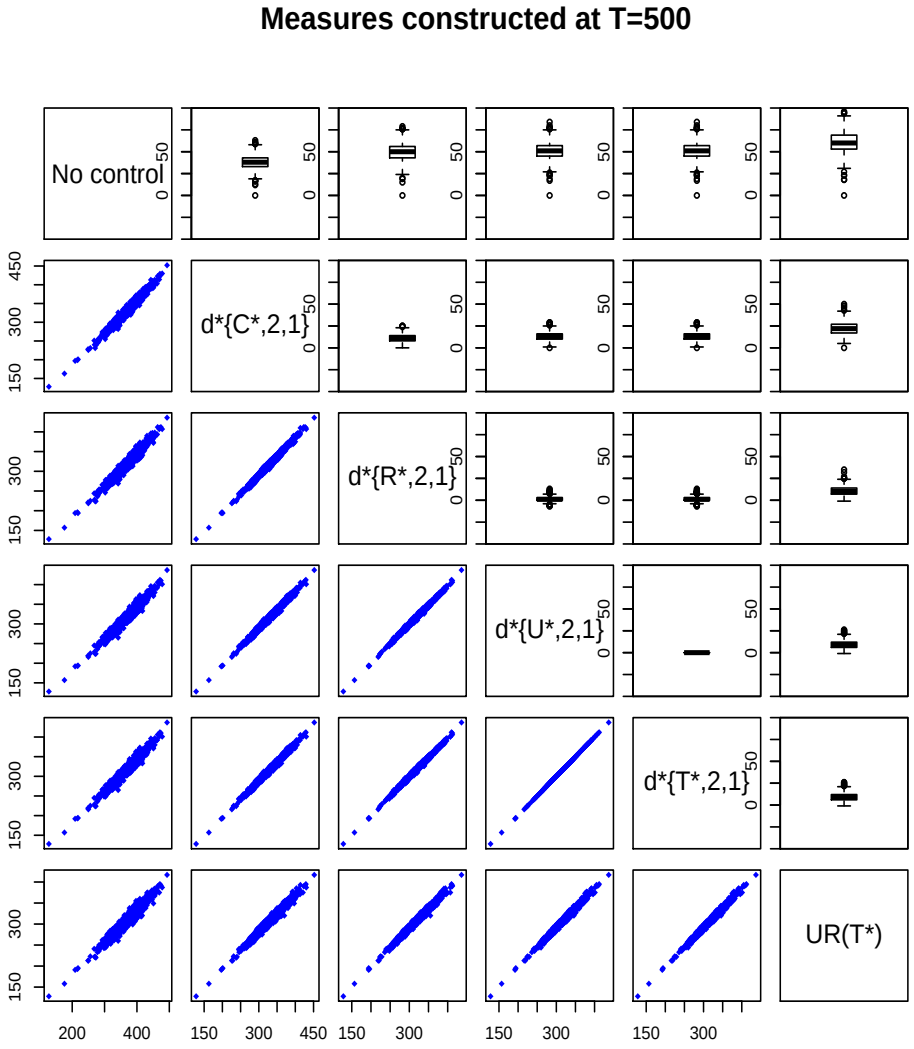


Figure 5.9: Pairwise joint posterior distribution of the epidemic size i.e. $\left(n\left(\underline{x}(\underline{Q},\underline{\theta},\boldsymbol{d},T_a)\right),\ n\left(\underline{x}(\underline{Q},\underline{\theta},\boldsymbol{d}',T_a)\right)\right.$ on the lower panel and the distribution of the difference between the outcomes of the paired controls $n\left(\underline{x}(\underline{Q},\underline{\theta},\boldsymbol{d},T_a)\right)$-$n\left(\underline{x}(\underline{Q},\underline{\theta},\boldsymbol{d}',T_a)\right)$ for $\boldsymbol{d}\neq d'$.

5.6.2 Application to structured populations: citrus locations from Florida

To illustrate the approach described above, we use data regarding citrus locations from Florida to mimic a realistic spatial distribution of hosts.

5.6.2.1 Introduction to the data

The data used for the analysis consist of the citrus locations from a site located in Broward county, labelled B_2 from the four sites in an urban region close to Miami

(Neri et al., 2014). A total of 18769 trees across the four sites were monitored with 1111 in B_2. In the four sites, the spatial locations of citrus were fully enumerated using a differential global positioning system. It is reported that over 16 inspectors, experienced at recognising citrus canker symptoms, visited each site at successive intervals between October 1997 and October 1999. The locations of the infected trees were identified and their corresponding infection times were estimated by experienced personnel from the lesions, size, tree age and other phenotypical characteristics.

5.6.2.2 Simulated data

The locations of the citrus population are then used to simulate epidemics governed by Equation (5.2.4). Two different normalised kernels with qualitatively different behaviour considered in Neri et al. (2014) are used for the simulations:

$$\textbf{Exponential:} \qquad K(d,\alpha) = \frac{1}{2\pi d}\frac{1}{\alpha}\exp(-d/\alpha) \tag{5.6.5}$$

$$\textbf{Cauchy:} \qquad K(d,\alpha) = \frac{1}{2\pi d}\frac{2}{\pi\alpha(1+d^2/\alpha^2)} \tag{5.6.6}$$

where d is the Euclidean distance between infected and susceptible hosts.

Note that the Exponential kernel characterizes a short-range disease spread while the Cauchy kernel exhibits a long-range dispersal. Three scenarios are investigated:

- Case (I): An exponential kernel with presence of primary infection

  We assume that the entire population is susceptible at time $t = 0$, the time corresponding to the introduction of the external source. The value used for the contact rate, the dispersal parameter and the primary infection rate are respectively $\beta = 7 \times 10^{-6}\,\text{days}^{-1}km^2$, $\alpha = 0.08\,km$, $\epsilon = 5 \times 10^{-6}\,\text{days}^{-1}$ and we observe the process up to time $t_{obs} = 460$ day where 169 were symptomatics with 235 cryptics. Figure 5.10 shows the progress of a simulation over time. These parameters are chosen from Neri et al. (2014) where they were estimated *via* MCMC using 12 months of the epidemiological data.

- Case (II): An exponential kernel with no primary infection.

  We repeat a similar experiment with $\beta = 8 \times 10^{-6}\text{days}^{-1}km^2$, $\alpha = 0.8km$ and $\epsilon = 0$ but assuming that $t = 0$ corresponds to the time of the initial infection. For convenience, we choose the first infection from the Canker data (Neri et al., 2014) to be the host initially infected. Here, we maintain $t_{obs} = 460$ but we observe 111 symptomatics and 124 cryptics (see Figure 5.11 for the progress of a simulation over time).

- Case (III): Cauchy kernel with presence of primary infection.

Again, we assume that $t = 0$ corresponds to the introduction of the external source of infection. Here, we use $\beta = 2 \times 10^{-5}\text{days}^{-1}km^2$, $\alpha = 0.2km$ and $\epsilon = 3 \times 10^{-5}\text{days}^{-1}$ and we observe the process up to $t_{obs} = 440$ days. We record 166 symptomatics and 339 cryptics at the end of the observation (see Figure 5.12 for the progress of the simulation).

In all cases, the cryptic period considered is $\Delta = 100$ days. The data consist of a sequence of snapshots taken at 30-day intervals over 360 days starting from day 130 in case (I) and (II) with 330 days starting from day 110 in case (III) to observe a similar number of symptomatic as in Case (I). This is summarised in Table 5.2.

It is worth noting that these kernels are merely used to illustrate our methodology. Details related to the kernel that best fits the citrus data can be found in Neri et al. (2014). It is known that the symptoms take, on average 100 days to emerge following infection (Parnell et al., 2009). Therefore, in this section we fix $\Delta = 100$, in line with the assumption by Parnell et al. (2009) and Neri et al. (2014).

Case	α	β	ϵ	t_{obs}	Infections observed	Cryptic	T
(I)	0.08	7.10^{-6}	0.00005	460	169	156	500
(II)	0.08	8.10^{-6}	0	460	111	124	500
(III)	0.2	2.10^{-5}	0.00003	440	166	339	500

Table 5.2: Summary of the parameters used and outcomes obtained for the simulations in the three cases.

5.6.2.3 Parameter estimation

For the parameter estimation, we adopt the MCMC algorithm described in 5.3. The estimation is done as in Section 5.6.1 varying T depending on the case considered (see Figures 5.13, 5.15 and 5.14). In all cases, the algorithm is run for 520000 steps with a burn-in period corresponding to the initial 20000 iterations.

In Figures 5.13, 5.15 and 5.14 we show the sample trace plots of the parameters in all cases. The convergence issues are of no concern as shown on the figures. Nevertheless, it is clearly apparent that the chain mixing depends on how far we augment the imputation period. For instance, the sample trace plot for parameters α in Figures 5.13a and 5.13b, and β in 5.13d and 5.13e suggest that complete knowledge on the number of infections ($T = t_{obs} - \Delta$) leads to a chain which mixes better than when precise times of future infections are imputed explicitly.

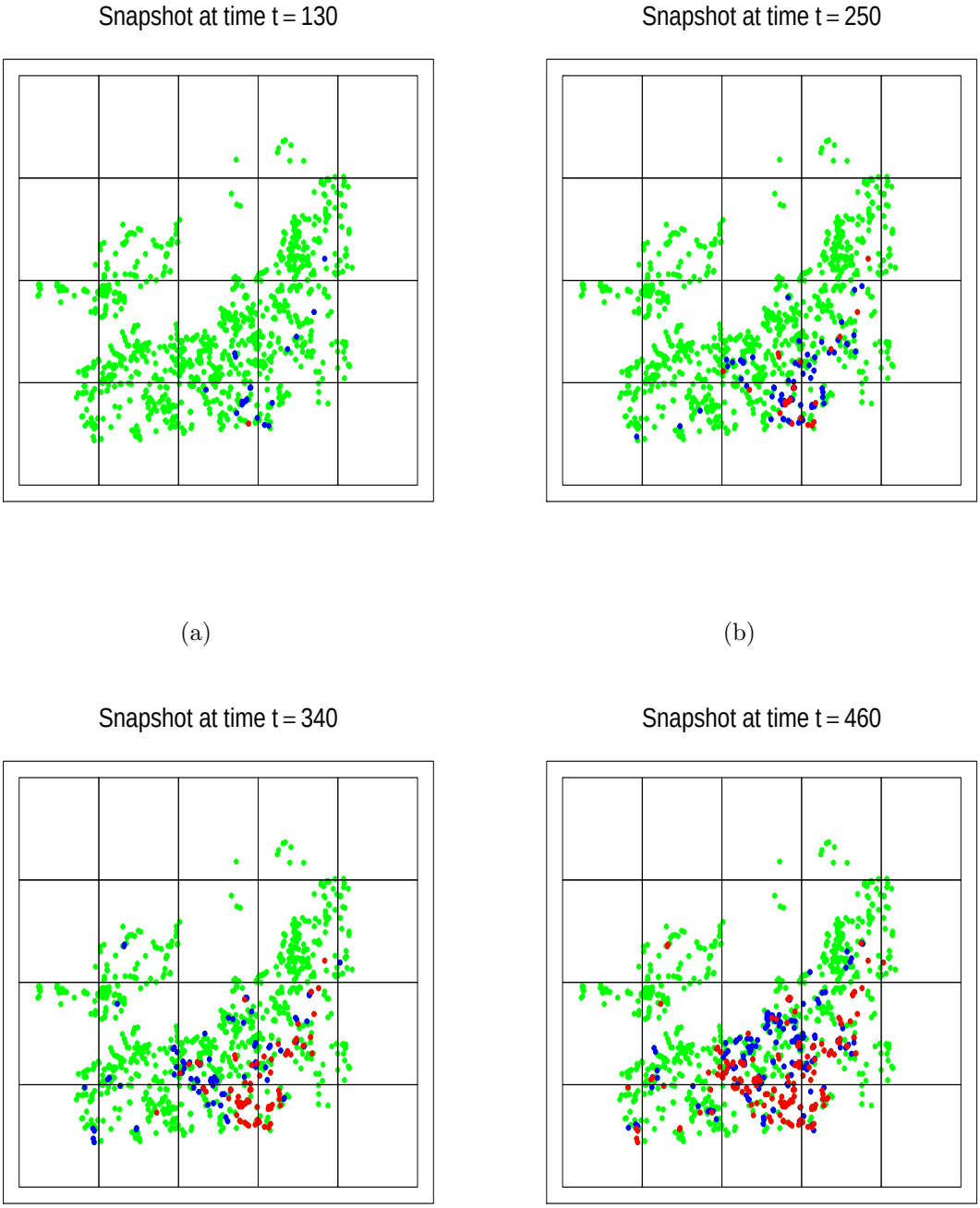


Figure 5.10: Case I: A sample of a realisation of the disease progress maps made at 30-day intervals from $t = 130$ up to $t = 460$, on the citrus population of size $N = 1111$ from a site located in Broward county. Symptomatic hosts, cryptic infections and susceptible hosts at the time of the snapshot are denoted by red, blue and green dots respectively.

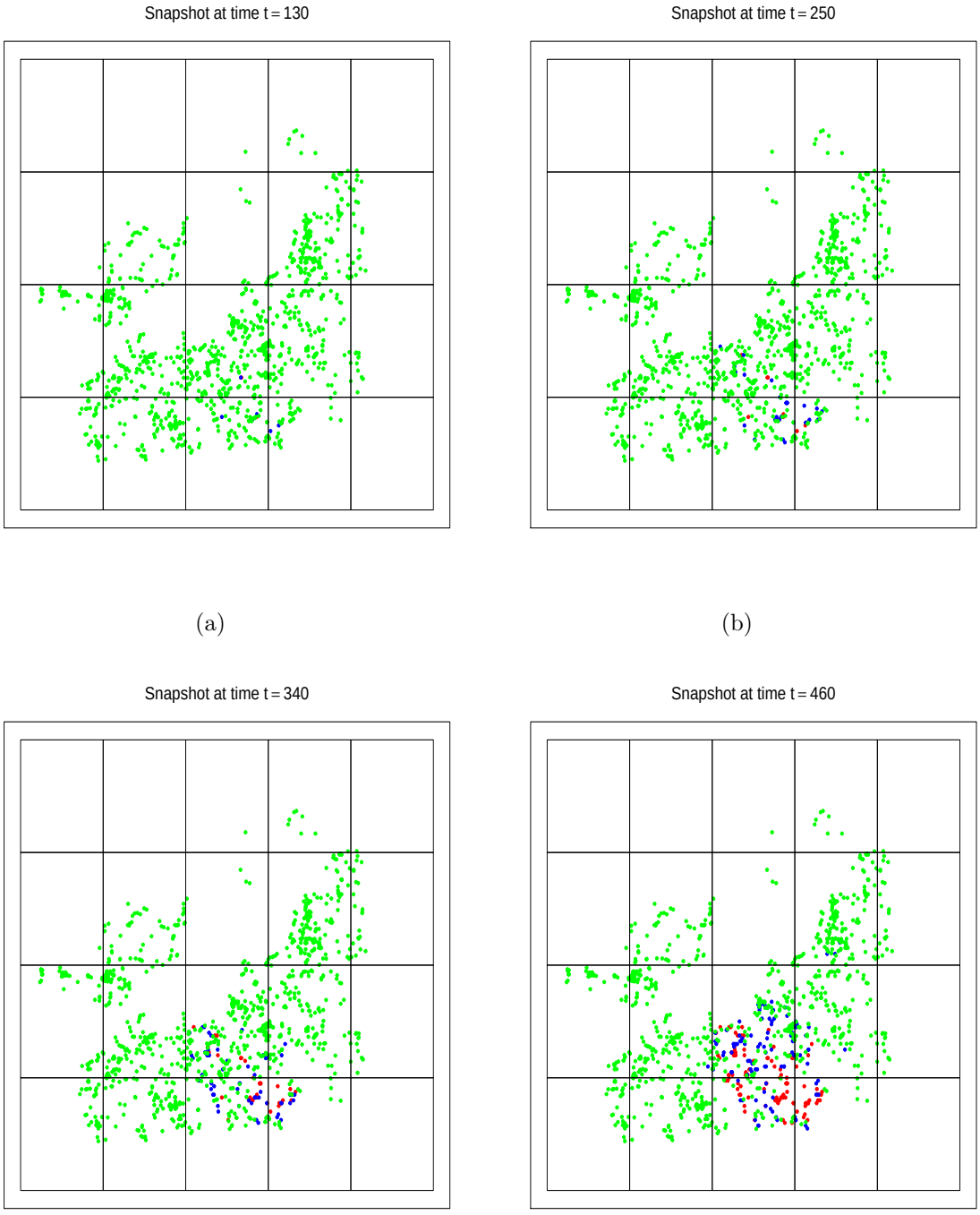


Figure 5.11: Case (II): A sample of a realisation of the disease progress maps made at 30-day intervals from $t = 130$ up to $t = 460$, on the citrus population of size $N = 1111$ from a site located in Broward county. Symptomatic hosts, cryptic infections and susceptible hosts at the time of the snapshot are denoted by red, blue and green dots respectively.

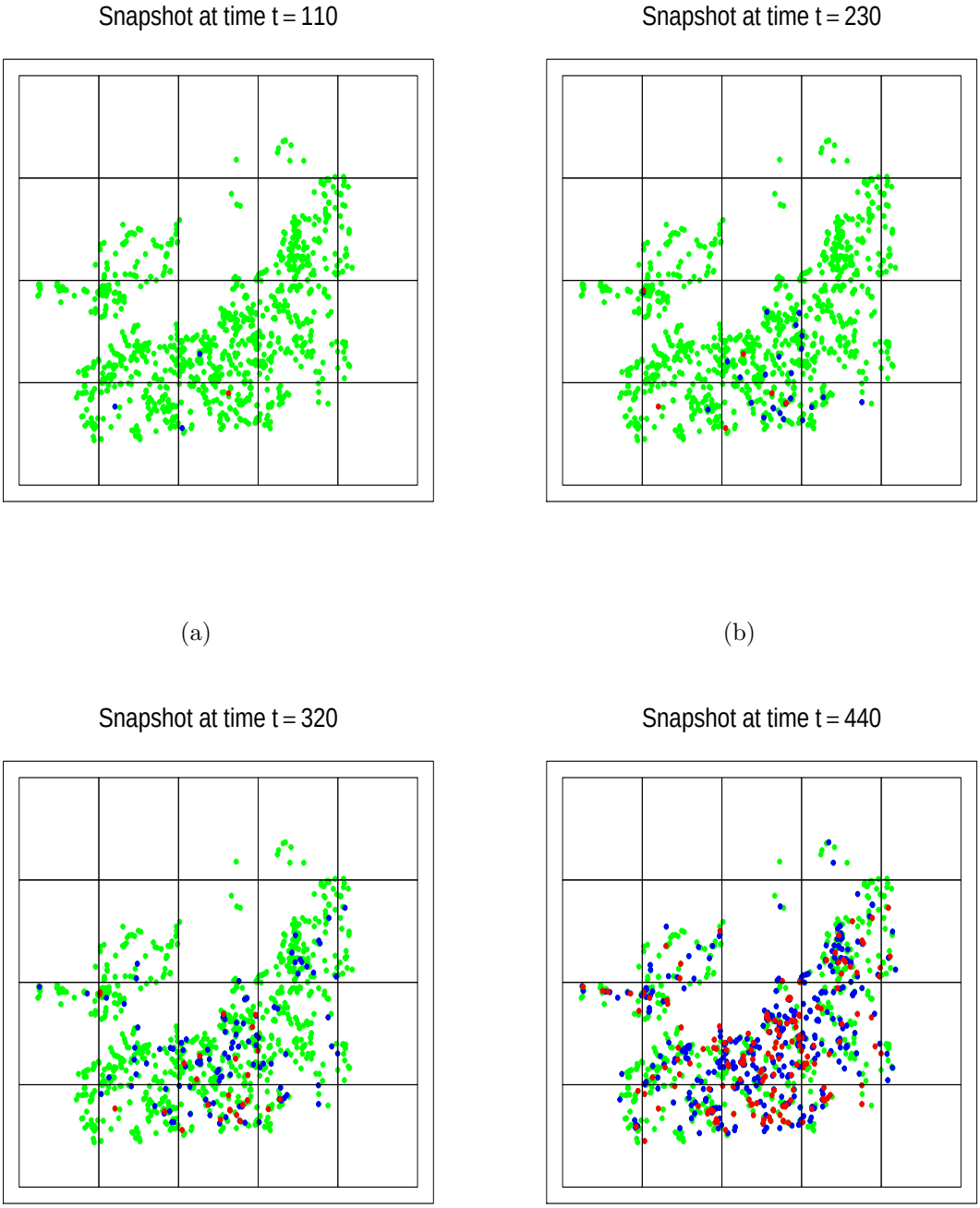
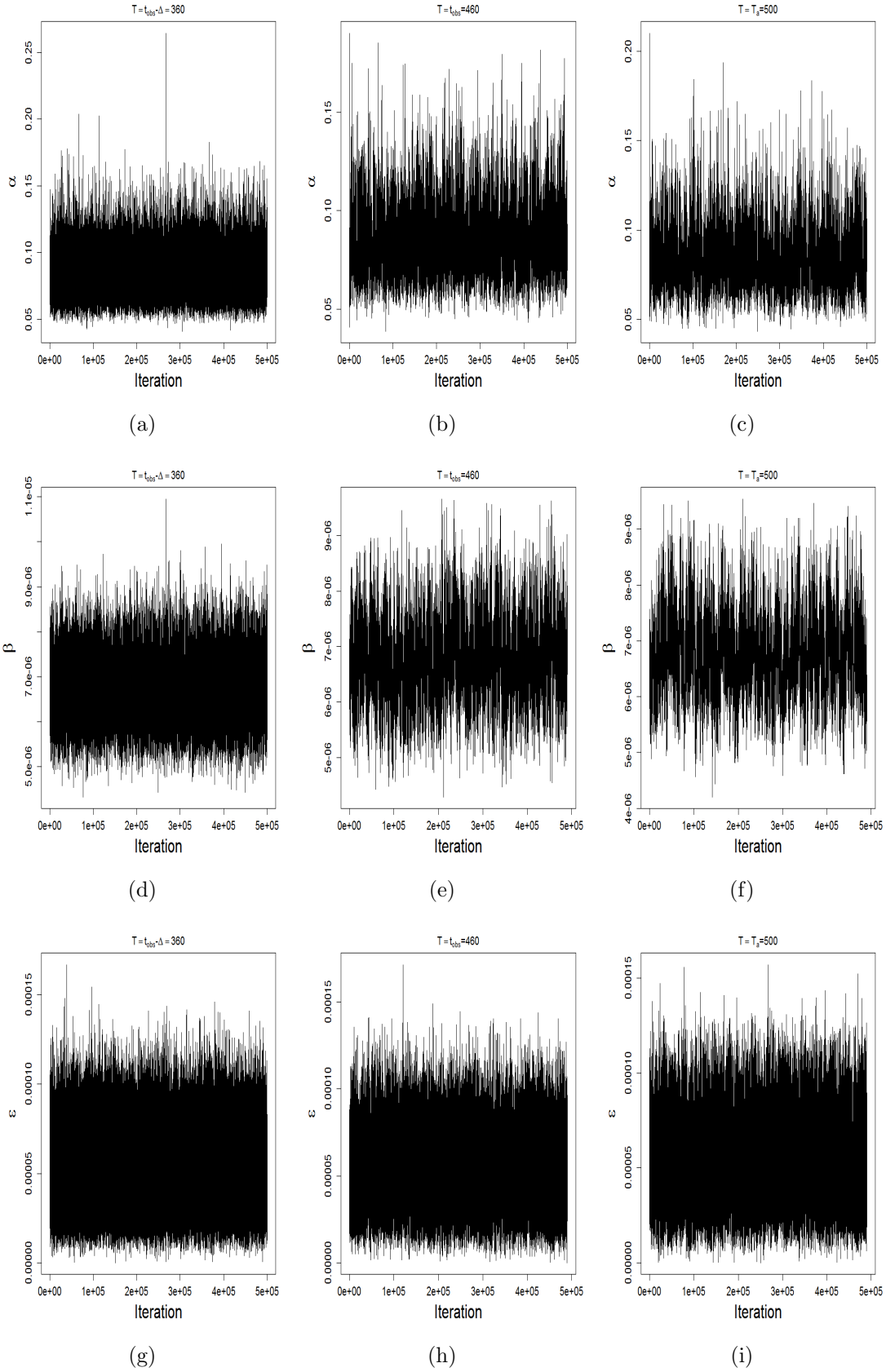


Figure 5.12: Case (III): A sample of a realisation of the disease progress maps made at 30-day intervals from $t = 130$ up to $t = 440$, on the citrus population of size $N = 1111$ from a site located in Broward county. Symptomatic hosts, cryptic infections and susceptible hosts at the time of the snapshot are denoted by red, blue and green dots respectively.

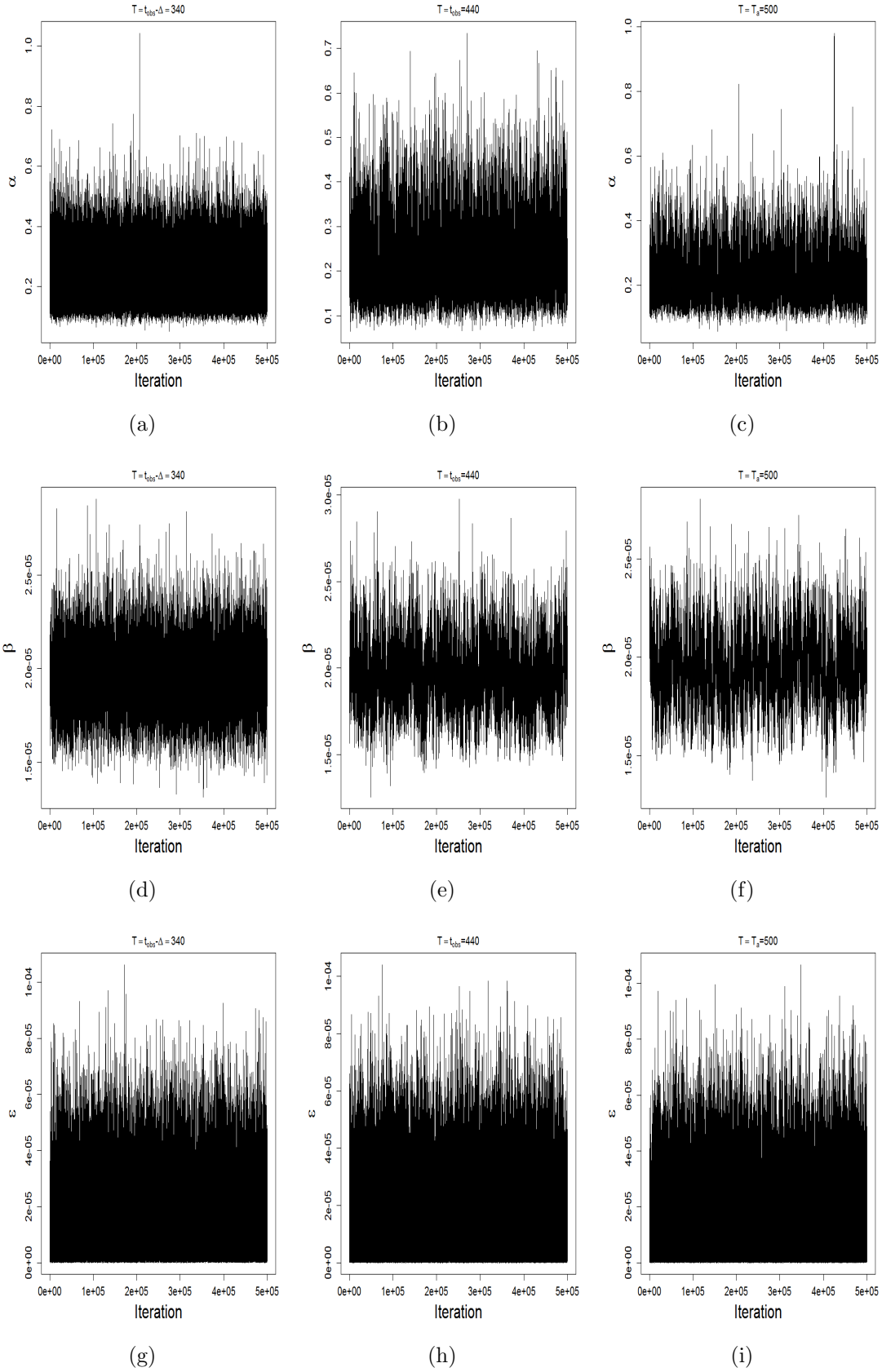


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 5.13: Case (I): Sample trace for the posterior distribution of parameters α β and ϵ after a burn-in period of 10000 iterations using the MCMC algorithms described in Section 5.3.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 5.14: Case (III): Sample trace for the posterior distribution of parameters α β and ϵ after a burn-in period of 10000 iterations using the MCMC algorithms described in Section 5.3.

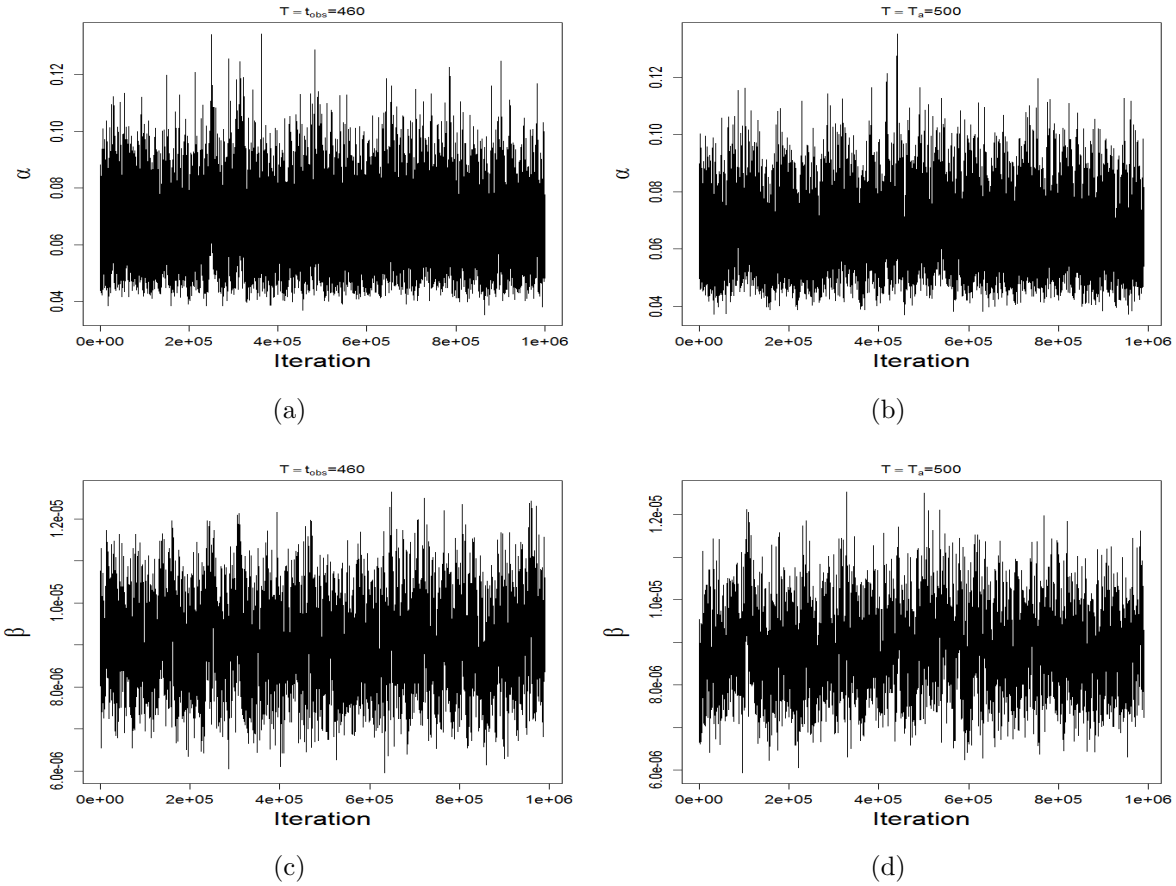


Figure 5.15: Case (II): Sample trace for the posterior distribution of parameters α β after a burn-in period of 10000 iterations using the MCMC algorithms described in Section 5.3.

The posterior distribution of the model parameters α, β and ϵ at various T for Cases (I), (II) and (III) shown respectively in Figures 5.16, 5.17 and 5.18 match regardless of how far we impute infection times beyond t_{obs}. This gives evidence that the algorithm give an accurate picture of the posterior distribution. In addition, it can be seen that the model parameters used for the simulation (dashed lines) are consistent with their respective posterior distributions highlighting the fact that the estimation is good. We estimate the epidemic size considering two different augmentation periods including $T = t_{obs}$ and $T = T_a$ as shown in Table 5.2. Figure 5.16d shows the 95% confidence band of the epidemic trajectory for Case (I). It can be seen that the actual trajectory of the epidemic is contained in the 95% credible regions. Also, the posterior distribution of the predicted epidemic sizes are shown in Figures 5.17c and 5.18d for case (II) and (III) respectively. Again, this suggests that the actual epidemic size lies within the range of the values supported by the predicted distribution. Moreover, we estimate the set of Sellke thresholds as part of the parameter estimation routine in order to compare different control strategies and identify the optimal control strategy conditional on this.

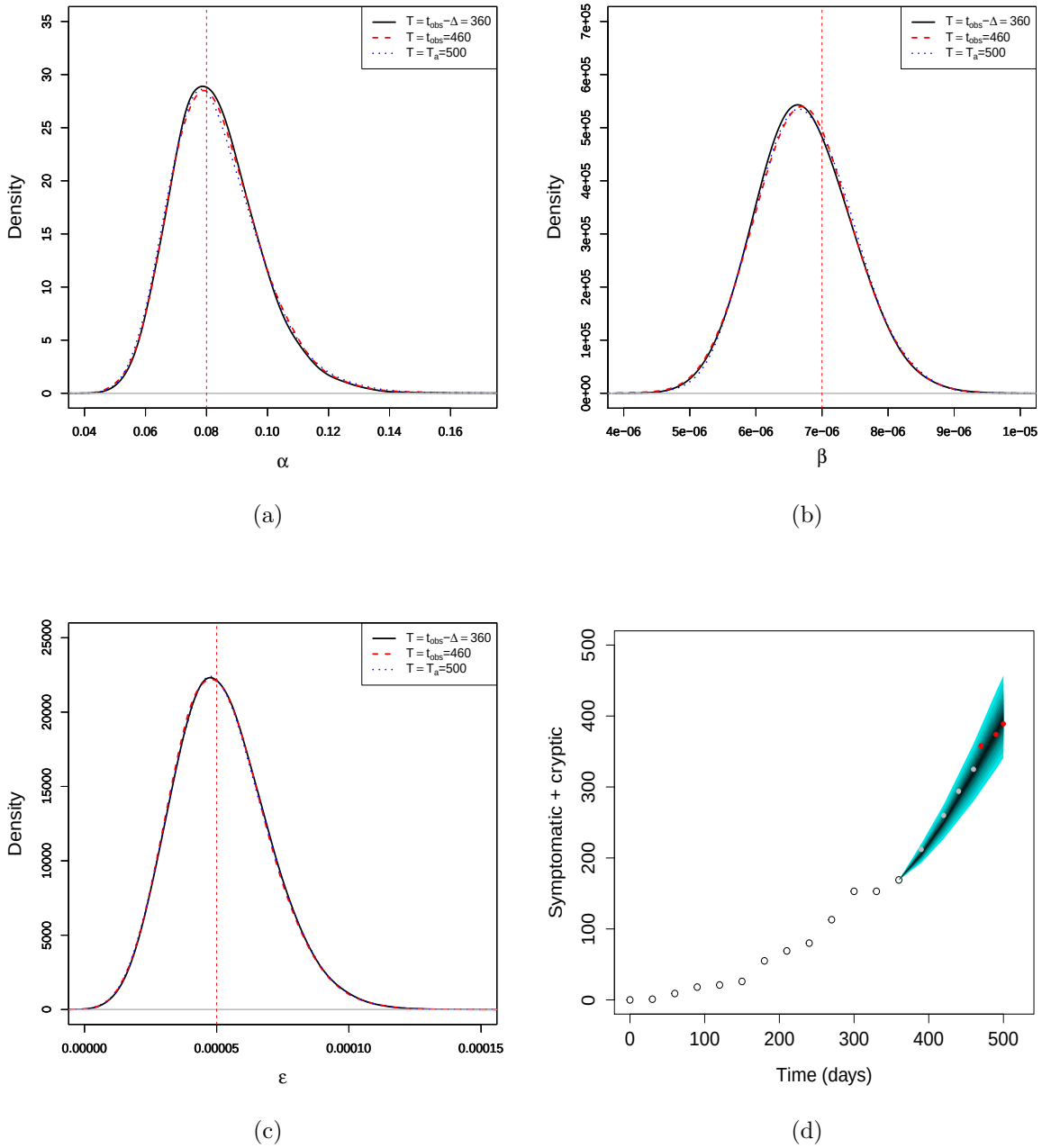
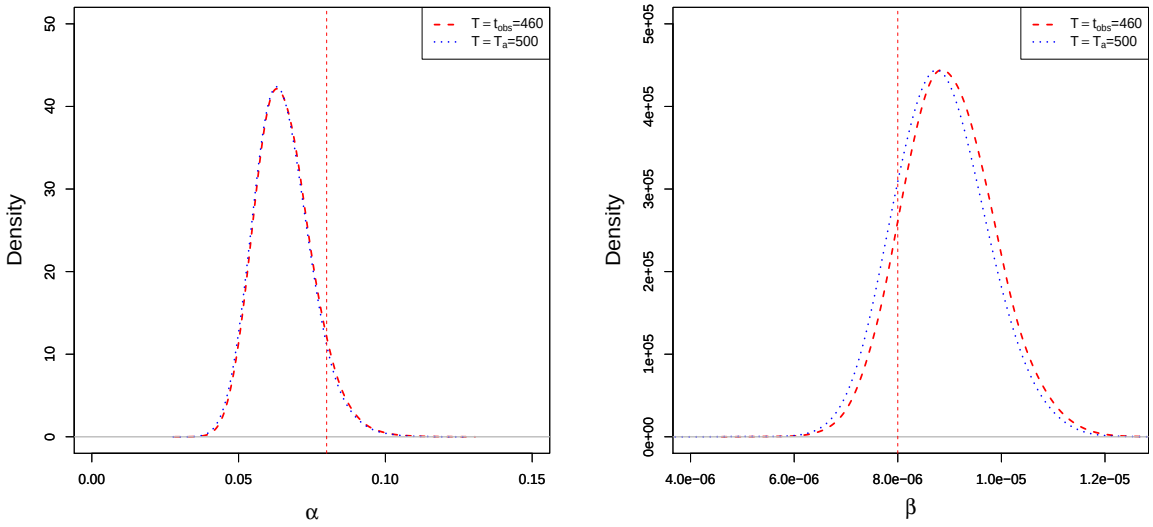
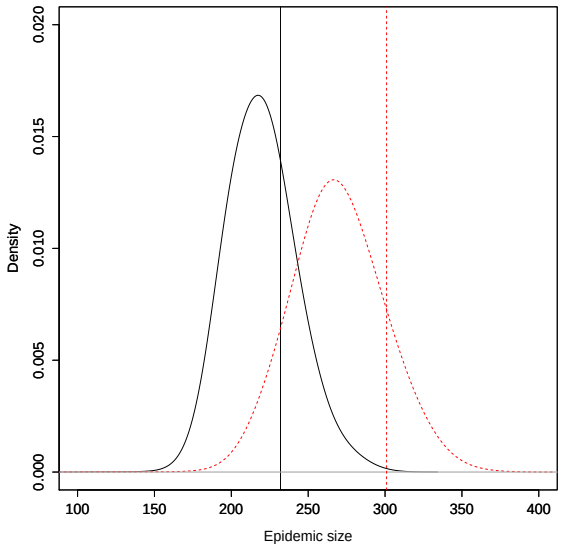


Figure 5.16: Case (I). The posterior distributions for Bayesian MCMC estimation of the model parameters including the dispersal rate α (a), the secondary infection rate β (b), and the primary infection rate ϵ (c), considering $T = 360$ (black full curve), $T = 460$ (red dash curve) and $T = 500$ (blue curve). Dashed lines correspond to the actual value used for the simulation. (d) 95% posterior credible band of the distribution of the disease progress during the period $[360, 500]$ (shaded region) compared to the actual disease progress (circles represent the observed trajectory at $t_{obs} - \Delta$, gray dots represent the unobserved trajectory at t_{obs} and red dots are the future trajectory at T_a).



(a)

(b)



(c)

Figure 5.17: Case (II) The posterior distributions for Bayesian MCMC estimation of the model parameters including the dispersal rate α (a), the secondary infection rate β (b) using $T = 460$ and $T = 500$. Dashed lines correspond to the actual value used for the simulation. (c) The posterior distributions of the epidemic size.

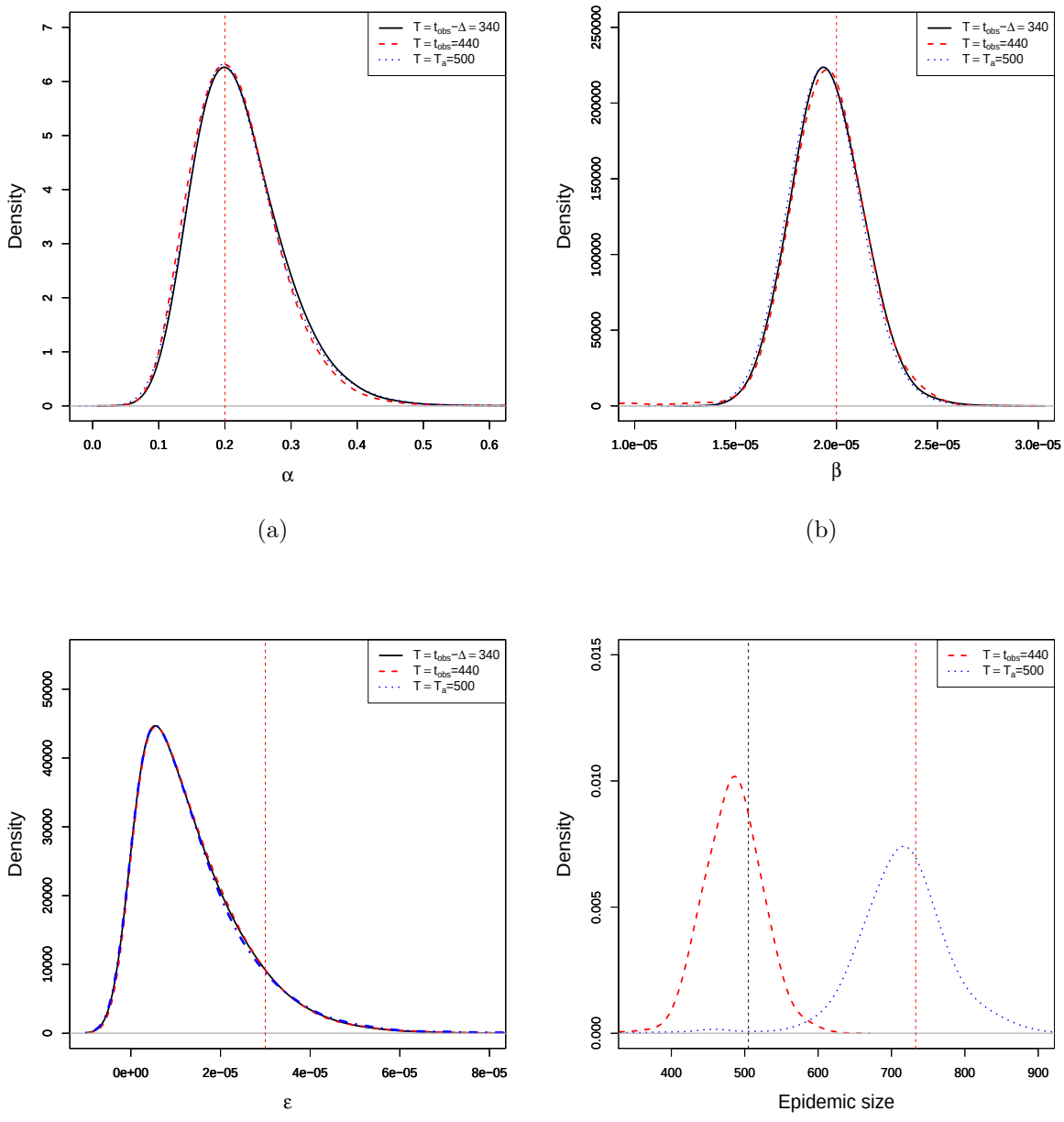


Figure 5.18: Case (III). The posterior distributions for Bayesian MCMC estimation of the model parameters including the dispersal rate α (a), the secondary infection rate β (b) and the primary infection rate ϵ (c) considering $T = 340$ (black full curve), $T = 440$ (red dash curve) and $T = 500$ (blue curve). Dashed lines correspond to the actual value used for the simulation. (c) The posterior distributions of the epidemic considering two different augmentation periods ($T = 440$ and $T = 500$).

5.6.3 Optimisation and results

We show the effectiveness of the control approach developed here on the citrus locations described previously. The intervention time is chosen to be $t_c = 461$ but could

lie anywhere beyond the day we last observe the system. We assume that the region is subdivided into 20 square regions as shown in Figure 5.19. As default, we construct different measures at $T_a = 500$ for the control as shown in Figures 5.19, 5.20 and 5.21 for all different cases and assume that the maximum number of removals $N' = 500$ unless otherwise stated. Although we observe approximatively the same number of symptomatic trees, the exponential kernel gives rise to more localised dispersal than the Cauchy kernel. Note that these prioritisation measures are constructed using 200000 samples from the joint posterior distribution $\pi_0(\underline{x}(T), \boldsymbol{\theta}|\mathbf{y})$ obtained from the MCMC scheme (see Section 5.3.2) where $T = 500$. Samples from the joint posterior distribution of the model parameters and the Sellke-thresholds ($\pi_0(\boldsymbol{\theta}, \boldsymbol{Q}|\mathbf{y})$) are then generated as part of the MCMC routine.

We sample $\{(\boldsymbol{Q}^i, \boldsymbol{\theta}^i)|i = 1, \ldots, 1000\}$, realisations from $\pi_0(\boldsymbol{\theta}, \boldsymbol{Q}|\mathbf{y})$ which are then used to design control strategies. We consider control strategies of the form $d_{\{\mathcal{M},2,1\}}$. We only describe the result for case (I), but the descriptions are qualitatively similar in the other two cases. Using the threat map $\mathcal{T}^*$, after 20 realisations, the Algorithm 2 (see Section 5.4) converges. The optimal control strategies using different measures are shown in Table 5.3. However, the optimal control obtained using $\mathcal{T}^*$ is $d^*_{\{\mathcal{T}^*,2,1\}} = (3; 166)$, which means that to maximize the impact of the epidemic conditioned on the threat map in the Figure 5.19c, the intervention must be deployed in the 3 highest ranked regions out of 20 where 166 hosts must be visited for control in each. Recall that the regions are ranked in term of w_i (see Equation (5.4.1)), the average value of the respective measure over the region i. The corresponding estimated distribution of the number of infections avoided $\{u\left(\underline{x}(\boldsymbol{Q}^i, \boldsymbol{\theta}^i, T)\right)\text{-}u\left(\underline{x}(\boldsymbol{Q}^i, \boldsymbol{\theta}^i, \boldsymbol{d}, T)\right)|i = 1, \ldots, 1000\}$ where $\boldsymbol{d} = d^*_{\{\mathcal{T}^*,2,1\}}$ is shown in Figure 5.25 (black curve), leading to an average of 45 infections avoided with this control. The latter represents about 57% of the actual number of infections (without control) as shown in Table 5.3. This highlight the fact that regions with high threat are regions where susceptibles are mixed with few infected or are close to densely infected regions.

Furthermore, we compare the impact of the control strategy $d^*_{\{\mathcal{T}^*,2,1\}}$ to that of $d^*_{\{\mathcal{R}^*,2,1\}}$, $d^*_{\{\mathcal{C}^*,2,1\}}$ and $d^*_{\{\mathcal{U}^*,2,1\}}$ on the same realisations $(\boldsymbol{Q}^i, \boldsymbol{\theta}^i)$, $i = 1, \ldots, 1000$. The distributions $\{u\left(\underline{x}(\boldsymbol{Q}^i, \boldsymbol{\theta}^i, T)\right)\text{-}u\left(\underline{x}(\boldsymbol{Q}^i, \boldsymbol{\theta}^i, \boldsymbol{d}, T)\right)|i = 1, \ldots, 1000\}$ where $\boldsymbol{d} \in \left\{d^*_{\{\mathcal{T}^*,2,1\}}, d^*_{\{\mathcal{U}^*,2,1\}}, d^*_{\{\mathcal{R}^*,2,1\}}, d^*_{\{\mathcal{C}^*,2,1\}}\right\}$ are shown in Figure 5.25. Unsurprisingly, using $\mathcal{R}^*$ or $\mathcal{C}^*$ as the prioritisation measures of intervention is less cost-effective in reducing the number of infections. This is simply due to the fact that the regions with high risk are those where the majority of individuals are already infected. Therefore the removed individuals may be surrounded by hosts already infected. This would be cost-effective if the number of regions to visit along with the individuals to remove are sufficiently large so that most of the infected hosts can be removed. This measure will be more appropriate when attempting to design surveillance with the goal

of detecting

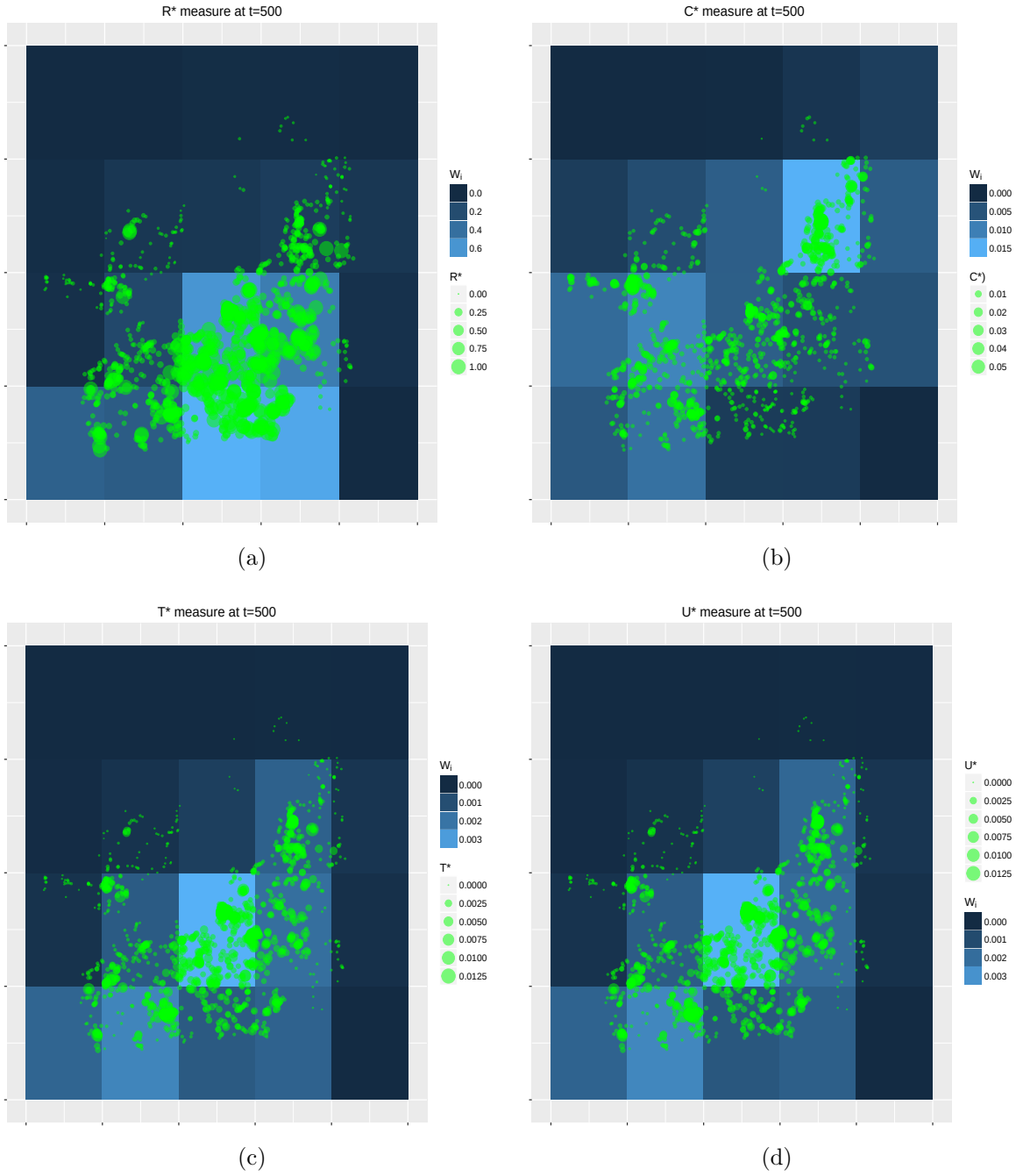
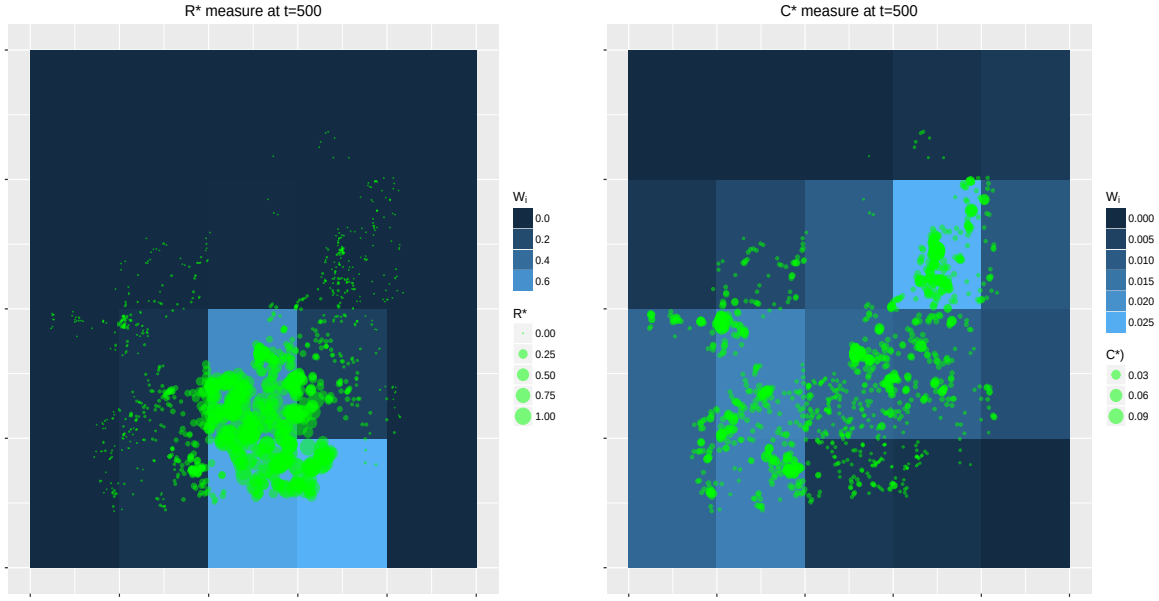
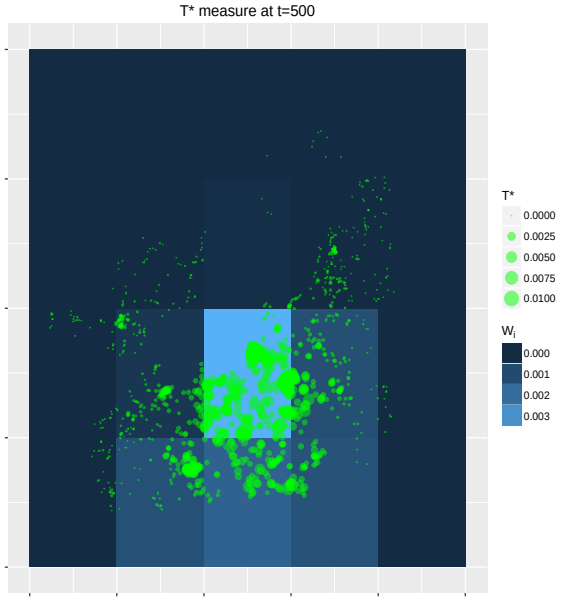


Figure 5.19: case (I): Plots showing the posterior expectation of the risk $\mathcal{R}^*$ (a), posterior expectation of the challenge $\mathcal{C}^*$ (b), the threat $\mathcal{T}^*$(c) and the product of the expectations $\mathcal{U}^*$ (d) maps constructed on the assessment time $T = 500$.

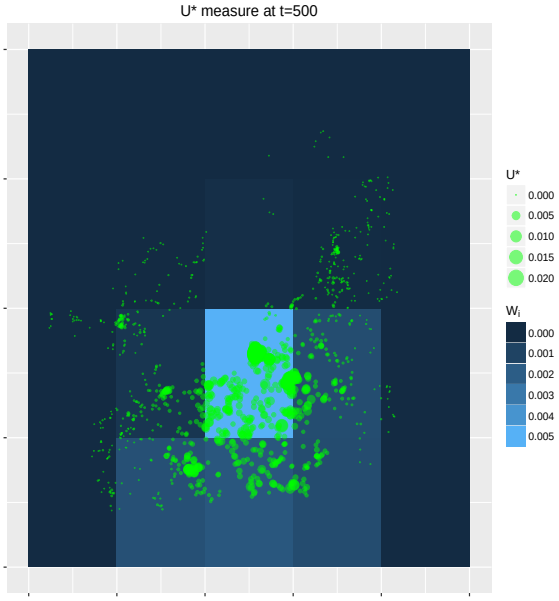


(a)

(b)



(c)



(d)

Figure 5.20: case (II): Plots showing the posterior expectation of the risk $\mathcal{R}^*$ (a), posterior expectation of the challenge $\mathcal{C}^*$ (b), the threat $\mathcal{T}^*$(c) and the product of the expectations $\mathcal{U}^*$ (d) maps constructed on the assessment time $T_a = 500$.

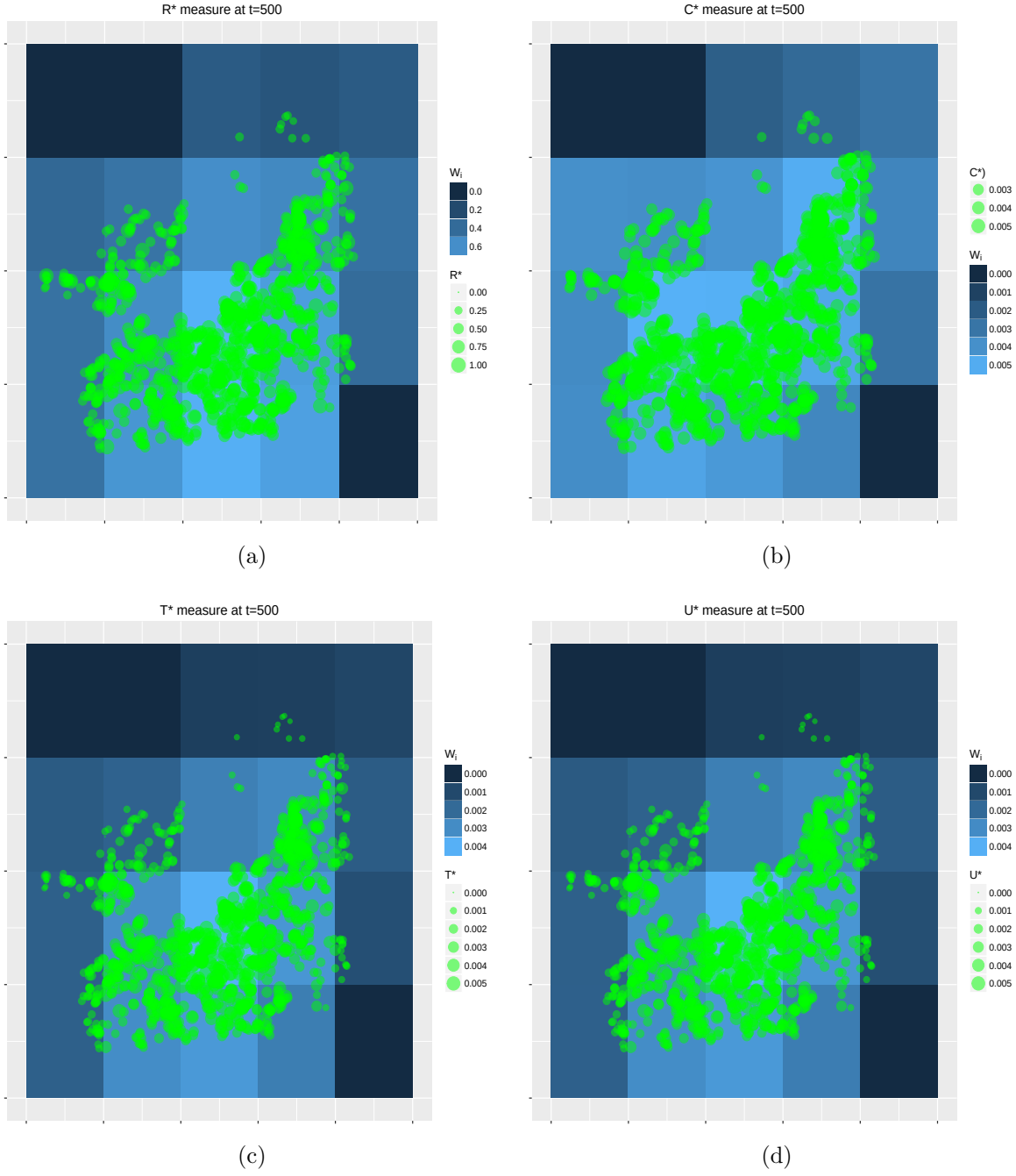
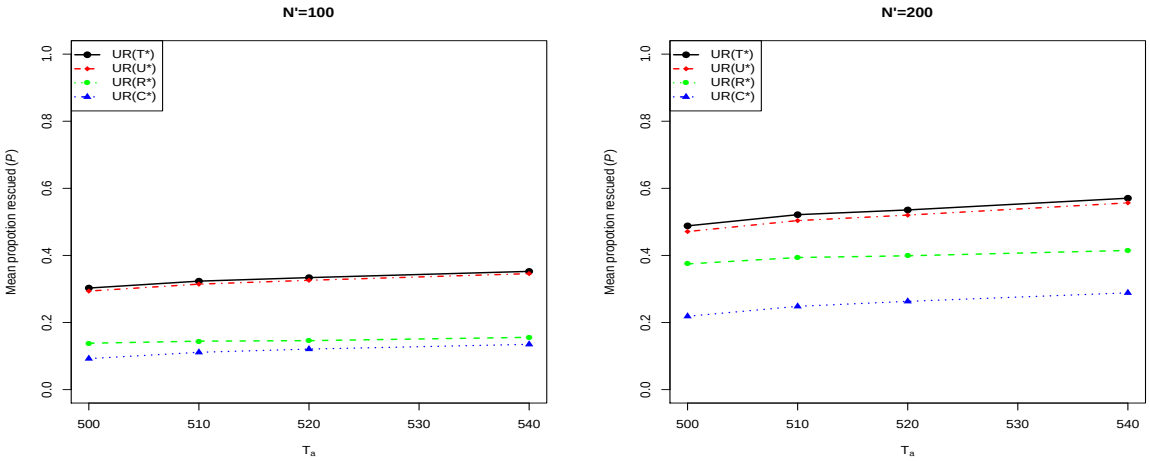
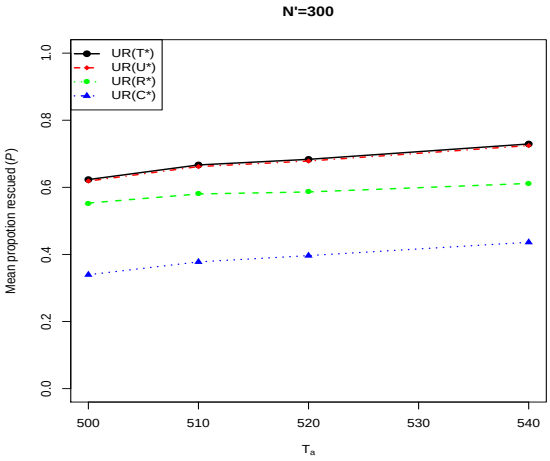


Figure 5.21: Case (III): Plots showing the posterior expectation of the risk $\mathcal{R}^*$ (a), posterior expectation of the challenge $\mathcal{C}^*$ (b), the threat $\mathcal{T}^*$(c) and the product of the expectations $\mathcal{U}^*$ (d) maps constructed on the assessment time $T_a = 500$.

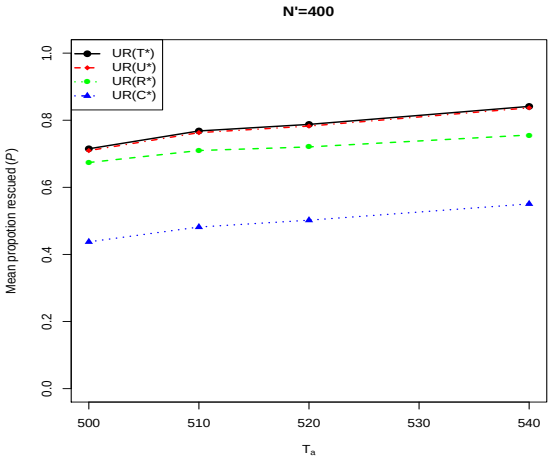


(a)

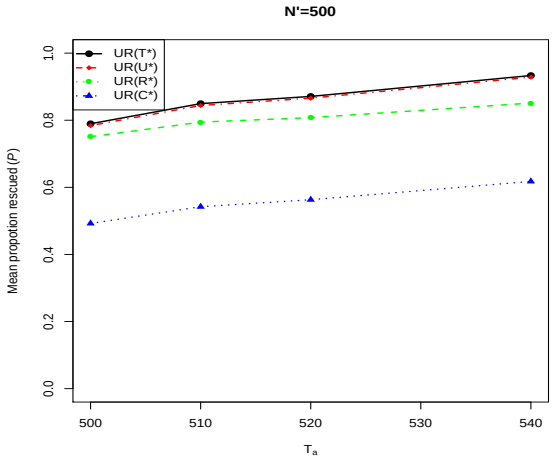
(b)



(c)



(d)



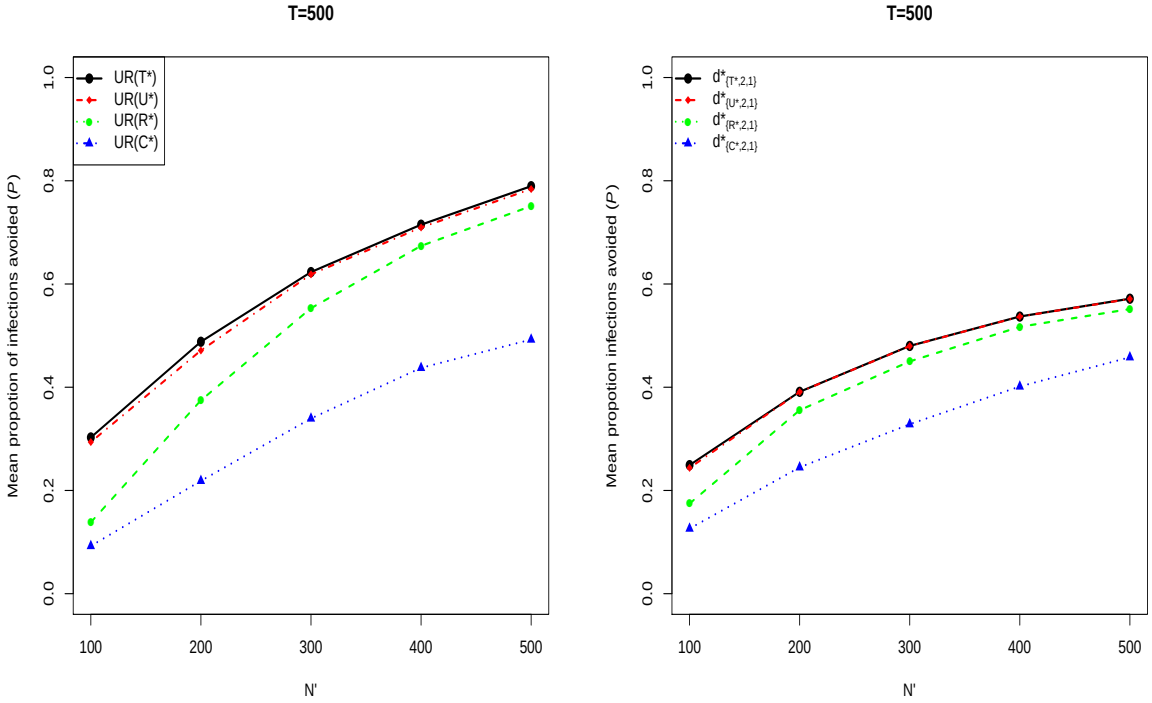
(e)

Figure 5.22: Effect of various control strategies (with no regional constraints) on the relationship between the assessment time T_a and the posterior mean of the number of infection avoided, $\mathcal{P}$, when deploying resources at $t_c = 461$ with $N' = 100, 200, 300, 400, 500$, given that the maps are constructed on the assessment time $T_a = 500$ using simulated data on citrus locations.

infected individuals. Figures 5.25, 5.24a and 5.24c revealed that infection avoided when deploying control optimally based on the threat map $\mathcal{T}^*$ out-performed other measures; though there is not much difference between control using measures $\mathcal{T}^*$ and $\mathcal{U}^*$. This is emphasized by the summary statistic of the proportion $\mathcal{P}$ of infections avoided shown in Table 5.3. It can be seen that the 95% credible intervals do not differ significantly from each other.

Strategy	Optimal control $(n_r; n)$	Mean	SD	0.025	Median	0.975
		$460-500$				
$d(\mathcal{C}^*,2,1)$	(5;100)	0.458	0.074	0.318	0.457	0.603
$d(\mathcal{R}^*,2,1)$	(7;71)	0.551	0.080	0.392	0.553	0.712
$d(\mathcal{U}^*,2,1)$	(3;166)	0.571	0.073	0.435	0.565	0.716
$d(\mathcal{T}^*,2,1)$	(3;166)	0.572	0.073	0.435	0.566	0.716
$UR(\mathcal{T}^*)$	-	0.789	0.059	0.667	0.791	0.9

Table 5.3: Summary statistics for the ratio $\mathcal{P}$ of the number of infection avoided using different interventions on day $t_c = 461$ relative to the actual number of infection during the period $460-500$, using different prioritisation measures. we refer the reader to 5.6 for the notation adopted here for the strategies.



(a) (b)

Figure 5.23: N' versus $E(\mathcal{P}|\mathbf{y})$ for various control strategies (a) unconstrained regions(b)Constrained regions.

We assess the performance of our proposed design by comparing the distribution

of the number of infections avoided

$$\{u\left(\underline{x}(\boldsymbol{Q}^i, \boldsymbol{\theta}^i, T_a)\right) - u\left(\underline{x}(\boldsymbol{Q}^i, \boldsymbol{\theta}^i, \boldsymbol{d}, T_a)\right) | i = 1, \ldots, 1000\}$$

using the control strategy $d^*_{\{\mathcal{T}^*,2,1\}}$, with the distribution of the outcome of the control strategy $d^*_{\{\mathcal{T}^*,3,1\}}$ (control obtained by optimising for each realisation $(\boldsymbol{Q}^i, \boldsymbol{\theta}^i)$) on the same 1000 realisations. The former distribution is shown in black and the latter distribution in yellow on Figure 5.25, 5.24a and 5.24c. By doing so, we hope to show that our optimal strategy is not far away from the global optimum. Results show that the two distributions differ little from each other, highlighting the effectiveness of our proposed approach.

In addition, we select N' hosts in the ordering with respect to their level of threat leading to the control strategy $UR(\mathcal{T}^*)$. Note that we employ the same technique for the control in that hosts are removed when diagnosed positive or symptomatic. We can see from Figure 5.25, 5.24a and 5.24c that the latter design provides a better control. This is unsurprising, given that the method will result in detection of hosts with high level of threat compared to the former method where *regions* with high threat are identified. These regions do not necessary coincide with regions comprised of hosts with the highest threat in the system. However, in practice it may be more convenient to adopt a strategy whereby the control team are sent to different areas (e.g. villages, farms etc), so that the number of regions that can be visited is constrained.

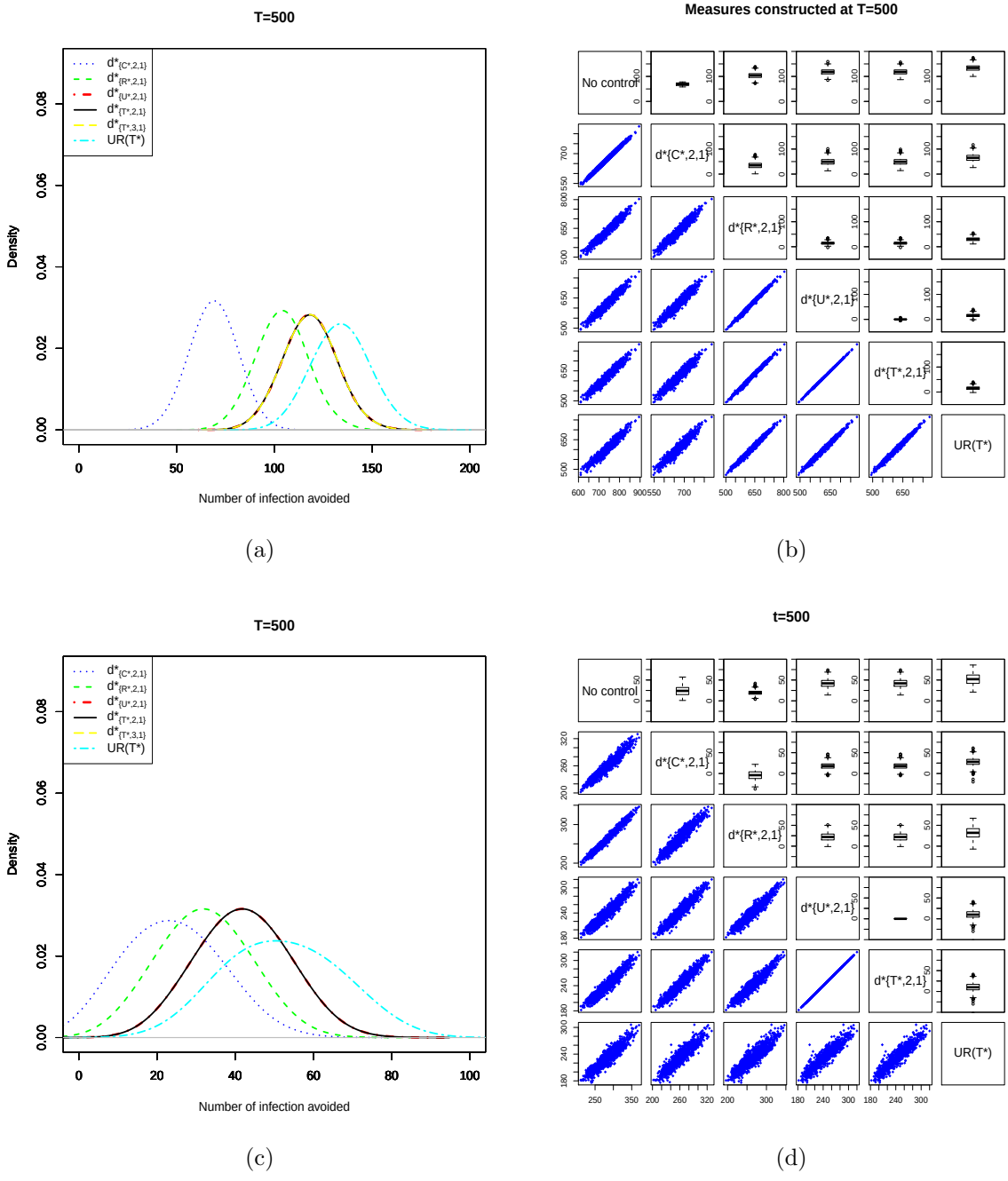
We display the joint posterior samples

$$\left(u\left(\underline{x}(\boldsymbol{Q}^i, \boldsymbol{\theta}^i, \boldsymbol{d}, T)\right), u\left(\underline{x}(\boldsymbol{Q}^i, \boldsymbol{\theta}^i, \boldsymbol{d}', T)\right)\right),\ i = 1, \ldots, 1000$$

for $\boldsymbol{d}$, $\boldsymbol{d}' \in \{\text{No control}, d^*_{\{\mathcal{C}^*,2,1\}}, d^*_{\{\mathcal{R}^*,2,1\}}, d^*_{\{\mathcal{U}^*,2,1\}}, d^*_{\{\mathcal{T}^*,2,1\}}, UR(\mathcal{T}^*)\}$, of the epidemic size on the assessment times $T_a = 500$ days in figure 5.26, 5.24b, 5.24d. Similar to the uniformly distributed population case, we can see a strong positive correlation between the outcomes of the epidemic for different controls. This results in the desired reduction in the variance of the difference between outcomes of the controls compared to independence sampling. Therefore, the coupling approach provides a more accurate estimate of the expected effect of different controls.

We also determine the dependence of the posterior mean proportion of infections avoided on both assessment time T_a and N'. Unsurprisingly, it can be seen in Figure 5.23a that, as N' increases, performance for prioritisation based on the risk map approaches that of the threat map. This observation is mainly due to the fact that more infected hosts are likely to be removed during the intervention, reducing the infection pressure on the remaining susceptibles. However, as the time that elapses between the assessment time and the time at which the maps are constructed increases, the outcome of the control strategy based on the risk map diverges from that

of the threat map as shown in Figure 5.23. The reason being that some infected hosts that present high challenge are not prioritised and therefore are not removed during the intervention based on the risk map causing future outbreak as a consequence.



(a) (b)

(c) (d)

Figure 5.24: Case (III) (a), (b) and Case (II) (c), (d). (a)(c) Posterior distribution of the number of infections avoided $\{u\left(\underline{x}(\boldsymbol{Q}^i,\boldsymbol{\theta}^i,T)\right)\text{-}u\left(\underline{x}(\boldsymbol{Q}^i,\boldsymbol{\theta}^i,\boldsymbol{d},T)\right)\,|\,i=1,\ldots,1000\}$ and (b)(d) pairwise joint posterior samples $\{(u\left(\underline{x}(\boldsymbol{Q}^i,\boldsymbol{\theta}^i,\boldsymbol{d},T)\right),\ u\left(\underline{x}(\boldsymbol{Q}^i,\boldsymbol{\theta}^i,\boldsymbol{d}',T)\right)),\ i=1,\ldots,1000\}$ on the lower panel and the distribution of the difference between the outcomes of the paired controls $\{u\left(\underline{x}(\boldsymbol{Q}^i,\boldsymbol{\theta}^i,T)\right)\text{-}u\left(\underline{x}(\boldsymbol{Q}^i,\boldsymbol{\theta}^i,\boldsymbol{d},T)\right)\,|\,i=1,\ldots,1000\}$ for $\boldsymbol{d}\neq d'$ when deploying resources at $t_c=461$, given that the maps are constructed at $T=500$.

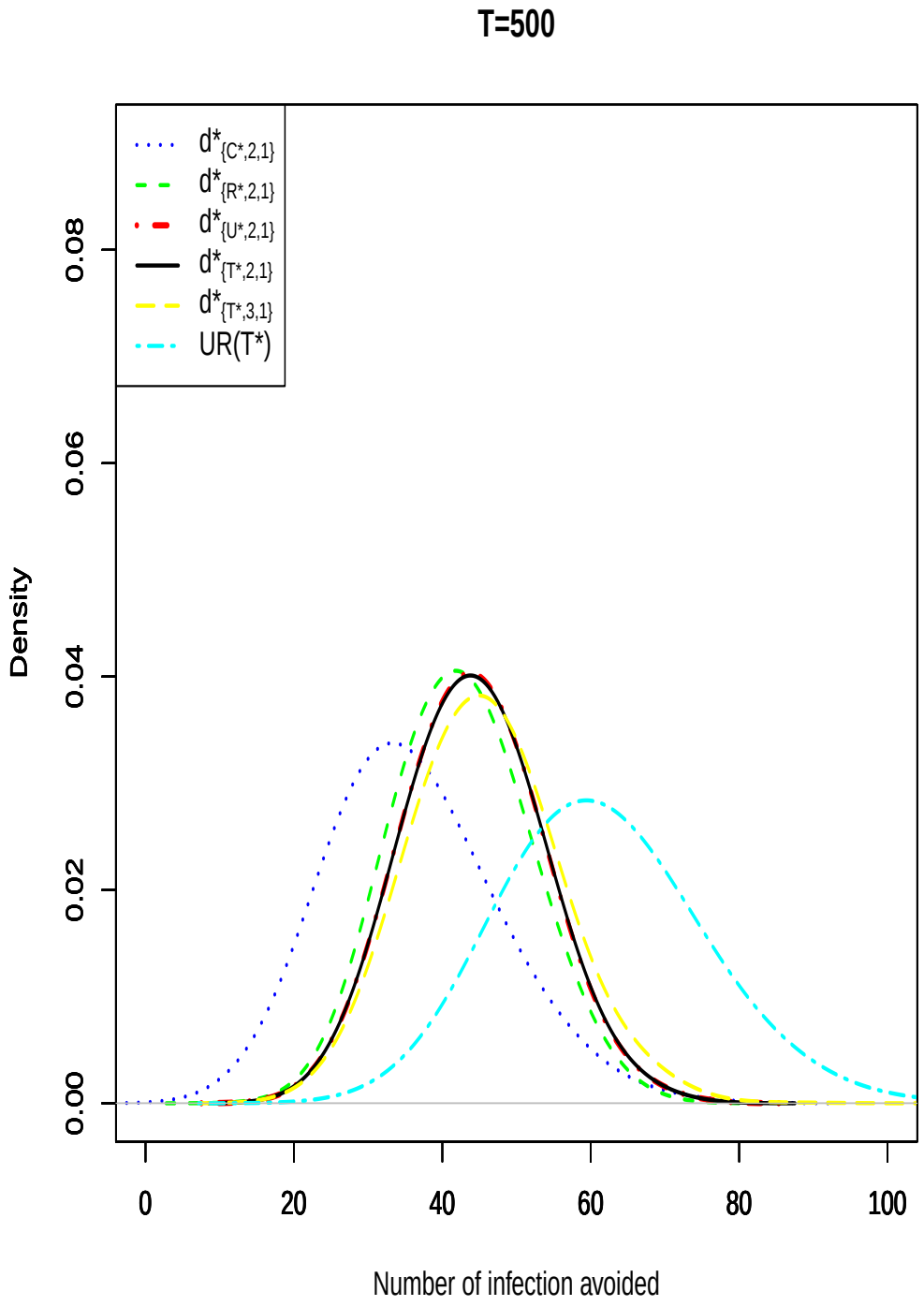
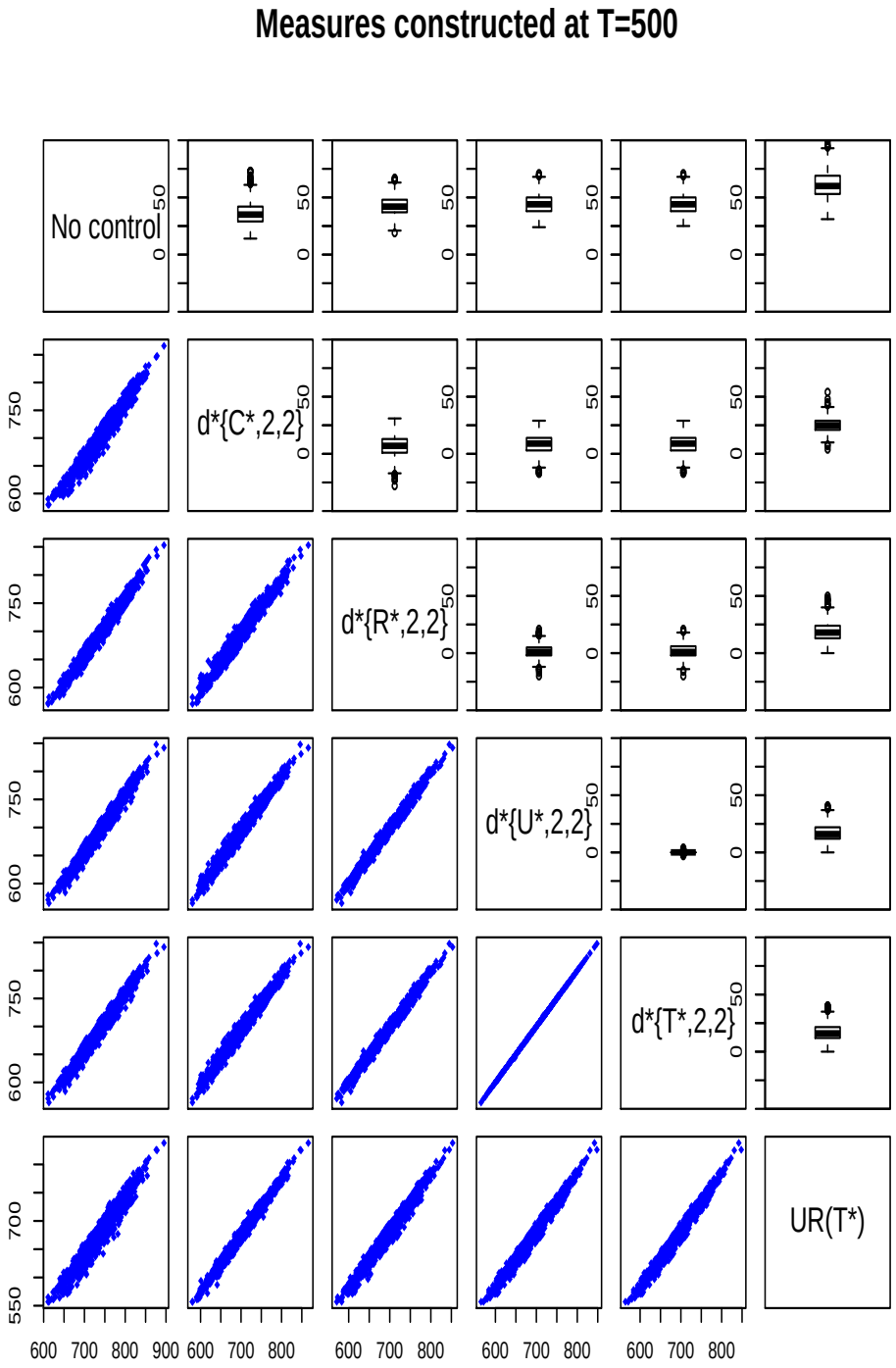


Figure 5.25: Case (I): Posterior distribution of the number of infections avoided where d takes respectively $d^*_{\{\mathcal{C}^*,2,1\}}$, $d^*_{\{\mathcal{R}^*,2,1\}}$, $d^*_{\{\mathcal{U}^*,2,1\}}$, $d^*_{\{\mathcal{T}^*,2,1\}}$, and $d^*_{\{\mathcal{C}^*,2,1\}}$ where the maps are constructed at the assessment day $T_a = 500$.



(a)

Figure 5.26: Case (I): Pairwise joint posterior samples $\{(u\,(\underline{x}(\boldsymbol{Q}^i, \boldsymbol{\theta}^i, \boldsymbol{d}, T))\,,\ u\,(\underline{x}(\boldsymbol{Q}^i, \boldsymbol{\theta}^i, \boldsymbol{d}'\ \ , T))),\ i = 1, \ldots, 1000\}$ and the distribution of the difference between these outcomes for d, $d' \in \left\{\text{no control}, d^*_{\{\mathcal{C}^*,2,1\}}, d^*_{\{\mathcal{R}^*,2,1\}},\ d^*_{\{\mathcal{U}^*,2,1\}},\ d^*_{\{\mathcal{T}^*,2,1\}},\ d^*_{\{\mathcal{C}^*,2,1\}}\right\}$ when deploying resources at $t_c = 461$, given that the maps are constructed at the assessemement day $T_a = 500$.

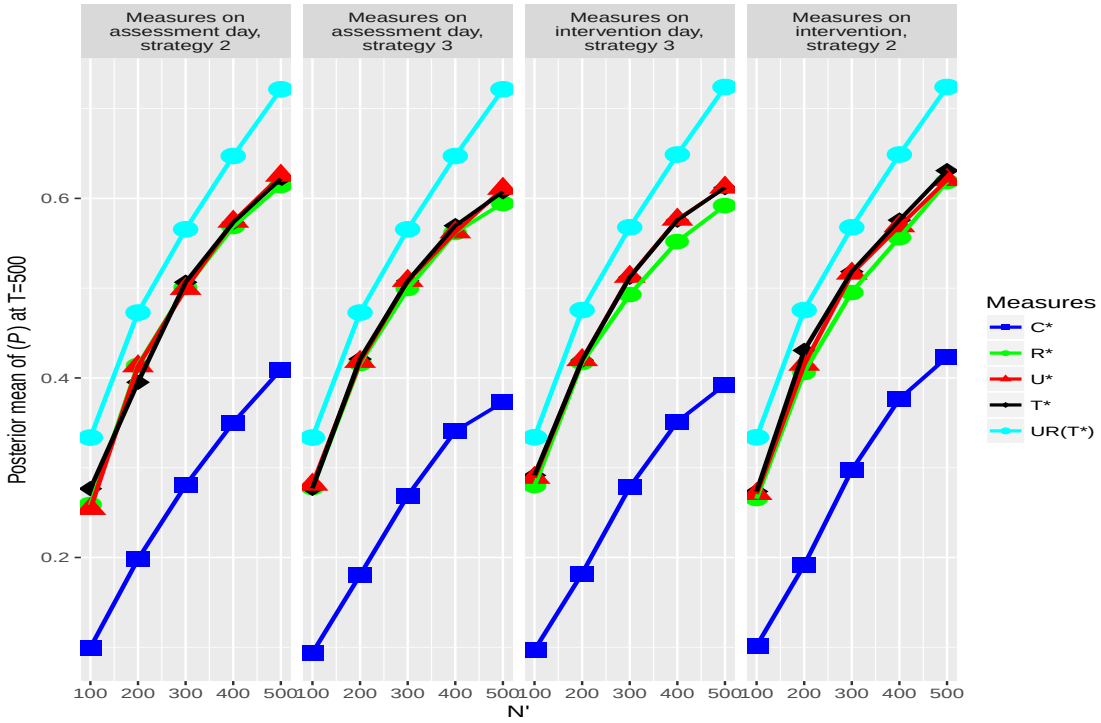
5.6.4 Effect of time of prioritisation map

So far, all controls are conditioned on maps constructed on the assessment time. However, an interesting question when deploying controls or when directing surveillance effort based on map generated during an outbreak would typically be either to construct the map on the intervention time (see for example the Scotland risk map generated in September 2007 (Woolhouse, 2011) during the FMD outbreak in Surrey in England (Bessell, 2009)) or on the assessment date (DEFRA, 2013).

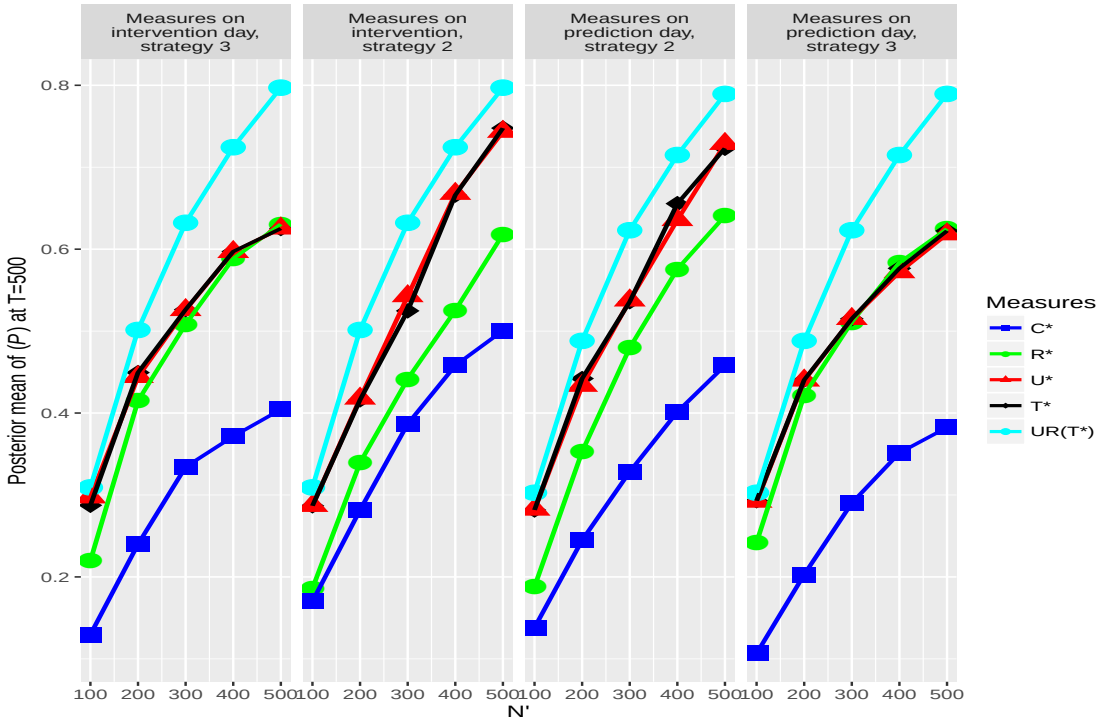
Alternative maps are therefore constructed at $t_c = 461$ for data simulated for uniformly distributed hosts and for hosts with the citrus locations. Again we run the optimisation with varying number of hosts to sample for the control ($N' = 100, 200, 300, 400, 500$) but in this case we adopt the control strategies of the form $d^*_{\{\mathcal{M},2,2\}}$ and $d^*_{\{\mathcal{M},2,3\}}$. Note that we constrain the number of regions to visit $n_r \leq C$ where $C = 50$ for the uniformly distributed population and $C = 10$ for the citrus locations case. These control strategies are obtained using $\{(\boldsymbol{\theta}^i, \boldsymbol{Q}^i)|i = 1, \ldots, 20\}$.

Figures 5.27a and 5.27b reveal that there is no clear difference between controls using maps constructed at the intervention date and the assessment date even though we observe some little variation when the control is conditioned on the risk and the challenge maps. It appears from the two graphs that it is more cost-effective to base a control on the challenge map constructed on the intervention date than one computed on the prediction date. This is due to the fact that hosts removed in the latter case are from regions where hosts are either clustered with some susceptibles and isolated from the rest of the system or are surrounded by susceptibles which are in their turn clustered or surrounded by infectives. Therefore, they constitute a barrier for disease transmission.

In addition, from the Figure 5.27b, the strategy 2 typically outperforms strategy 3, while the difference in Figure 5.27a is more evident when the control is conditioned on the risk and the challenge maps. This observation shows that host distribution impacts on the relative efficacy of different approaches to control, with the heterogeneity of the citrus data accentuating the superiority of the threat map as a tool to guide the design of control.



(a)



(b)

Figure 5.27: Effect of the control strategies used and the maps constructed at $t_c = 461$ and $T = 500$ on the relationship between the proportion of infections avoided $\mathcal{P}$ and the maximum number of hosts visited (N'). (a) Uniformly distributed host (b) the citrus location data.

5.7 Conclusion

The removal of infected hosts during the course of an epidemic is considered as the most efficient control strategy to halt epidemics of highly infectious diseases (Cook et al., 2008; Cunniffe et al., 2014). This chapter introduces a prospective approach to target control of highly infectious disease with spatial dynamical transmission. It introduces a prioritisation scheme based on the idea that hosts with the highest threat defined as the expected product of risk of being infected and the challenge presented if infected, should be considered for removal first under the control.

The first important feature of our approach is that we have embedded it entirely in the Bayesian framework. The second important feature of our approach is the use of latent processes to couple epidemics. By so doing, we preserve the patterns inherited from the data (epidemic trajectory up to the final observation time) when comparing different control strategies with the aim to select the optimal one. The latent process used here is the Sellke thresholds which is easy to construct as part of the parameter estimation using MCMC and data augmentation methods. We revealed that by coupling epidemics the joint distribution of the outcomes of paired control strategies (including no-control) shows a strong positive correlation (see Figures 5.9 and 5.26). As a result, the variability in the difference between the outcomes is then reduced, which indicates a better estimate of the average causal effect defined by Holland (1986).

The stochastic optimisation method adopted here (Simulated Annealing) to optimising control strategies appears to be efficient in finding the optimal distribution of resources. However, strategy 2 (probability based allocation) seems to have outperformed strategy 3 (adaptation of Neyman allocation). Nevertheless, all simulations show that a control conditioned on the threat measures (conditional expectation of the product of risk and challenge) reduces considerably the expected number of infections independently of the spatial transmission mechanism of the disease and outperformed controls conditioned on the risk and the challenge measures.

Chapter 6

Case study and extension to heterogeneously mixing population

6.1 Case study: Citrus Canker data from Florida

6.1.1 Introduction to Citrus canker disease

Citrus canker is a serious leaf, fruit and stem blemishing disease caused by a bacteria pathogen *Xanthomonas axonopodis pv. citri* (Xac). The disease causes major economic losses in many citrus-growing countries including the USA (Gottwald et al., 2001a, 2002).

6.1.1.1 Mode of transmission

The bacterium is believed to be transmitted over short to medium distances through wind-driven rain (Kuhara, 1978; Stall et al., 1980; Serizawa, 1981; Gottwald et al., 1988; Gottwald and Timmer, 1995), circumstantially up to at least $10km$ during severe rainstorms such as tornadoes and tropical storms (Gottwald et al., 1997b,a) and is exacerbated by the citrus leafminer, an insect which occurs in most citrus producing areas in most parts of the world and whose larvae attack all citrus varieties particularly affecting the younger leaves (Gottwald et al., 2002).

Moreover, long-distance spread is believed to occur *via* contaminated equipment, by exposure to diseased citrus, or by transfer by humans or exposed plant material (Civerolo, 1984; Schubert et al., 2001).

6.1.1.2 Symptoms

Infection of citrus canker causes a variety of symptoms on leaves, fruits and stems. Symptoms on leaves first appear as small, raised lesions that resemble blisters. After some time, they might turn tan and brown with the edges appearing water-soaked

and developing a yellow halo (Cabi, 2016). The cankers can be seen on both sides of the leaf. While newly formed lesions could be difficult to spot, older lesions take on a raised, 'corky' appearance and sometimes fall out, leaving behind a shot-hole effect (Gottwald et al., 2002).

In the other hand, symptoms on fruits begin with dark raised lesions having a similar yellow halo as on the leaves. This causes premature fruit drop and with associated reduction in fruit quality and yield (Gottwald et al., 2002).

Finally, symptoms on twig and stem are typically corky and raised lacking the chlorotic halo as on leaves and fruits, surrounded by an oily or water-soaked margin. If twigs are not killed back by girdling infections, they may store inoculum that re-infects new tissues as it is produced by the citrus tree. Note that symptoms on the stem and twigs show that the tree may have been infected for a long time given that stem and twig are not as vulnerable as leaves (Gottwald et al., 2002).

6.1.1.3 Eradication programs

In spite of all regulation imposed by several countries, the disease continues to spread and increase its geographic range. In effect, citrus canker is present in over thirty countries in Asia, the Pacific and India Ocean islands, South America, South-eastern USA and South Africa (Gottwald et al., 2002). To date, there is no effective control of citrus canker. However, the control is better achieved through prevention. Typically this means that when canker is detected in an area, measures need to be taken to protect the canker-free areas. Copper-based fungicides can provide a barrier against canker spread through defoliated trees (Gottwald et al., 2002). Moreover, elimination of inoculum by removal and destruction of infected and exposed trees is the most common practice to eradicate or contain the disease to stop future spread. For example, the successful eradication program during the 1995 citrus canker outbreak in Florida, USA involved removal of millions of commercial, nursery and residential trees including infected and healthy trees at a cost of $>$ \$1 billion (Gottwald and Irey, 2007) and was subject to controversy since certain owners were extremely against the removal of healthy trees. Early eradication proved to be more cost-effective as shown by the eradication program during the Australian citrus canker outbreak in 2004. Indeed, the Australian Bureau of Agricultural and Resources Economics estimated the net benefit of eradicating the disease from Queensland alone was about \$70 million (Gambley et al., 2009).

6.1.2 Bayesian analysis of the Citrus Canker data

6.1.2.1 Data and assumptions

The data used for the analysis is partially introduced in chapter 5. In addition to the 1113 citrus trees located in Broward county and labelled B_2 described in Section 5.6.2.1, the locations of the infected trees are identified and their corresponding infection times are estimated by experienced personnel from the lesions, size, tree age and other phenotypes characteristics. The infection times are not recorded precisely but are grouped into 24 successive 30-day intervals. The data obtained is equivalent to the data described in chapter 5 if the observation times shifted back by the asymptomatic period Δ so that all infections are observed. Here, the data consist of spatial snapshots of locations of susceptibles and infected trees at 30-day intervals. A comprehensive description of these data is found in Gottwald et al. (2001b).

In Neri et al. (2014), the analysis of these data was performed with the assumption that an infected tree remains infected over the entire process, which is not the case. In reality, once an infected tree is identified as such, it is removed within two to four weeks after the detection (Gottwald et al., 2001b). Based on that, we assumed for simplicity that an infected tree is removed three months (90 days) following its infection. By considering this time lag, we combine the detection and the removal period into a single removal period. This assumption is equivalent to a system where hosts take three months to develop symptoms after infection and immediately at the onset of the symptoms, they are instantaneously detected and removed. To illustrate the methodology described in this thesis, we only consider the first 6 months of the data i.e. we assume that the final observation time is $t_{obs} = 180$ days where in total 55 trees were detected during that period among which 18 removed.

6.1.2.2 Model

We consider a spatially-explicit SIR model for the spread of the disease over time and space. Indeed, a host in the susceptible state (S) until it is infected (I) either by infected host or by external inoculum. A newly-infected host (I) is immediately infectious and is removed (R) after a three-month lag and does not play any further part in the epidemic transmission.

The transition from S to I in the model considered here is the same as the one described in the previous chapter. Recall that the model represents two sources of infection: primary (external sources) with a fixed rate ϵ and secondary (host-to-host infection) with rate β. The transition from S to I is governed by the equation (Equation 5.2.4) i.e.

$$P\left(i \text{ infected in } [t, t+dt]\right) = \lambda_i(t)dt + o(dt), \tag{6.1.1}$$

$$\lambda_i(t) = \left(\beta \sum_{j \in I(t)} K(d_{ji}, \alpha) + \epsilon \right) \tag{6.1.2}$$

It is worth noting that the latent period for citrus canker is short, averaging 14 days (Gottwald et al., 2001b), therefore we did not take into account latency for infection in the analysis here. As stated above, the removal period is kept fixed (90 days); hence the transition from (I) to (R) is assumed not to be stochastic.

Furthermore, for the dispersal kernel, we focus solely on a short-range transmission since we do not intend here to conduct a full analysis of these data. For a full analysis of this data we refer the reader to Neri et al. (2014) where choice of kernel function is discussed. We consider the exponential dispersal kernel with a short-distance correction used in Neri et al. (2014), defined as follows:

$$K(d, \alpha) = \begin{cases} \dfrac{1}{\pi(2\alpha + d_o)d_o} & \text{if } d \leq d_o \\ \dfrac{1}{\pi(2\alpha + d_o)d} \exp\left(-\dfrac{d - d_o}{\alpha}\right) & \text{if } d > d_o \end{cases} \tag{6.1.3}$$

with $d_o = .005$km. Due this parameterisation of the kernel, the parameters α, β and ϵ are respectively measured in km, $days^{-1}km^2$ and $days^{-1}$.

6.1.2.3 Parameter estimation

The parameters are estimated using the MCMC algorithm described in Section 5.3 considering $T = 180$ and $T = T_a = 270$ (the assessment time) to generate the joint distribution $\pi_0(\boldsymbol{\theta}, \underline{x}(T)|\mathbf{y})$ with $T = 270$. We assume independent non-informative uniform priors for all parameters with support coinciding with the interval $[0, 1]$. As described in chapter 5, a Metropolis-Hastings algorithm with independent Gaussian proposal distribution is used. We run the chain for 510000 iterations with the first 10000 discarded. Posteriors distributions of the model parameters are shown in Figure 6.1. The posteriors distributions match regardless of how far beyond the observation time we impute infection times providing reassurance that the methods are correctly implemented. In addition, we observe that the primary infection rate and the dispersal kernel parameter match those reported in Neri et al. (2014). The posterior distribution of the contact rate β however scaled by a factor about 2. This is due to the inclusion of the removal process into the analysis. Intuitively, removal of trees during the process decreases the infectious pressure on susceptibles. Given that the two processes are conditioned on the same data, an increase in β will increase the infection pressure to maintain the trend in the data.

We then draw 100000 samples from the joint posterior distribution $\pi_0(\boldsymbol{\theta}, \underline{x}(T)|\mathbf{y})$ to construct the prioritisation measures discussed below. Moreover, the joint posterior

distribution $\pi_0(\boldsymbol{\theta}, \boldsymbol{Q}|\mathbf{y})$ is ultimately generated as part of the MCMC scheme using samples from $\pi_0(\boldsymbol{\theta}, \underline{x}(T)|\mathbf{y})$. Again, this posterior distribution is then used to generate pre-epidemics on which different competing control strategies are compared.

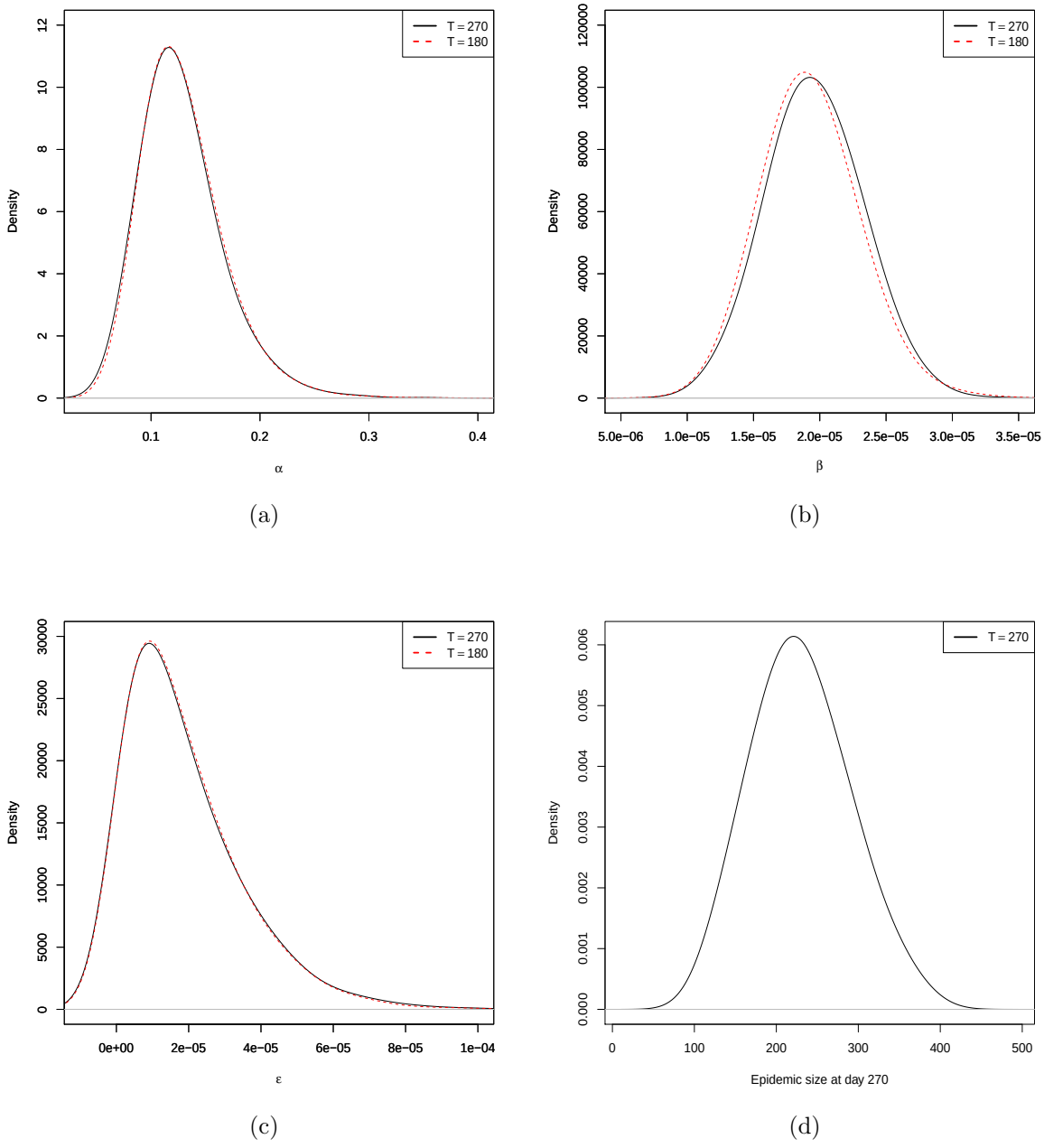


Figure 6.1: Posterior distributions of α, β, ϵ for $T = 180$ and $T = 270$ and the epidemic size considering $T = 270$ days.

6.1.3 Control strategies and results

Similar to chapter 5, different measures defined in Equations (5.3.9-5.3.12) are constructed including the risk ($\mathcal{R}^*$), the challenge ($\mathcal{C}^*$), the threat ($\mathcal{T}^*$) and the product

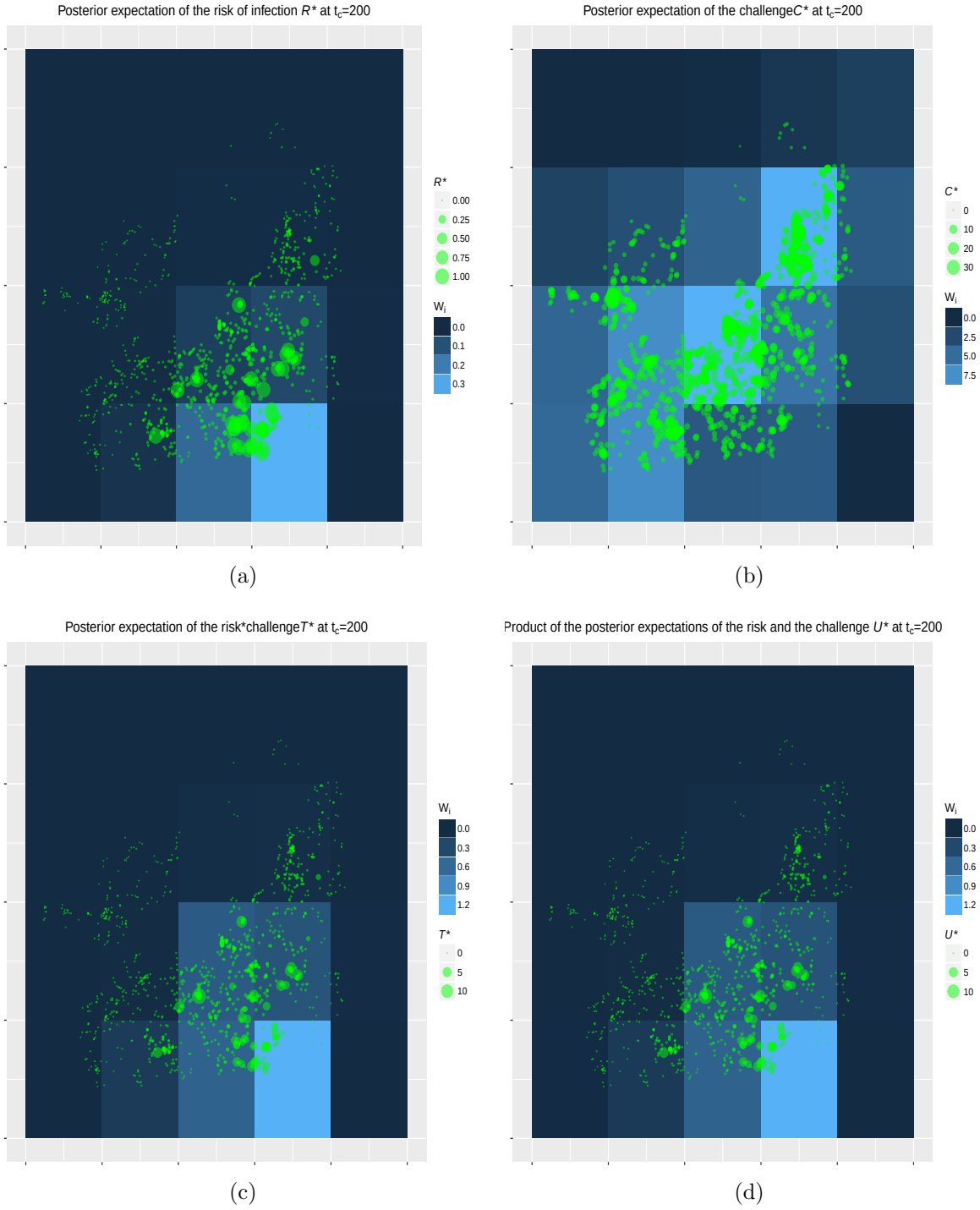


Figure 6.2: Posterior expectation of the risk ($\mathcal{R}^*$)(a), the challenge ($\mathcal{C}^*$) (b), the expectation of the product of risk and challenge ($\mathcal{T}^*$) (c) and the product of the expectations of the risk and the challenge ($\mathcal{U}^*$) (d). Prioritisation measures constructed at $t_c = 200\,days$, the intervention day.

of expectations ($\mathcal{U}^*$) (see Figures 6.2 and 6.3). In addition the definition given for the risk and challenge is subject to little change. Here, the presence of the host is taken into account, in other words, hosts removed before the measures construction time are not taken into account. Only active hosts are considered. Two different times

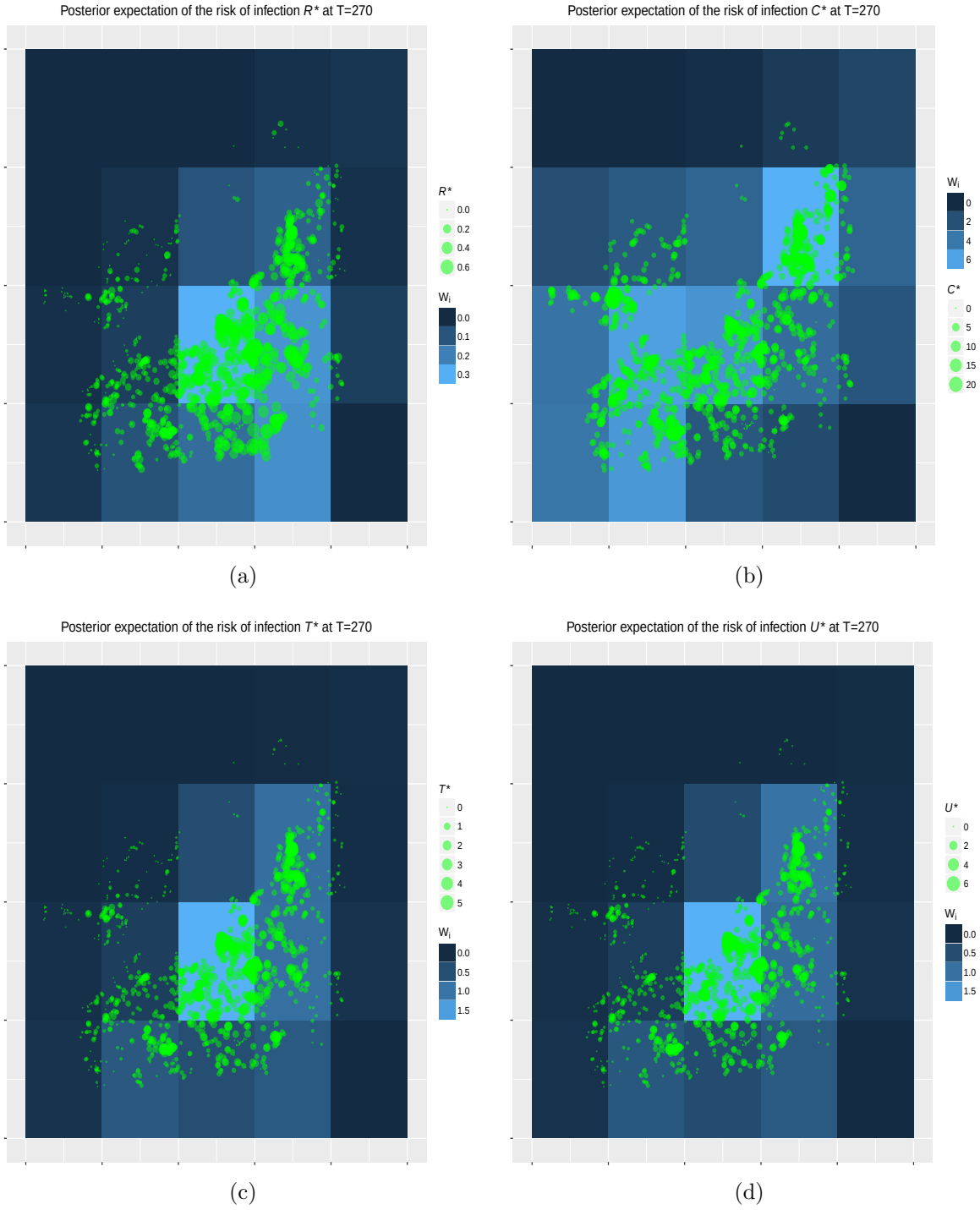


Figure 6.3: Posterior expectation of the risk ($\mathcal{R}^*$)(a), the challenge ($\mathcal{C}^*$) (b), the expectation of the product of risk and challenge ($\mathcal{T}^*$) (c) and the product of the expectations of the risk and the challenge ($\mathcal{U}^*$) (d). Prioritisation measures constructed at $T = 270\,days$, the assessment day

are chosen to construct the measures: the intervention day $t_c = 200$ (see Figure 6.2) and the assessment day $T_a = 270$ (see Figure 6.3). Note that regions on Figures 6.2 and 6.3 are shaded according to their level of measures used (w_i) and used for the prioritisation of the regions. Also, trees are removed once diagnosed infected among

the list of prioritised hosts. We consider various resource constraints (the maximum number of removals) $N' = 100, 200, 300, 400, 500$ and we assume that the number of regions to visit n_r must satisfy $1 \leq n_r \leq 20$. In addition, we assume that the intervention will take place at a future time $t_c = 200$. We wish to estimate the impact of different control strategies conditioned on the available resources at the assessment time $T_a = 270$.

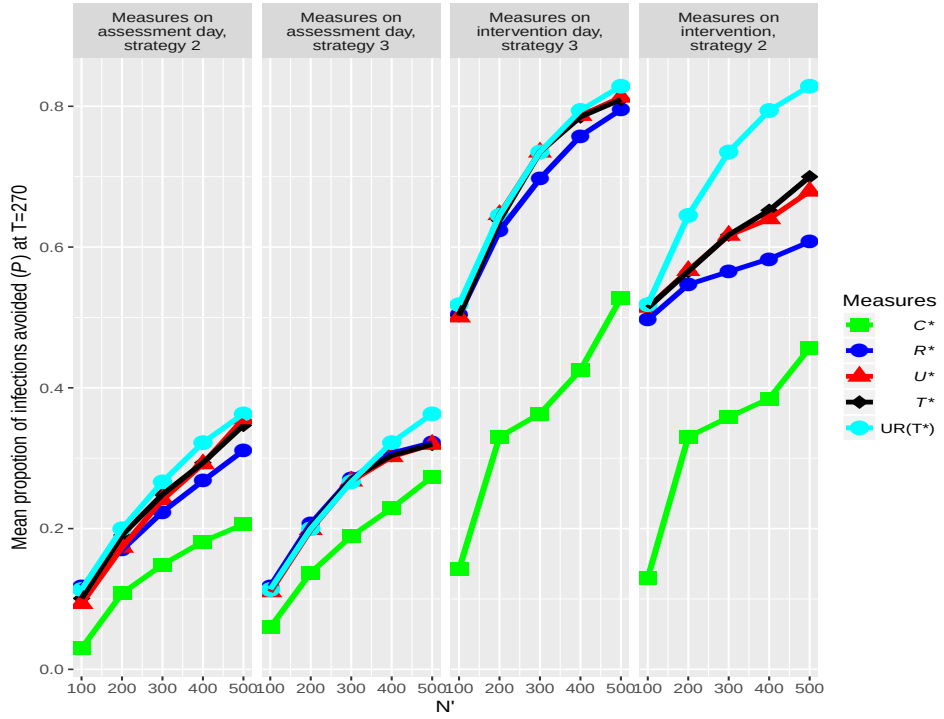
For every measure, we consider the control strategies of the form $d^*_{\{\mathcal{M},2,2\}}$ and $d^*_{\{\mathcal{M},2,3\}}$, the optimal controls obtained using the Simulated Annealing Algorithm 2 (incremented sample size) conditioned on the Strategy 2 and 3 respectively (see 5.6). After incrementing simulated realisations $\{(\boldsymbol{\theta}^i, \boldsymbol{Q}^i)|i = 1, \ldots, 20\}$, the algorithms converge for different measures (assuming that the algorithms stop if 10 successive realisations do not improve the control strategy) using different various values of N'. We compare the proportion of infections avoided by day 270 when applying the above optimal controls on the same 1000 realisations randomly drawn from the posterior distribution $\pi_0(\boldsymbol{\theta}, \boldsymbol{Q}|\mathbf{y})$.

Results show that the prioritisation using the threat measure ($\mathcal{T}^*$) is more cost-effective than those using $\mathcal{C}^*$ and $\mathcal{R}^*$ in reducing the impact of future outbreak when constructed at both intervention day and the assessment day (see Figures 6.4a). This is also emphasized with the distribution of the number of infections avoided using $N' = 500$ with the maps constructed at $T_a = 270$ shown in Figure 6.4b. In contrast to the observation made in the chapter 5, Figure 6.4a shows that conditioning on measures constructed on the intervention day is more cost-effective than the one using measures on the assessment day. This observation is a direct result of the removal process introduced here. Indeed, most hosts that spread the disease by time t_c are likely to be removed before the assessment time T_a. In this case, their contributions are not taken into account when constructing the maps at T_a, therefore are not removed during the intervention t_c (they are not used for prioritisation) thus continue to spread the disease.

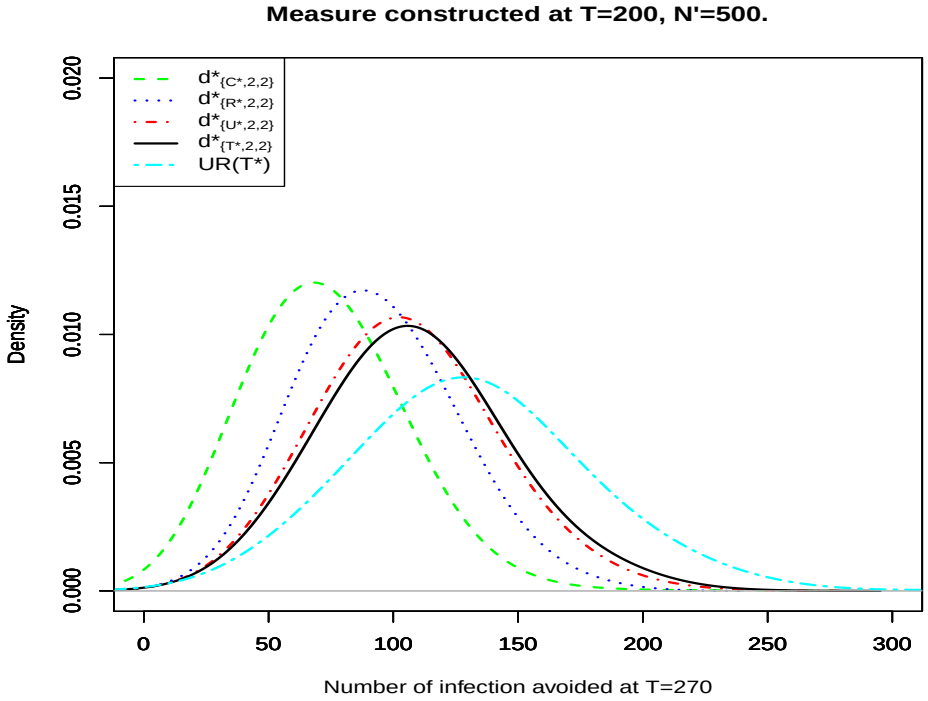
In contrast to the observation made in chapter 5, Strategy 3 outperforms strategy 2 when the maps are constructed on the intervention day. The reason being that, the first ranked region (see 6.3a and 6.3c for example) only comprises very few trees; therefore, the strategy 2 will typically allocate more resources than needed in that particular region which will surely impact on regions densely populated. However, given the randomness of Strategy 2 inherited from the multinomial distribution, it will eventually outperform Strategy 3 in a long run. In addition, we observe on the second and third panel of the Figure 6.4a where the strategy 3 is used that, the outcome of the prioritisation using $\mathcal{R}^*$ approaches the one of $\mathcal{T}^*$. The main reason for that is that high risk and threat regions coincide as shown in Figures 6.3a and 6.3c leading to the detection and removal of approximately similar number of infectious

hosts across these regions. Also, with the latter strategy, the proportion of infections avoided by day 270 using the measure $\mathcal{T}^*$ approximates the one obtained from the control strategy $UR(\mathcal{T}^*)$ (selecting the first N' trees in that ordering with respect to $\mathcal{T}^*$). It confirms that prioritisation based on the strategy 3 could be used as an alternative strategy in a large scale system where resources can only be deployed in limited numbers of subregions.

In reality, when comparing different control strategies in term of their expected outcomes (epidemic size for example), we are interested in estimating the average effect of one strategy over an alternative control strategy (Holland, 1986). Like any unbiased estimate, a reduction in the variability is always preferred for a better accuracy of the estimate. The coupling technique adopted in this thesis provides just that. Indeed, Figures 6.5b and 6.5a show the pairwise joint distributions of the outcomes (epidemic size) of various control strategies on the lower-panel and the box plots showing the variability in the difference of these outcomes. As expected, the outcomes are strongly correlated as observed in the previous chapters. Hence, the approach gives a reduction in the variability between outcomes. Thus, the approach provides a better estimate of the expected difference between the outcome of the controls. In addition, we notice that as the correlation with the outcome from the unconstrained control ($UR(\mathcal{T}^*)$) for example is getting stronger, the variability becomes smaller, therefore gives a much better estimate of the average causal effect. This could be used as criterion for selecting an alternative strategy among different strategies if an optimal strategy is known in advance but could not be used in practice (e.g. in a large system). Finally, all control strategies were coupled on the same 1000 realisations of the epidemic, which reduces the amount of realisations needed compared to independent sampling of outcomes over controls.

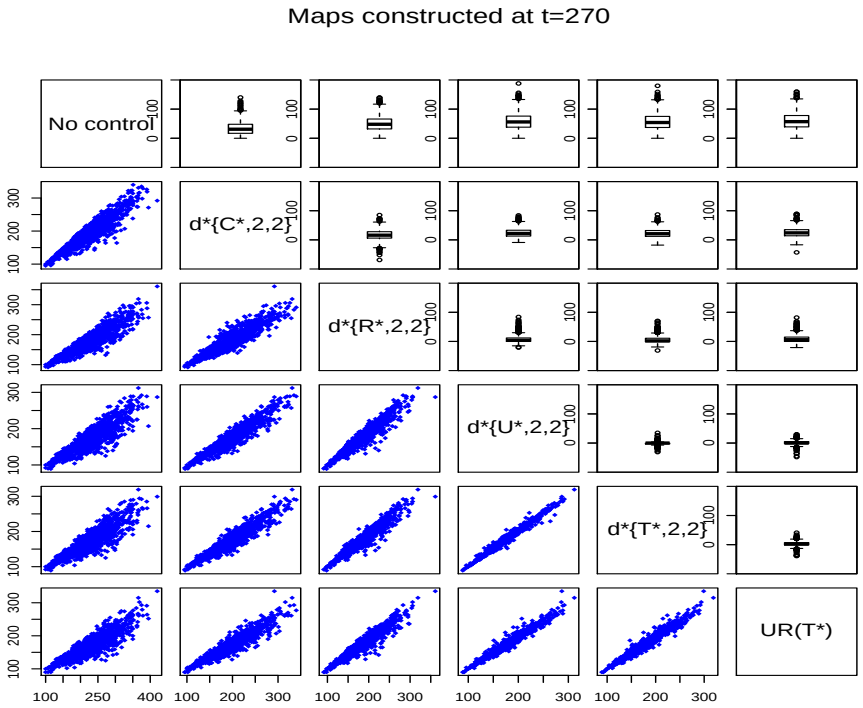


(a)

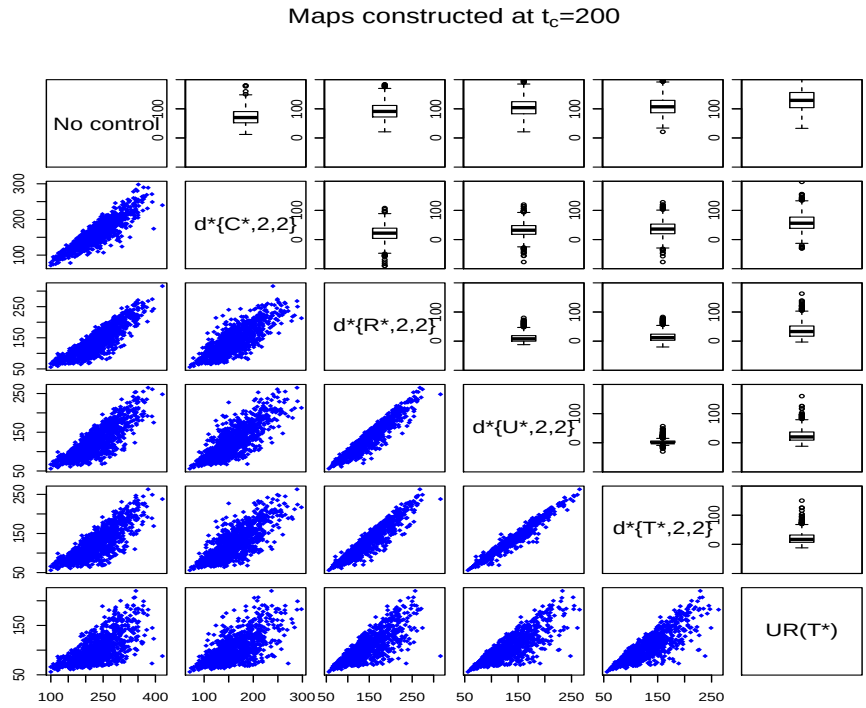


(b)

Figure 6.4: (a) Effects of the maps and different strategies (Strategy 2 and 3) on the relationship between the mean proportion of infections avoided and the maximum number of hosts to visit. (b) Posterior distribution of the number of infections avoided by $T_a = 270$ when $N' = 500$ and the maps are constructed at $t_c = 200\,days$.



(a)



(b)

Figure 6.5: Pairwise joint posterior samples of various control strategies on the lower panel and the distribution of the difference between the outcomes of the paired controls when the maps used are constructed on (a) the intervention day $t_c = 200\,days$ and (b) the assessment day $T_a = 270\,days$

6.1.4 Retrospective analysis of the impact of controls used prospectively

6.1.4.1 Motivation

So far, the control strategies are deployed in the light of what was known at $t_{obs} = 180$, to reducing the impact of the epidemic at a future time. This is a prospective approach common in epidemiology where decisions must be taken at early stage of an epidemic. However, when looking back at an historic epidemic, a sensible question to answer could be 'what would have been the impact of the decision made in the light of what is known now?'.

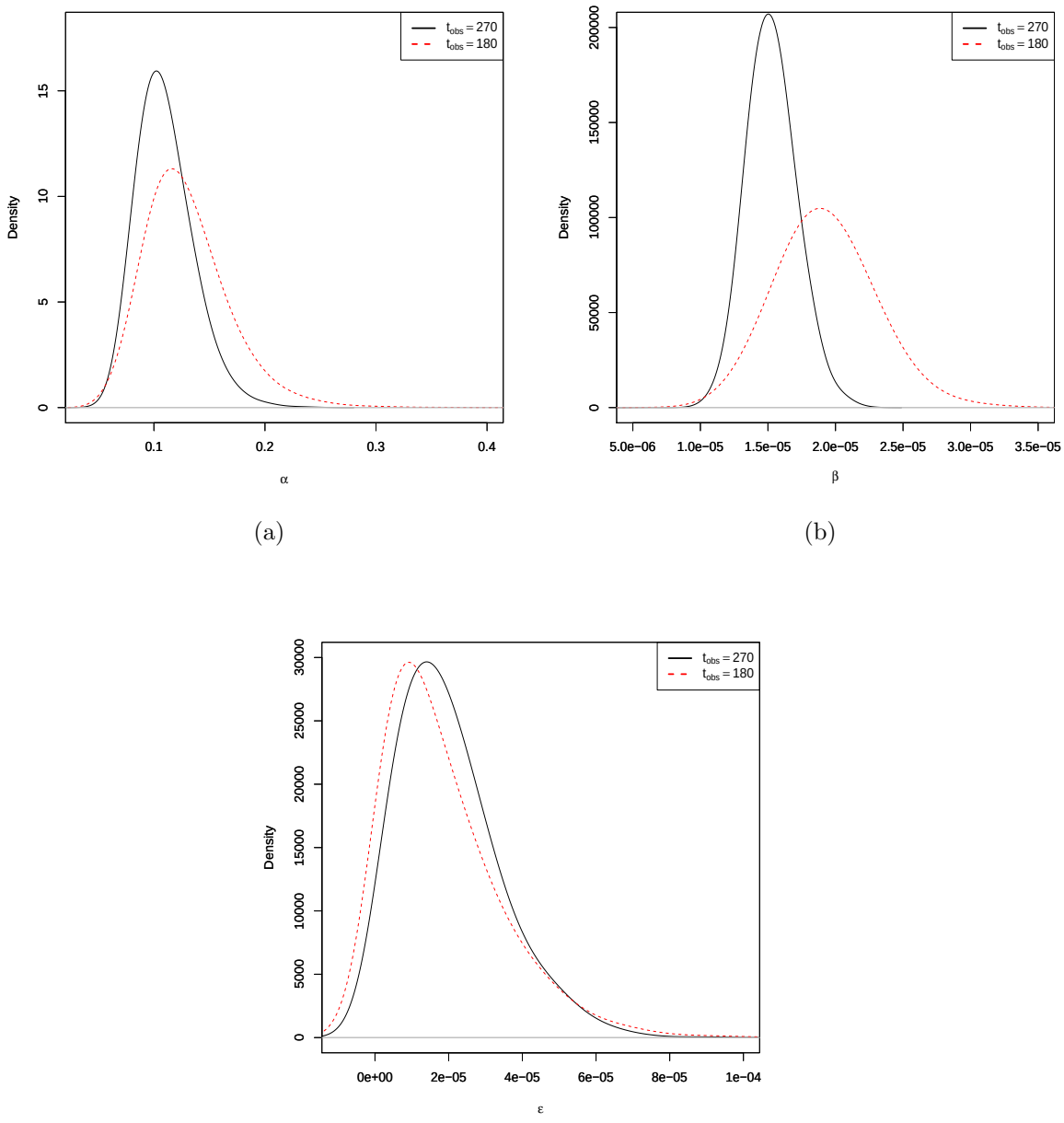
The retrospective questions have been tackled in different ways in the literature. A semi-retrospective approach whereby alternative controls are compared on independent realisations of the epidemic has been mostly used in the literature (Ripley et al., 2003; Keeling et al., 2001). However, implementing a control of this type can lead to a significant probability of a worse outcome than was actually observed (Woolhouse, 2003). To answer the above retrospective question, we use the approach of Cook et al. (2008) where epidemics are coupled. We couple epidemic under different controls using the Sellke thresholds. An implicit assumption in this kind of analysis is that the control measures do not affect the Sellke threshold $\boldsymbol{Q}$ or the system parameter $\boldsymbol{\theta}$.

6.1.4.2 Application

We consider the observation (citrus data) up to $t_{obs} = 270$. This will correspond to the disease progress maps made at 30-days intervals over 9 months, yielding a total of 113 infections among which 55 were removed.

We use the MCMC routine developed in 5.3.2 to sample from the posterior distribution $\pi_0(\boldsymbol{\theta}, \underline{x}(t_{obs})|\mathbf{y})$ using independent uniform priors with support $[0, 1]$ on all parameters. Again we consider a Gaussian proposal distribution for all parameters and a uniform proposal distribution over the censored intervals for the infection times. We run the chain for 510000 with the first 10000 discarded and we thin the chain each 100 iterations. Figure 6.6 shows the posterior distributions of the model parameters. Unsurprisingly, the variance in the posterior estimates decreases if $t_{obs} = 270$ compared to the case where $t_{obs} = 180$ as more information is available. We then draw $(\boldsymbol{\theta}^i, \underline{x}^i(T))$, $i = 1, \ldots, 1000$ from $\pi_0(\boldsymbol{\theta}, \underline{x}(T)|\mathbf{y})$ where $T = 270$ to generate $\pi_0(\boldsymbol{\theta}, \boldsymbol{Q}|\mathbf{y})$.

Figures 6.7 and 6.8 show the distributions of the proportion of infection avoided with different control strategies where the maps are constructed on the intervention day $t_c = 200$. There are a number of observations to be made from these figures. First, we observe that the control strategies obtained in the light of what was known at $t_{obs} = 180$ perform effectively well on the actual trajectory of the epidemic. Second, it can be seen that, while there are differences between the outcomes of the strategy



(a) (b)

(c)

Figure 6.6: Posterior distributions of α, β, ϵ using the complete data at $t_{obs} = 270$ and $t_{obs} = 180$.

2 (probability based allocation) and strategy 3 (adaptation of Neyman allocation) when conducting the prospective analysis for different maps, the distributions of the proportion of infections avoided is similar for the retrospective approach. In addition, there is no clear difference in the outcomes of the control strategies for measures $\mathcal{R}^*$, $\mathcal{U}^*$ and $\mathcal{T}^*$. Finally, results suggest that with $N' = 100$, the posterior distribution of the proportion of infections avoided approaches the one of $N' > 100$ when the control is conditioned on $\mathcal{R}^*$, $\mathcal{U}^*$ and $\mathcal{T}^*$ measures. Therefore, resources would be wasted if

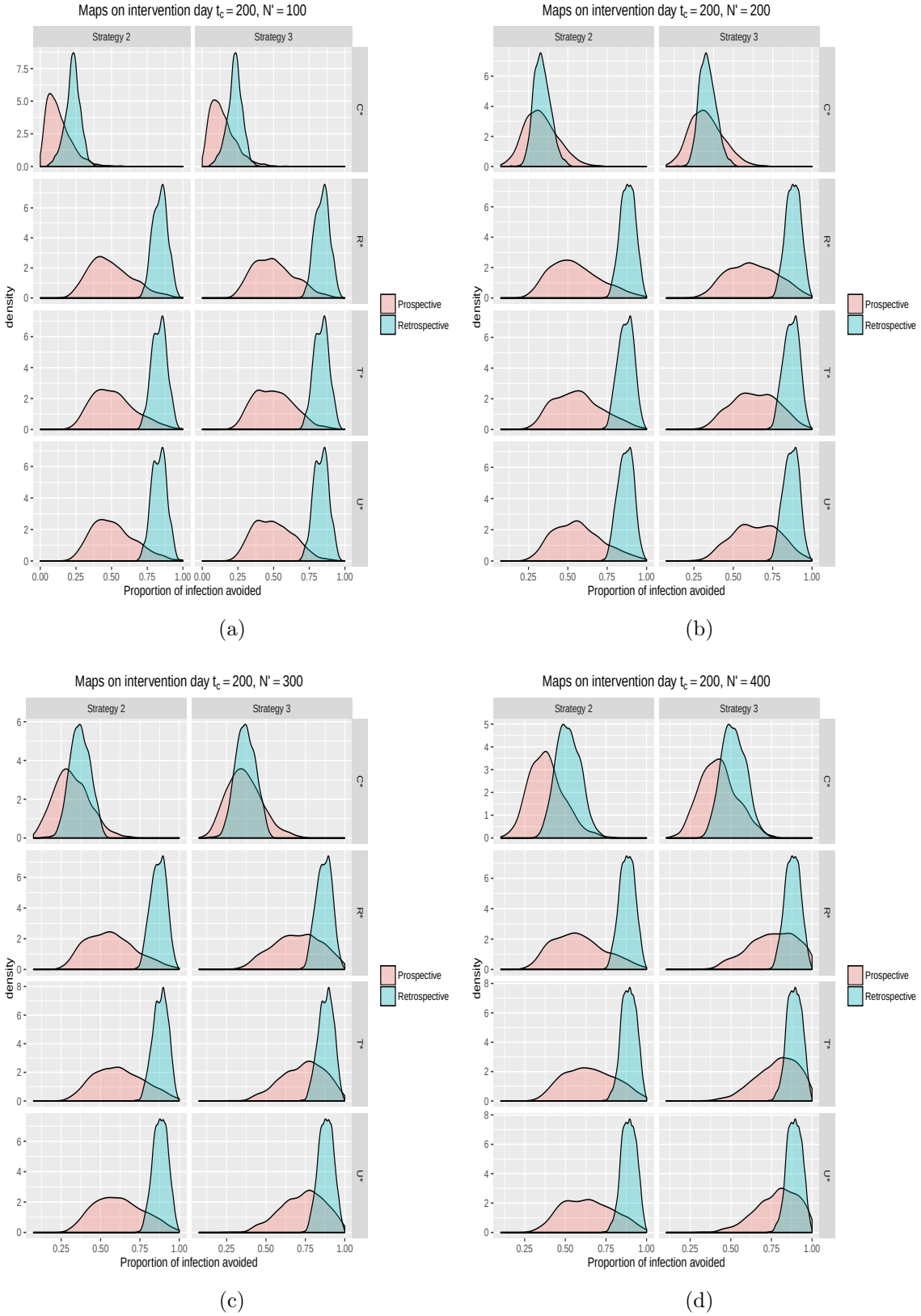


Figure 6.7: Comparison of the posterior distribution of the proportion of infections avoided for the prospective and retrospective approaches considering $N' = 100, 200, 300, 400$.

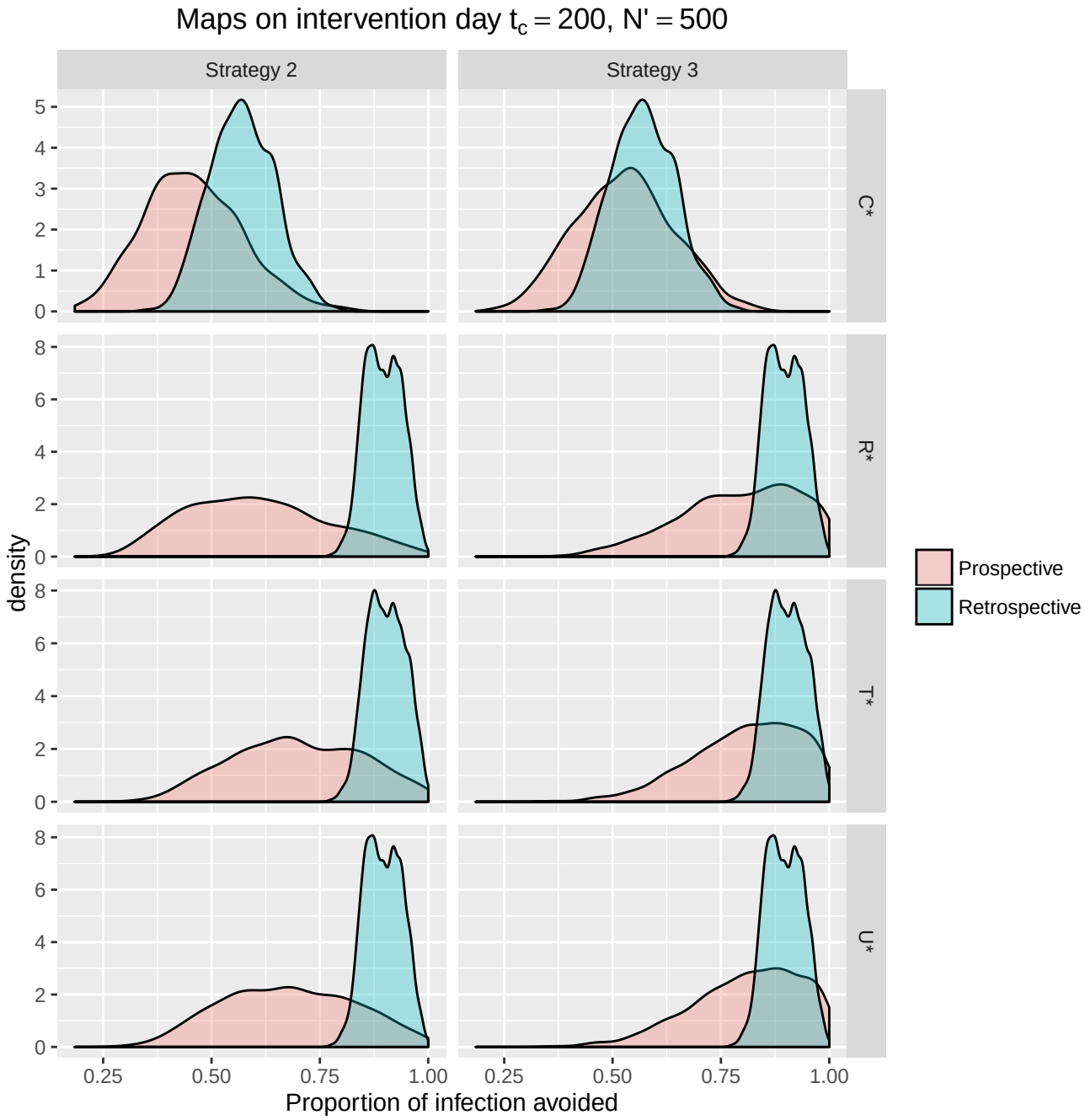


Figure 6.8: Comparison of the posterior distribution of the proportion of infections avoided for the prospective and retrospective approaches for $N' = 500$

the control was carried out using $N' > 100$. These observations are mainly due to the fact that based on what was known at t_{obs}, we have predicted scenarios that are typically worse than that which eventually occurred.

6.2 Extension to spatio-temporal epidemics in heterogeneously mixing population

6.2.1 Introduction

6.2.1.1 Why models with heterogeneously mixing population are necessary

In the chapter 5 and earlier in this chapter, we have focussed on models for populations of identical, homogeneous hosts. That is hosts are identically uniform to each other. This assumption is unrealistic for most real-life epidemics. For example in human population all members are inherently different from each other (Mayr, 1982) so that every member in that population could respond differently to a specific disease. Having said that, it is important to consider source of heterogeneity when modelling epidemics.

Different factors could be sources of variation within hosts in the population. An obvious candidate of such variation is host age (Isham, 2005). For example, the period from infection with HIV to diagnosis of full AIDS is known to vary with host age (Billard et al., 1990). In addition, age plays a key role in the spread of some plants diseases. Indeed, while estimating parameters for models of an emerging epidemic subject to control, Parry et al. (2014) pointed out that the spread of Huanglongbing is highly dependent on citrus age. The introduction of host mixtures to support disease control is an obvious source of diversity in some crop populations (e.g barley). Following the definition given to cultivar mixtures in Wolfe (1988) as "mixtures of cultivars that vary for many characters including disease resistance but have sufficient similarity to be grown together", it is crucial to consider structure in the population when modelling with a disease outbreak in such a system.

6.2.1.2 Review of some modelling approaches

The general approach considered when analysing heterogeneity is to split the population into groups (types) within which hosts mix homogeneously (no variation in the capacity of hosts to become infected, their ability to infect other hosts in the population (O'Neill, 2010)) within each group. In this case, a mixing matrix whose element r_{ij} specifies the rate of disease transmission from an i-host (hosts in group i) to a given susceptible host at each group j is usually considered (Ball et al., 1997b). This mixing matrix is in general not symmetric given that r_{ij} depends on the infectivity of the infective and the susceptibility of the susceptible. More complex structure of mixing matrices have been consider in Koopman et al. (1989).

During the last three decades or so researchers have developed various models for epidemics in structured populations. Longini and Koopman (1982) studied models

in which individuals reside in households and are subject to two potential sources of infection: local source (infection from the same household) and external source (infection from another household). They assumed that infections occur within the household independently of the dynamics of the community. Similar work was developed by Addy et al. (1991) who extended the approach of Ball (1986). The model developed by Longini and Koopman (1982) is used by Britton and Becker (2000) to estimate the critical vaccination coverage required to prevent outbreaks in households. O'Neill et al. (2000) on other hand, applied the MCMC methods to analyse temporal and final-size data from model in population partitioned into households. Ball and Lyne (2002); Becker et al. (2003) compared various vaccination strategies through their effect on a population subdivided into households. Ball et al. (1997a) introduced an epidemic models with two levels of mixing, along with methods for statistical inference and discussed various vaccination strategies. The model described assumes local and global contacts among individuals. Becker et al. (2005) drew inferences from models of this type in a Bayesian setting. Britton et al. (2011) have extended the two-level mixing framework to three levels.

Despite all these aforementioned models, researchers pointed out that it is practically impossible to capture every aspect of the population structure since individuals interact with each other in diverse ways. In an attempt to address this issue, models where the population structure is represented as a random network are considered. Britton and O'Neill (2002) applied the MCMC techniques to make inferences using models where individuals interact through social networks according to a Bernoulli random graph. Also, heterogeneity has been considered when modelling the 2001 Foot and mouth disease in UK where infectivity and susceptibility of farms are assumed to vary with the species present in the farm (Ferguson et al., 2001; Keeling et al., 2001; Jewell et al., 2009).

6.2.2 Model description

We assume that the population is subdivided into k groups each of size N_i, $i = 1, \ldots, k$ so that the size N of the population is $N = \sum_{i=1}^{k} N_i$. Let u_{lm} be the rate at which a host of type l contacts a host of type $m, l = 1, \ldots, k$. The generalisation of the homogeneous model is done in a straightforward way to the case where there is heterogeneity within the population hosts by assuming that an infectious host i makes contact with a susceptible j at rate h_{ij}.

In this chapter, we adopt the case where the general contact parameter h_{ij} can be represented as a product $h_{ij} = \beta_{g(i)}\mu_{g(j)}$, known under the name of *proportionate mixing* (Andersson and Britton, 2000; Isham, 2005), where g is a function that maps any host to its type so that i and j are of type $g(i)$ and $g(j)$ respectively. The parameters β_l and μ_m, $l, m = 1, \ldots, k$ represent infectivities and susceptibilities respectively.

Moreover, the probability that a host i becomes infected is given as follows:

$$P\left(i \text{ infected in } [t, t+dt]\right) = \lambda_i(t)dt + o(dt), \tag{6.2.1}$$

$$\lambda_i(t) = \beta_{g(i)} \left(\sum_{j=1}^{N} \mu_{g(j)} K(d_{ji}, \alpha) + \epsilon \right) \tag{6.2.2}$$

where $\lambda_i(t)$ is the hazard or the force of infection on host i at time t, ϵ the primary infection rate i.e. the rate at which the individual i contracts the disease from an external source of the population (e.g. environment drivers). In addition, recall that $K(d_{ji}, \alpha)$ is a non-negative function characterizing the challenge posed by the host j to i as a function of the inter-host distance d_{ji} and known as the transmission kernel which typically shows how infectivity decreases with distance d_{ji}. The parameter α represents the dispersal parameter.

6.2.2.1 Data and assumptions

Similar to Chapter 5, the data considered consist of a sequence of snapshots from which the type and the location of symptomatic hosts are observable, while the infection times are not. Only the period in which the host contracts the disease is known. The unknown infection times are then treated as additional parameters to the model using Bayesian methods coupled with data augmentation method. Again, we define the time origin by setting $t = 0$ to be time of introduction of the primary infection source, the data then consist of sequence of interval on the form $\left[t_l^{p-1}, t_u^p\right]$, $p = 0, 1, \ldots, n$ in which symptomatics hosts with their locations and type are identified. We also assume that $t_l^{-1} = 0$ and $t_u^n = t_{obs}$ is the final time of the observation. Similar to Chapter 5, we consider the period it takes for an infected host to display symptoms to be fixed Δ. Therefore for any host i detected as symptomatic for the first time in $\left[t_l^{p-1}, t_u^p\right]$ is surely infected in $\left[t_l^{p-1} - \Delta, t_u^p - \Delta\right]$.

6.2.2.2 Inference of model parameters

We again use Bayesian methods and MCMC modified to take account of the additional complexity that arises from the inclusion of heterogeneity.

Likelihood

Following the description given in Section 5.3, the likelihood of the vector parameters $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}, \boldsymbol{\mu}, \epsilon)$ at any time $T \geq t_{obs}$ where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$ can be written in a more simple way as follows:

$$L(\boldsymbol{\theta}; \underline{x}(T)) = \prod_{i \in \mathcal{I}} \lambda_i(s_i^-) \exp\left(- \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{I} \cup \tilde{\mathcal{I}}} A_{ij}(t_j - t_i) \mathbb{1}_{t_j > t_i} + \epsilon \sum_{j=1}^{N} \beta_{g(j)} t_j \right) \quad (6.2.3)$$

where $A_{ij} = \beta_{g(j)} \mu_{g(i)} K(d_{ij}, \alpha)$ is the pressure of host i on j and $\mathbb{1}$ the indicator function.

Recall that under Bayes' theorem,

$$\pi(\boldsymbol{\theta}, \underline{x}(T) | \mathbf{y}) \propto L(\boldsymbol{\theta}; \underline{x}(T)) \times \pi(\boldsymbol{\theta}). \quad (6.2.4)$$

Parameter estimation using MCMC

Authors have developed routines for spatio-temporal models to fit homogeneous host-types (Cook et al., 2007; Cunniffe et al., 2015; Neri et al., 2014) and heterogeneous host-types (Cook et al., 2007; Lau et al., 2014) where hosts vary in term of their suitability. Ferguson et al. (2001); Keeling et al. (2001); Jewell et al. (2009) have also used MCMC techniques of this type to estimate parameters from the 2001 Food and Mouth disease data from UK but assumed that hosts (farms) vary in terms of their infectivities and susceptibilities. We adapt these approaches and describe our own routine following the approach given in section 5.3.2. This yield samples from the posterior distribution $\pi_0(\boldsymbol{\theta}, \underline{x}(T) | \mathbf{y})$. Recall that one particularly attractive feature of drawing a sample to estimate the posterior distribution of the parameters is that we can apply any transformation to estimate the posterior distribution of any function of those parameters.

6.2.3 Modelling control of the epidemic in heterogeneously mixing population

The generalisation of the measures used to design controls in the case of the homogeneously mixing population in Section 5.2.5 is done considering each host's type. While the challenge that a host presents to others in the population when it becomes a new focus of infection is somewhat modified, the definition of the risk of a host to be infected at a certain time t remains the same.

Specifically the Equations 5.2.6 and 5.2.7 become

$$G_R^j(\underline{x}(t)) = \mathbb{1}_{\{x_j \leq t\}} \quad (6.2.5)$$

$$G_C^j(\underline{x}(t)) = \mu_{g(j)} \sum_{i \in S(t)} \beta_{g(i)} K(d_{ij}, \alpha) \quad (6.2.6)$$

Intuitively, hosts with high susceptibility are more likely to be infected and from the definition of the *challenge* in Equation (6.2.6), hosts surrounded by highly susceptible individuals will be highly ranked according to that measure. In this case, regions vastly populated by highly susceptible hosts might be prioritised when deploying controls.

Recall that the control used here is again based on the removal of infected hosts in the sampled population with respect to the resources.

6.2.4 Application

6.2.4.1 Data and inference

We use simulated epidemics to illustrate our methodology. We consider a population of size $N = 1000$ hosts identified by their locations sampled independently from a uniform distribution over 1000×1000 square region. We assume that the population can be partitioned into two groups ($k = 2$). The values considered here for $\beta_1 = 0.5, 1.5$ and $\beta_2 = 0.7, 3$ $\mu_1 = 1, 0.8$ and $\mu_2 = 1, 0.3$ for the susceptibilities and infectivities for group 1 and 2 respectively. The entire population is assumed to be disease free at $t = 0$, time corresponding to the introduction of the external source of infection with parameter $\epsilon = 10^{-4}, 25 \times 10^{-5}$. Observations are accepted up to $t_{obs} = 460$. The data consist of a sequence of snapshots of symptomatic hosts taken from $t = 130$ at 30-days intervals. Note that we use a short range exponential dispersal kernel of the form $K(d, \alpha) = \exp(-d\alpha)$ and the value of the dispersal parameter used is $\alpha = .2$ in both cases. This is summarised in Table 6.1.

Case	α	ϵ	β_1	β_2	μ_1	μ_2	t_{obs}	Infection in group 1	Infection in group 2
(I)	0.2	0.0001	0.5	0.7	1	1	460	66	187
(II)	0.2	0.000025	1.5	3	0.8	0.3	460	48	141

Table 6.1: Summary of the parameters used for the simulations in the three cases.

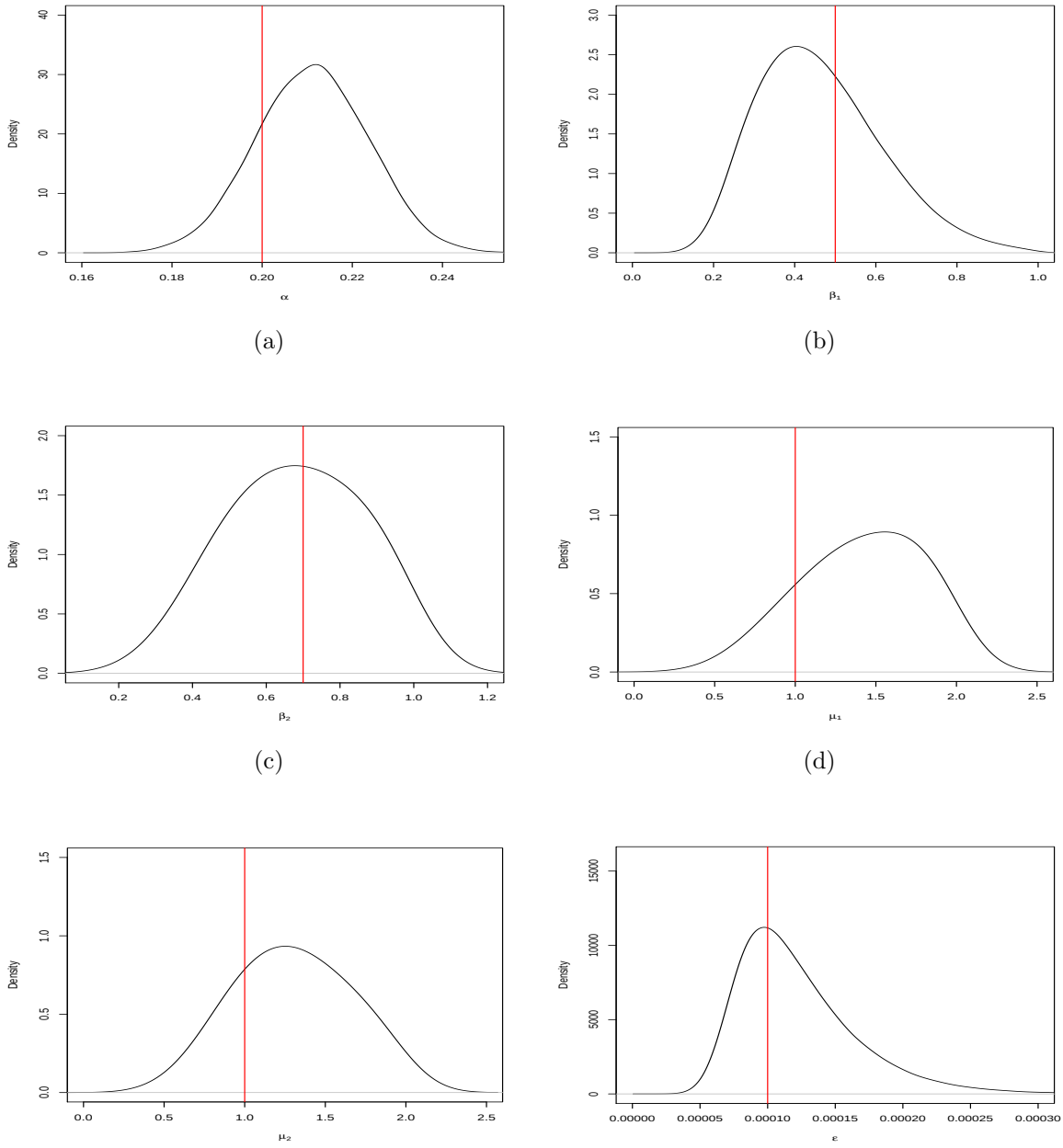
The case (I), where the infectivities are identically 1 for every group, is a special case of the two-level mixing. This is equivalent to the assumption that the contact parameters $\{h_{ij}\}$ satisfy $h_{ij} = \beta_{g(i)}$. This assumption implies that all infectives are equally infectious and hosts only vary in term of their susceptibilities. That is each susceptible of type (group) m is equally infected by any infective host irrespective of its group (Andersson and Britton, 2000; Cook et al., 2007).

For the parameter estimation, we adopt the MCMC algorithm described in section 5.3.2 to generate samples from the posterior distribution $\pi(\boldsymbol{\theta}, \underline{x}(T)|\mathbf{y})$. We assume non-informative uniform priors on α, β_1, μ_2 and ϵ, whereas we consider informative uniform priors $[0, 5]$ and $[0, 1]$ respectively for β_2 and μ_1. We run the MCMC for

500000 iterations considering $T = 550$ for case (I) and for 1000000 iterations for Case (II) using various imputation period considering $T = 360, 460, T = 550$. Note that $T = 460, 550$ involve the use of the reversible-jump techniques.

We only discuss results for case (II), as results for case (I) are qualitatively similar. The trace plots in Figure 6.10 show no sign of non-convergence. Also, the posterior distributions of the model parameters at various T shown in Figure 6.12 match regardless of how far beyond t_{obs} we impute infection times. Again this confirms that the algorithms are performing well. It is worth noting that the small deviations observed on the plots are due to the sampling effect. In addition, Figures 6.9 and 6.12 show that the parameters used for the simulation (dashed lines) are consistent with their respective posterior distributions highliting the fact that the estimation is good.

We then draw 200000 samples from the joint posterior distribution $\pi(\boldsymbol{\theta}, \underline{x}(T)|\mathbf{y})$ to produce measures for each host and the geographical maps at two different times (on the intervention day $t_c = 461$ and the assessment day $T_a = 550$). Figures 6.13 and 6.14 show the measures constructed on the assessment day and the priority regions. Each region is coloured according to its level of measure w_i computed as the average of all measures of hosts presented in the region (see equation (5.4.1)).



(a) (b) (c) (d) (e) (f)

Figure 6.9: Simulated data Case (I): The posterior distributions for Bayesian MCMC estimation of the model parameters where $t_{obs} = 460$ and imputing infection times up to $T = 550$ including the dispersal rate α (a), the susceptibilities of group 1 and 2 respectively β_1 (b) and β_2 (c), their corresponding infectivity μ_1 (d) and μ_2 (e) and the primary infection rate ϵ (f). Red lines correspond to the actual value used for the simulation.

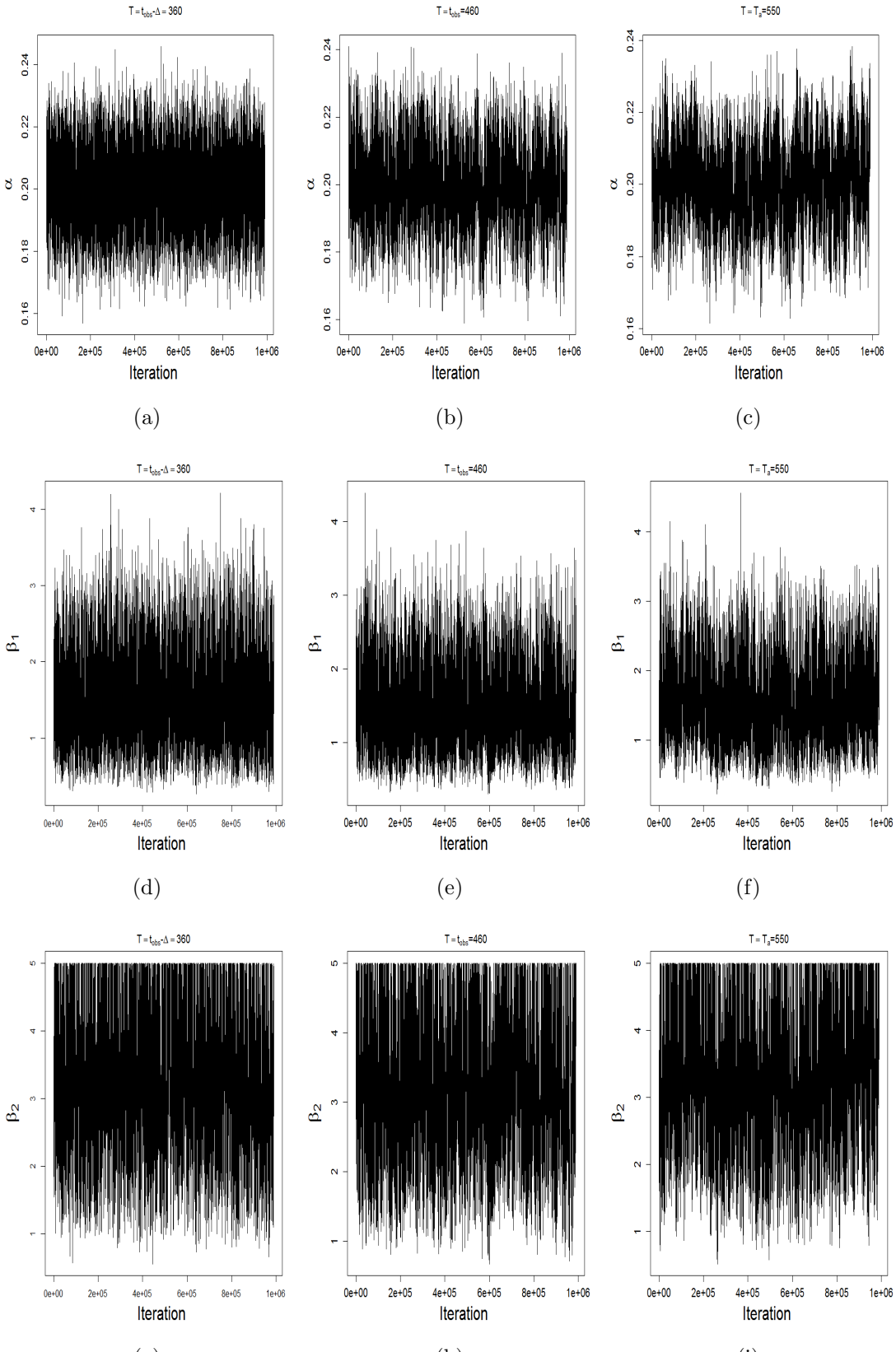
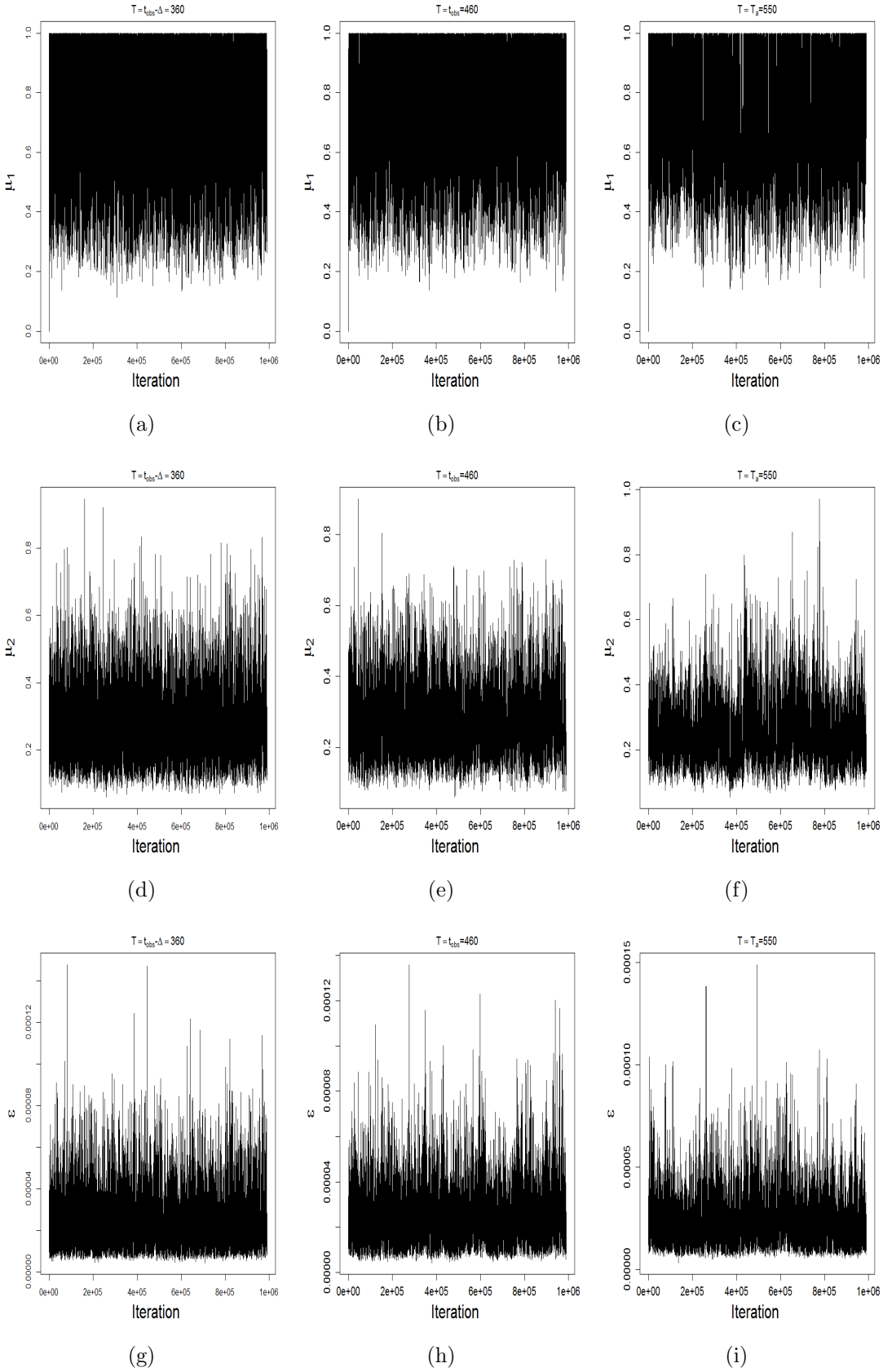


Figure 6.10: Sample trace plots for α, β_1, β_2 after a burn-in of 10000 iterations at different imputation periods.

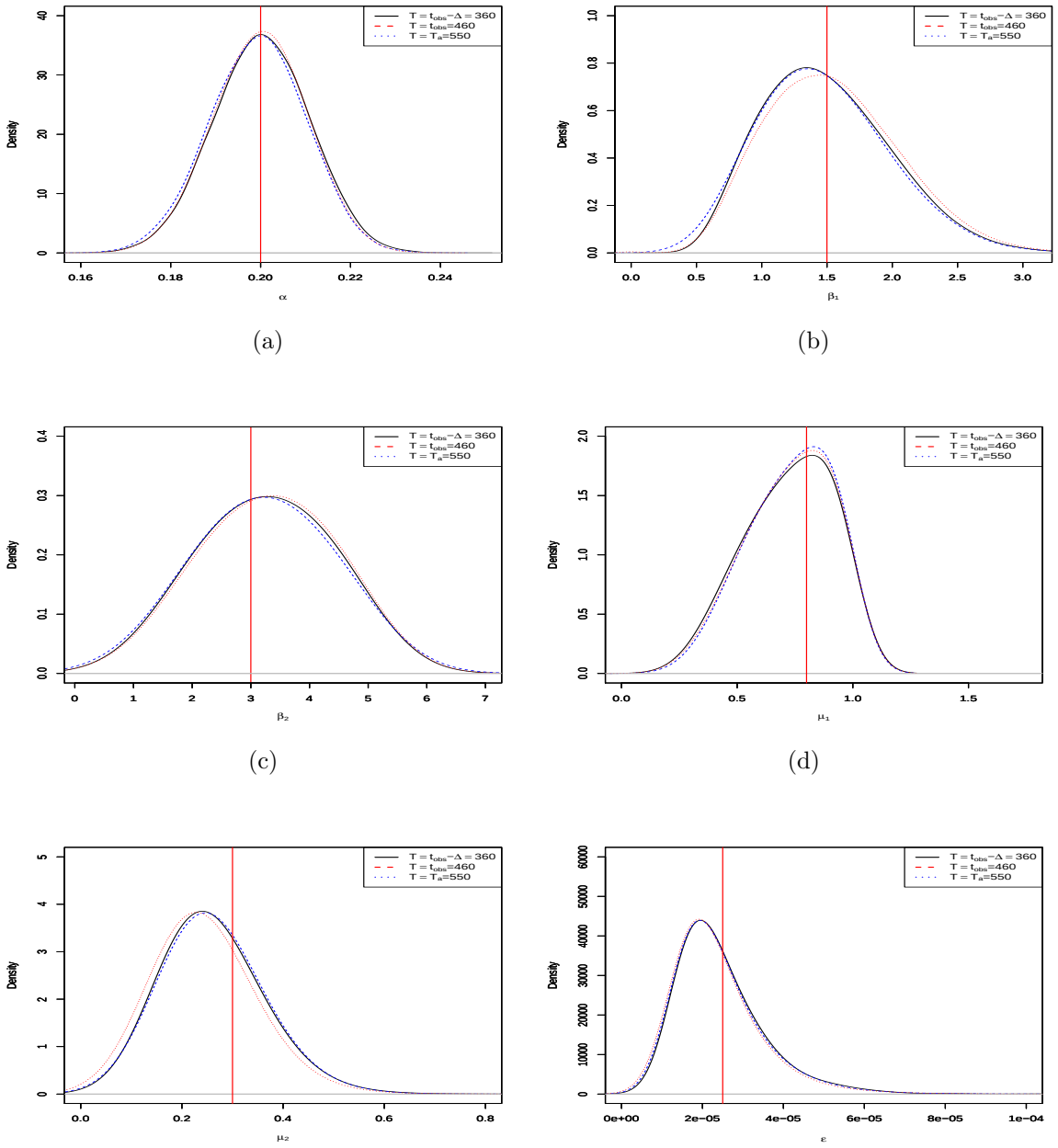


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 6.11: Sample trace plots for μ_1, μ_2 and ϵ after a burn-in of 10000 iterations at different imputation period.



(a) (b) (c) (d) (e) (f)

Figure 6.12: Simulated data Case (II): The posterior distributions for Bayesian MCMC estimation of the model parameters $\pi_0(\boldsymbol{\theta}, |\mathbf{y})$ obtained from $\pi_0(\boldsymbol{\theta}, \underline{x}(T)|\mathbf{y})$ where $t_{obs} = 460$ and $T = 360, 460, 550$ including the dispersal rate α (a), the susceptibilities of group 1 and 2 respectively β_1 (b) and β_2 (c), their corresponding infectivity μ_1 (d) and μ_2 (e) and the primary infection rate ϵ (f). Red lines correspond to the actual value used for the simulation.

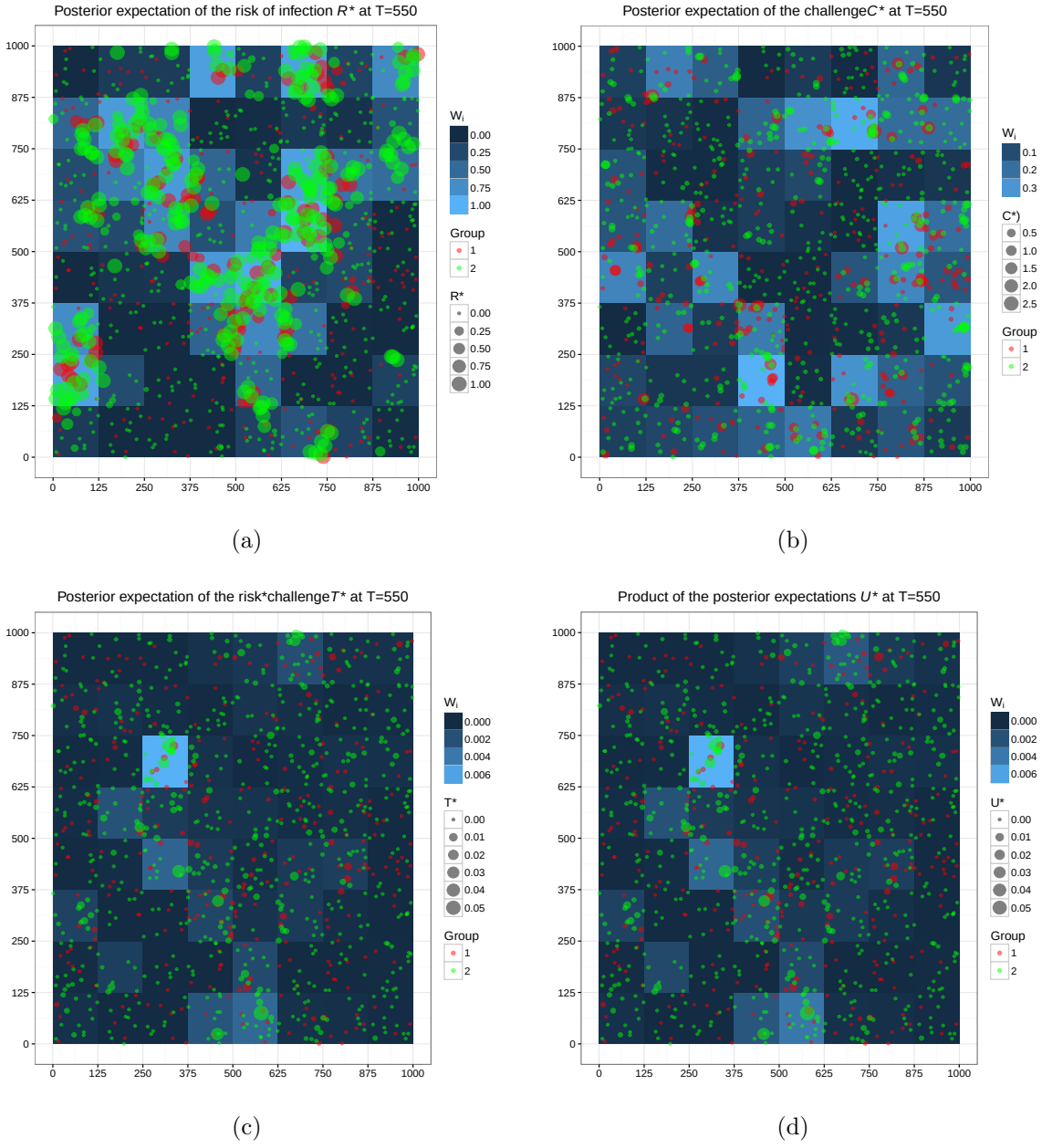


Figure 6.13: Case (II): Priority regions and the posterior expectation of the risk measure (a), posterior expectation of the challenge measure (b), expectation of the threat or expectation of the product of the risk and the challenge (c) and the product of the expectations (d) by day $T_a = 550$ using 200000 samples from $\pi_0(\boldsymbol{\theta}, \underline{x}(T_a)|\mathbf{y})$.

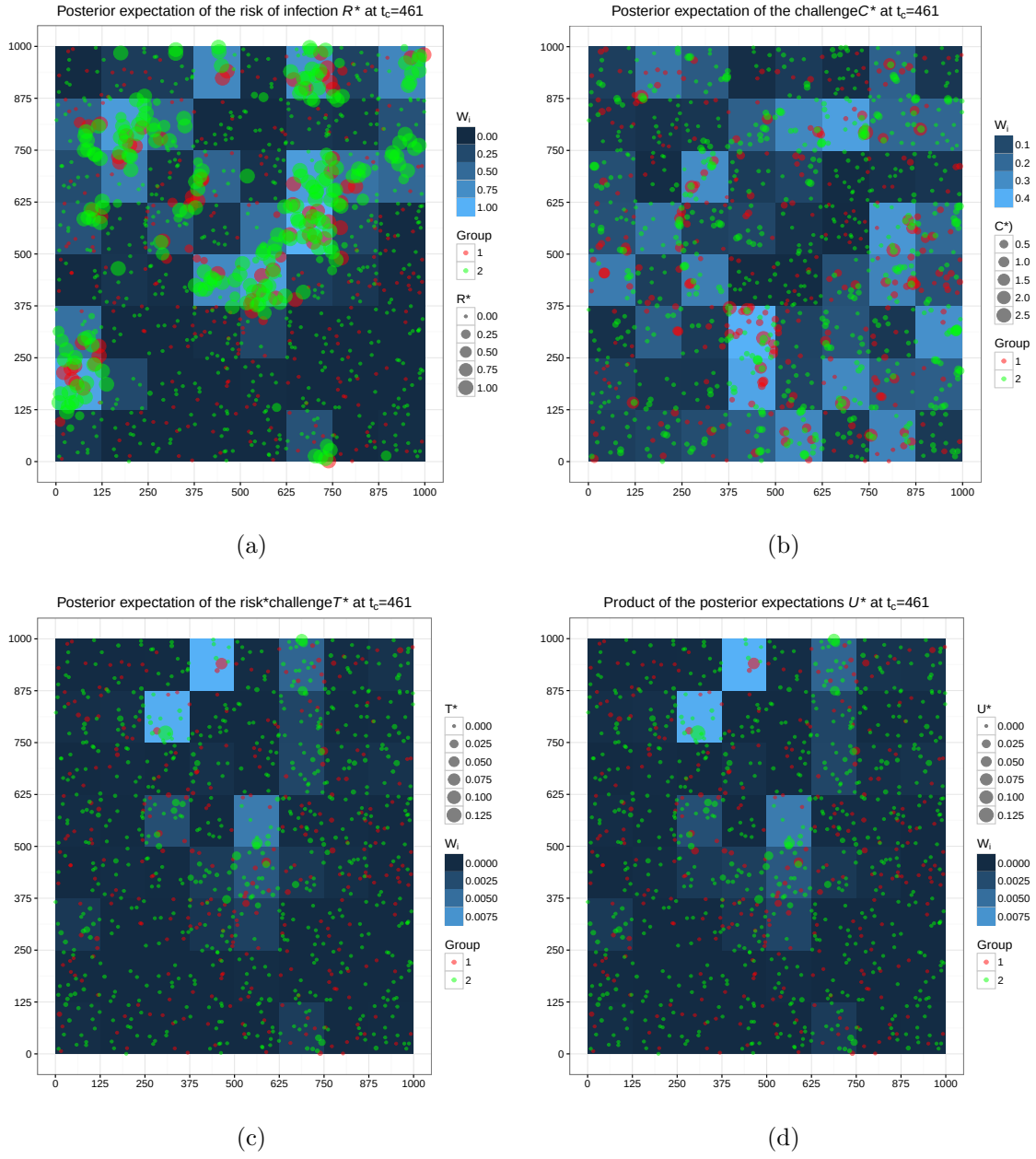


Figure 6.14: Case (II): Priority regions and the posterior expectation of the risk measure (a), posterior expectation of the challenge measure (b), expectation of the threat or expectation of the product of the risk and the challenge (c) and the product of the expectations (d) by day $t_c = 461$ using 200000 samples from $\pi_0(\boldsymbol{\theta}, \underline{x}(T_a)|\mathbf{y})$.

6.2.4.2 Controls and results

To test the efficiency of our control model, we partition the region into 100 and 64 squares regions (Figure 6.13 shows the latter case). We draw $\{(\boldsymbol{\theta}^i, \underline{x}^i(T))|i = 1, \ldots, 1000\}$ to generate $\{(\boldsymbol{\theta}^i, \boldsymbol{Q}^i)|i = 1, \ldots, 1000\}$ (samples from $\pi_0(\boldsymbol{Q}, \boldsymbol{\theta}|\mathbf{y})$). We assume various values of $N' = 100, 200, 300, 400, 500$, the maximum number of removals. In addition, the time of the intervention is assumed to be $t_c = 461$. For the optimisation, we consider the control strategies of the form $d_{\{\mathcal{M},2,2\}}$ and $d_{\{\mathcal{M},2,3\}}$

for different measure $\mathcal{M}$ constructed at both t_c and T_a as shown in Figures 6.13 and 6.14. For these measures, we use 30 realisations randomly drawn from $\pi_0(\boldsymbol{Q}, \boldsymbol{\theta}|\mathbf{y})$ to reach the convergence, leading to the optimal control strategies $d^*_{\{\mathcal{M},2,2\}}$ and $d^*_{\{\mathcal{M},2,3\}}$. Note that in each case, we constrain the number of regions to visit such that $n_r \leq 50$. These optimal control strategies along with the control $UR\{\mathcal{M}\}$ (obtained by ordering the first N' with respect to their measure $\mathcal{M}$) are then applied on the same 1000 realisations of the epidemic randomly sampled.

The resulting posterior distributions of the outcomes of each intervention for $N' = 500$ only are paired. Their joint distribution along with the box plots showing the variability of the expectation of their difference is shown on Figures 6.15. Similar to the homogeneous population case, the outcomes of the controls are strongly correlated, reducing the variability between their outcomes, leading to a more accurate estimation of the expectation of the difference between outcomes. Again, as a result, the sample size required to compare the effect of the competing strategies during the experiment is reduced.

We compare the effect of different maps on the posterior mean of the proportion of the number of infections avoided $\mathcal{P}$ considering the subregions described above (100 for Figure 6.17a and 64 for Figure 6.17b) and the unconstrained regions in Figure 6.16. Similar to the homogeneously mixing population case, results show that prioritisation based on the *threat* measures ($\mathcal{T}^*$) is more cost-effective in reducing the impact of the epidemic compared with other measures especially the risk ($\mathcal{R}^*$) and the challenge ($\mathcal{C}^*$) measures. We notice that there is no clear distinction between the effect of the prioritisation using either of $\mathcal{T}^*$ and $\mathcal{U}^*$. This is clearly shown in Figure 6.16. The reason being that even though the measures for each host are not equal, the priority regions coincide as shown in their respective maps in Figures 6.13c and 6.13d, and Figures 6.14c and 6.14d.

In addition, we compare the effect of the time at which the measures are constructed on the relationship between the posterior mean of $\mathcal{P}$ and N'. Similar to the homogeneous population without removal (chapter 5), we can see from the Figures 6.17b and 6.16 that, if the prioritisation is based on the $\mathcal{T}^*$ and $\mathcal{U}^*$, it is cost-effective to construct the measures at the assessment time. On the other hand, a prioritisation scheme using $\mathcal{R}^*$ and $\mathcal{C}^*$ offers greater reduction in the epidemic size when constructed at the intervention time. Similar to the homogeneous case in chapter 5, strategy 2 typically outperforms the strategy 3. Again this emphasizes the fact that host diversity impacts on the relative efficacy of the approaches proposed in this framework. Also, it shows that regardless of the approach used for the control, the threat map is a valuable tool to use to inform control policies.

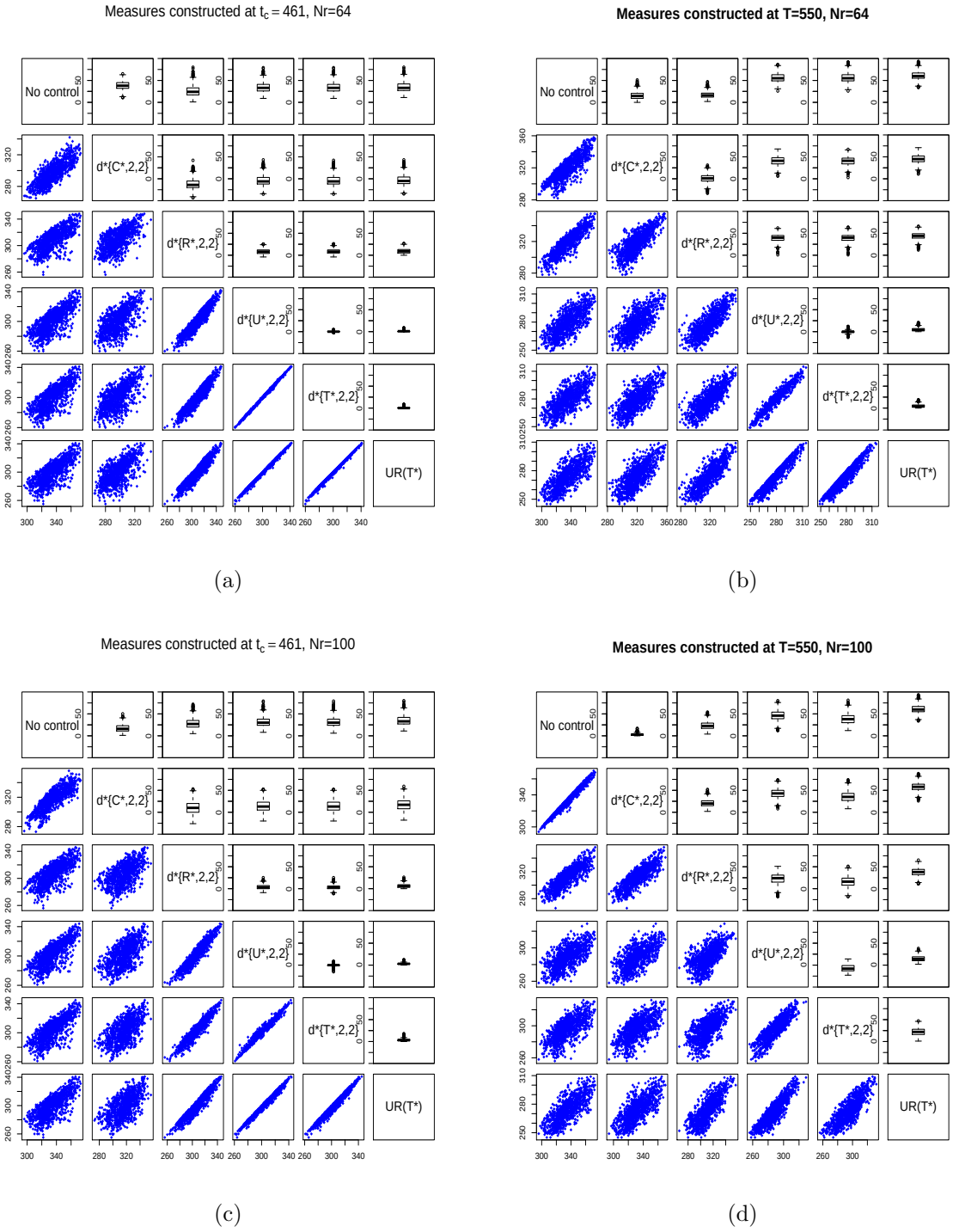


Figure 6.15: Pairwise scatter plots on the lower panel and box plots of the difference between the outcomes of the paired controls when the maps used are constructed at the intervention day $t_c = 461\,days$ and the assessment day $T_a = 550\,days$. (a) and (b) correspond to the 100 subregions (c) and (d) 64 subregions.

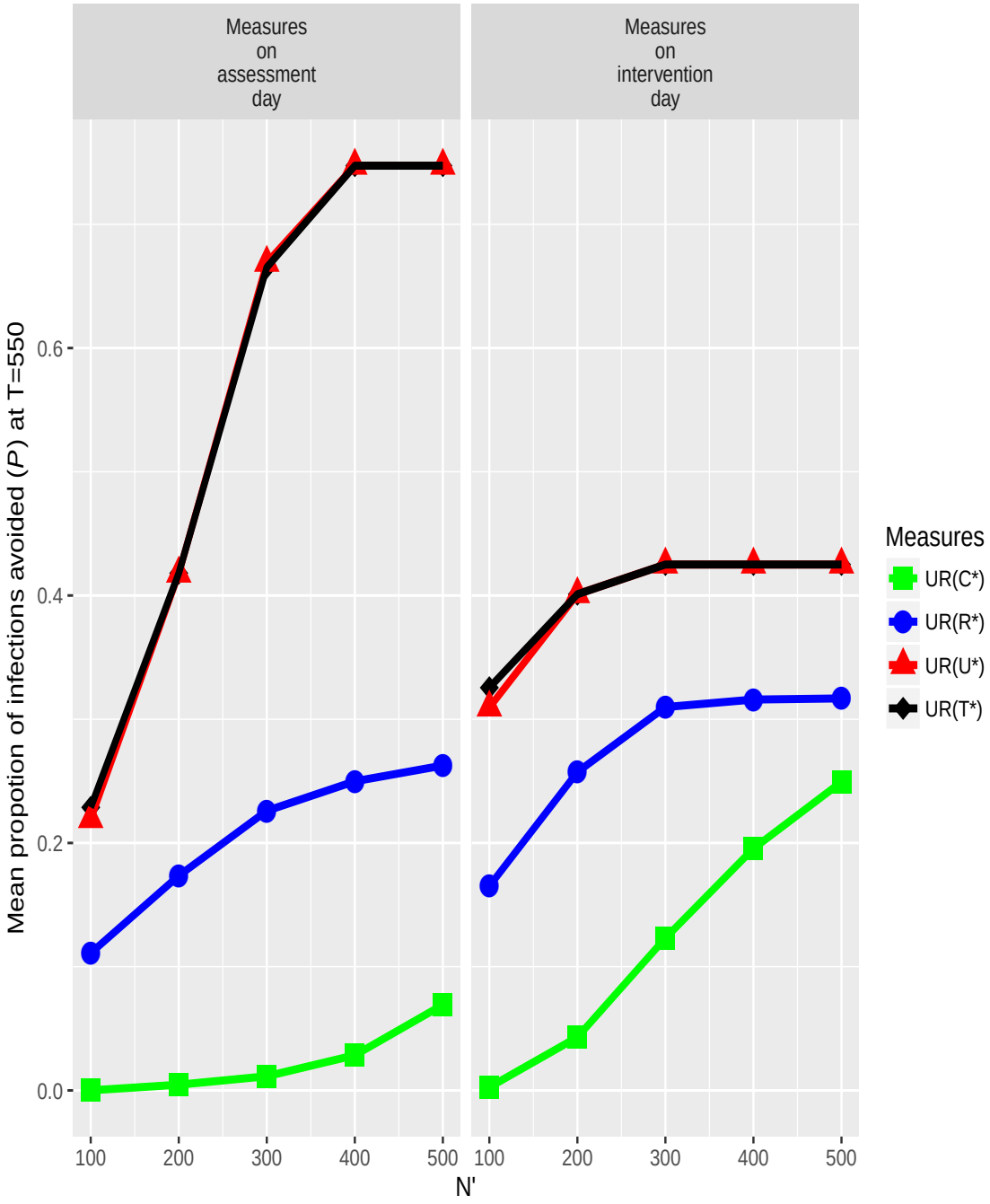
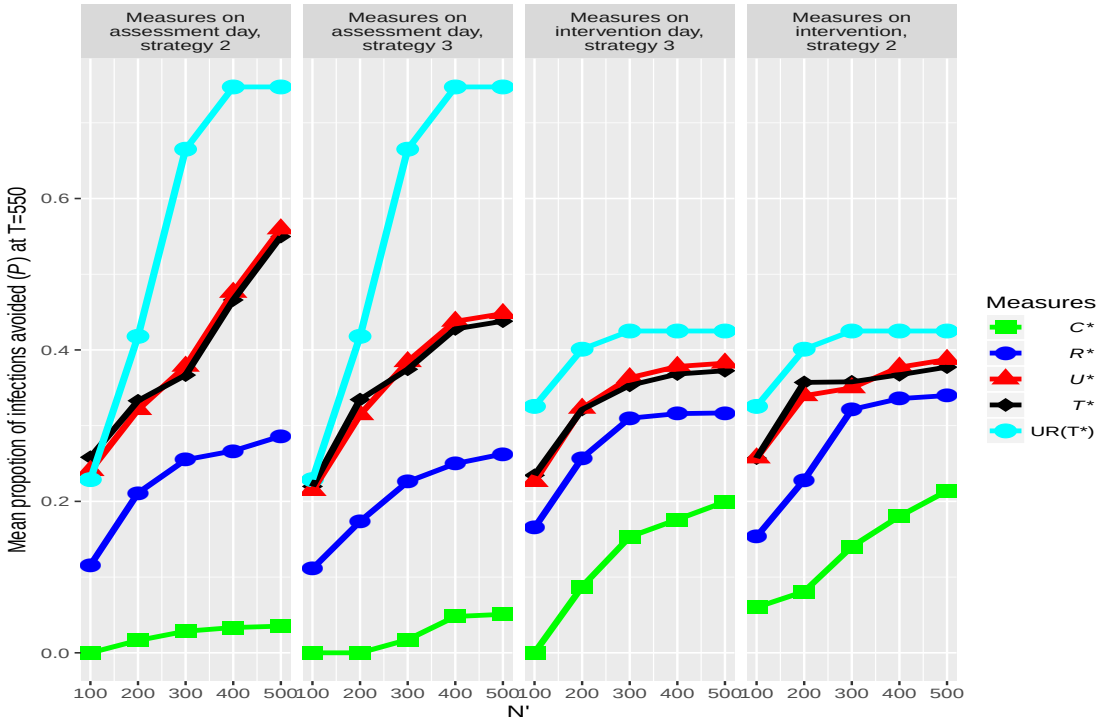
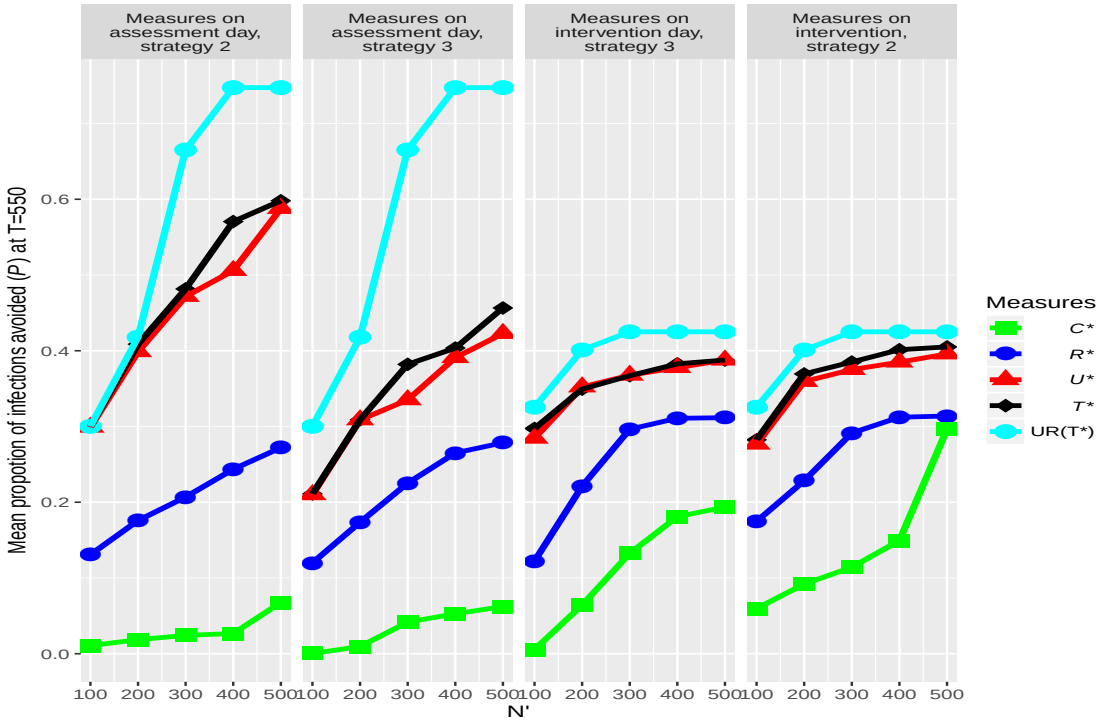


Figure 6.16: Effects of different measures and the size of the resources on the mean of proportion $\mathcal{P}$ using the unconstrained control.



(a)



(b)

Figure 6.17: Effects of prioritisation measures and the size of the resources on the posterior mean proportion of hosts rescued. (a) Considering 100 regions and (b) considering 64 regions

6.3 Conclusion

In this chapter, we first applied the coupling method described in Chapter 5 to a real-life epidemic data set regarding the 1995 Citrus Canker outbreak in Florida (Gottwald et al., 2002; Neri et al., 2014; Cunniffe et al., 2015) before moving to extend the approach to a more realistic framework assuming an heterogeneous population where individuals differ not only in term of their susceptibilities to contract the disease, but also their ability to transmit it. In this study, the primary aim was to identify the benefit of coupling epidemics with different control strategies.

One of the more significant findings to emerge from this study is that, the outcomes of epidemics under different control strategies are highly positively correlated (see Figures 6.5a, 6.5b, 6.15) leading to a reduction in the number of simulations required to compare different strategies. Taken together, these results suggest that the variability in the difference between outcomes of paired controls is reduced compared to independent sampling. This will provide a better estimate of their expected difference.

This study also suggests that regardless of host diversity, the prioritisation for control based on removal when identified as infective is more cost-effective in term of reducing the expected outcome of the epidemic if the control is designed using the *threat* map which is the posterior expectation of the the risk of being infected and the challenge it will pose once infected. It also suggests that both host diversity and removal process play apart in the choice of when to construct the maps.

Chapter 7

Modelling and control of the Pest des Petits Ruminants (PPR)

7.1 Introduction

This chapter describes work undertaken for the European Food Safety Authority (EFSA) under the project MAPRA: Modelling Animal Pathogens: Review and Adaptation.

The ultimate purpose of the modelling from the EFSA perspective is to understand the risk of an outbreak of Peste de Petits Ruminants (PPR) occurring in a given country and to understand the risk of subsequent spread to other countries in the network. Although the goals relate to the dynamics of the disease in a large-scale network of interacting countries, to understand transmission from one node to another in this network requires consideration of the spatio-temporal characteristics of spread within the country corresponding to the node. The challenge is then to see how such an understanding at the local level might be scaled-up to predict the risk of introduction across countries. The immediate goals of the work reported here were to provide a modelling framework that could be used to integrate data, much of which may not yet be available, from various sources to build a quantitative understanding of PPR dynamics in a network of countries to support the longer term aim of building a predictive model of spread that can be used to assess the risk of incursion into control strategies.

Data on the disease considered in this chapter come from Tunisia and consist of information on the occurrence of outbreaks at farm level, with each outbreak being characterised by the start time, end time, size of outbreak and size of susceptible population. The locations of infected premises were also available for each outbreak. Unlike all the other examples considered in previous chapters the data on PPR do not include details on the locations of the uninfected susceptible population. The impact of not considering unobserved susceptibles in spatio-temporal analyses of epidemics

has been demonstrated by Lau et al. (2015) and a particular goal was the development of methods that could accommodate the lack of information on the susceptible population in a systematic way, so that spatio-temporal models could nevertheless be fitted to the outbreak data. In Section 7.2 we describe PPR briefly before describing the spatio-temporal model used to model within-country spread and, in particular, how we attempt to accommodate the lack of information on the susceptibles using contact distribution models in Section 7.3. Bayesian methods for fitting the model to data are described in Section 7.4. In 7.5 we use the fitted model to generate a predictive distribution of wave speeds comparing our results with the results of an earlier analysis by EFSA of the data from Tunisia using an ad hoc method that did not attempt to model the dynamics of spread. In 7.6 we describe a network model for the spread of PPR which represents a framework for scaling the within-country models up to the continental scale. In particular we show how algorithms from graph theory can be used to give efficient methods for simulating the network model and present some initial results using the model. Finally in 7.7 we discuss how control measures may potentially be compared with the network by coupling epidemics using latent processes (in a manner analogous to earlier chapters).

7.2 Background

Peste des petits ruminants (PPR), also known as 'goat plague' (OIE, 2015) is a highly contagious disease of sheep, goats and, occasionally, wild small ruminants caused by a *Morbillivirus* in the family of Paramyxoviridae. It was first reported in Ivory Coast in 1942 and subsequently spread into other West African countries (OIE, 2015). Figure 7.1 shows the geographical distribution of Peste des Petits Ruminants for the period $2005 - 2013$. At a time, peste des petits ruminants was thought to be a West African disease but later has been reported in North Africa, Middle East as well as Asia. PPR was originally considered as rinderpest due to the similarity between their symptoms; however, this consideration was ruled out when it was noticed that cattle do not become diseased when exposed to ruminants infected with rinderpest (FAD PReP and USDA, 2013).

7.2.1 PPR characteristic

7.2.1.1 Transmission

The virus is mainly contained in discharges from eyes, nose and mouth, secretions from coughing and in faeces of infected animals. PPR virus is also transmitted through inhalation of fine droplets released into the air, particularly when infected animals cough and sneeze. Therefore, close contact between animals is the main route of

spread of the virus. However, fomites such as bedding, water and feed troughs can also be contaminated by urine and secretions and, therefore, become a source of infection (FAO, 1999; FAD PReP and USDA, 2013; OIE, 2015), though the virus only survives in the environment for a short period of time (FAD PReP and USDA, 2013).

7.2.1.2 Incubation period and symptoms

It is reported that the incubation period ranges between 2 to 10 days (EFSA, 2015; CFSPH and IICAB, 2008), while the symptoms can take 2 to 6 days to be become visible (FAO, 1999; CFSPH and IICAB, 2008; EFSA, 2015). This is followed by the sudden onset of fever, severe depression, respiratory symptoms, congestion and necrosis of mucous membranes. Severe diarrhoea is developed in some animals causing dehydration and weight loss. In most cases, animals can have a nasal discharge causing respiratory distress. In the advanced stage of the disease, pneumonia is common among the diseased animals. Abortion can be noted in pregnant animals. Death may occur within 2 to 10 days following the date of onset of symptoms.

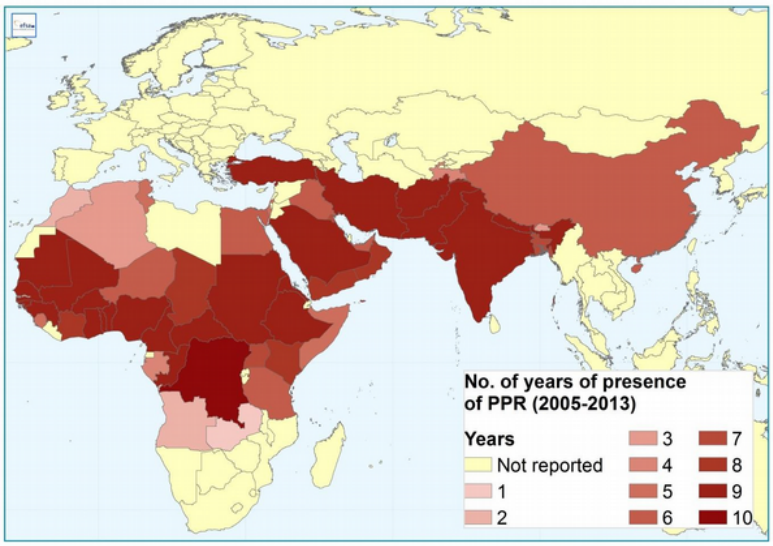


Figure 7.1: Number of years of presence of PPR in different countries as reported to OIE during 2005-2013 EFSA (2015).

7.2.1.3 Control

PPR is classified by FAO as one of the Transboundary Animal Diseases (TADs) defined by FAO as those diseases that are of significant economic, trade and/or food security importance for a considerable number of countries; which can easily spread to other countries and reach epidemic proportions; and where control/management, including exclusion, requires cooperation between several countries. As a TAD, it is

one of the priority animal diseases, whose control is considered important for poverty alleviation in the infected regions. There are no vaccines available that can eradicate the disease but there are means to decrease its spread. Since the transmission of the virus only occurs through direct contact with infected animals, the main control approaches include a combination of quarantine, pre-emptive culling, animal movement control and the use of 'ring' vaccination and vaccination of the high-risk population (FAD PReP and USDA, 2013). Because of its similarity with the Rinderpest virus, rinderpest vaccine has been used on some occasions to attempt to slow the spread.

7.2.2 Challenge of modelling PPR

The work focuses on the use of data on the disease in Tunisia which consist of information on occurrence of outbreaks at the farm level with each outbreak being characterised by the start time, end time, size of outbreak and size of susceptible population. Data on PPR outbreaks in other countries were also available through EFSA including, for example, spatial locations of reported outbreaks in Turkey from 2011 to 2014 along with information on patterns of animal movement. However, given time constraints, we focus on the Tunisia data here.

The ultimate purpose of the modelling from the EFSA perspective is to understand the risk of an outbreak occurring in a given country and to understand the risk of subsequent spread to other countries in the network. Although the goals relate to the dynamics of the disease in a large-scale network of interacting countries, it is arguable that to understand transmission from one node to another in this network requires consideration of the spatio-temporal characteristics of spread within the country corresponding to the node. The challenge is then to see how such an understanding at the local level might be scaled-up to predict risk of introduction across countries.

7.2.3 Previous PPR modelling

In contrast to other diseases considered in this thesis, there was little prior modelling on which the models developed here could be built, given the incomplete nature of the data. Also, the application of vaccine in most regions where PPR is reported (in most African countries) complicates any modelling. The only attempt at mathematical modelling to our knowledge is the one in EFSA (2015). They used the method derived by Koeijer et al. (1998) to estimate the basic reproduction number R_0 given as

$$R_0 = \frac{(1-f)\log(y)}{\log(x)} \tag{7.2.1}$$

where f is the mortality of infected animals, x the fraction of the population that survives the epidemic and y the fraction of the population that does not get infected.

However, the authors themselves recognised that such a model does not reflect reality. Also, this model implies complete knowledge of the initial population which is not often the case.

Also, another modelling attempt was implemented in EFSA (2015) where some quantitative work on estimating wave speeds using the Tunisia data and modelling transmission through animal movements was carried out. Even though the results presented are sensible, there is no clear mathematical description of how they were derived; more precisely, an explicit spatio-temporal dynamical model was not formulated. In this chapter not only do we consider such models but also consider how they may be used to inform the construction of network models for the spread of PPR through a network of countries at the continental scale.

7.3 Identifying Modelling Approaches

Consideration is given to identifying general aspects of modelling approaches that would most likely be appropriate for PPR modelling, taking account of the goals outlined in the previous Section, the available data and, importantly, the unobserved processes on which conclusions will depend. It was decided that attention should focus on modelling spread within a country and on identifying methods of building these models into a model for PPR spreading through a network of countries, as the incursion of PPR into uninfected countries was seen as the process of prime interest for this case study.

7.3.1 Model structure

Concerning the spread of PPR within a country there are several ways in which the susceptible population may be modelled. Data from Tunisia provide information on sub-populations infected, suggesting that a meta-population approach where susceptibles are modelled as a network of interconnected nodes would be appropriate. Although not pursued in this study, we note that this framework would offer the scope to capture heterogeneity across nodes, representing the differing species make-up of individual farms. The strength of connections in this network would then be influenced by the perceived strength of the infectious challenge between nodes. A natural initial first approach to this would be to represent the strength of interactions as being monotonically related to Euclidean distance.

Anticipating the problems regarding lack of knowledge on the susceptible population, it is also beneficial to consider frameworks where the susceptible population is represented via a continuous spatial intensity, which then moderates the occurrence of new infections. In particular it is possible that a framework in which spatial contact

distribution models (described later) for the initiation of new infections are used to formulate a marked point process model in which marks are drawn from a distribution representing the diversity of units forming the susceptible population may be appropriate, and may offer a way of accommodating the lack of information on the susceptible population.

7.3.2 Inference and parameter estimation

It was identified as important that parameter uncertainty can be treated properly for any models that are fitted to data on PPR, so that these uncertainties can be propagated through to predictions. With this in mind particular consideration was given to finding models that can be fitted using likelihood-based methods (to extract maximum information from data) or a fully Bayesian approach, for uncertainty propagation. As already mentioned it was considered important that any approach to model fitting could cope with unobserved processes, or with the increased model complexity that arises for example from including parameters for probability of non-reporting of infections, with methods that use data on final size for the inference of epidemic parameters potentially relevant.

7.4 Models for within-country spread

7.4.1 Model structure

We represent the spread of PPR through a network of premises where only infections are recorded, so that no information on susceptible premises is available. We treat the farm or premises as being the population unit so that, in effect, we are considering a meta-population model for the system. A key feature of our modelling approach is that we avoid the need to represent the susceptible population by, in a sense, combining the generation of the susceptible population with the spread of the disease within that population. This greatly simplifies the challenge of constructing and fitting a model.

The main assumptions in the contact distribution model are as follows:

- On infection, premises remain infectious for a fixed period, γ (which is unknown) for all infection premises.
- Infected premises cause new infections according to a Poisson process with fixed rate, β.
- Each new infection is located at random distance r from the source premises, drawn from density, $r \sim f(r; \alpha)$ placed at a uniformly distributed bearing from the source.

- The unknown susceptible population is implicitly modelled in this approach as a spatial Poisson process.

We note that the representation of the susceptible population means that the effect of depletion of the susceptible population is not explicitly represented in the population, and ignores certain temporal dependencies that should occur in a given neighbourhood, since the number of premises in a neighbourhood remains constant. However, the approach is valid for the early stages and linear phase of the epidemic and is arguably appropriate for the estimation of speeds of the disease wave-front.

For a given choice of function f, there are three key parameters in this model the dispersal rate β, the infectious period γ and the parameter α that controls the spatial dispersal process. All of these combine in the estimation of spatial spread rates. Note that β combines the effect of the rate at which new infectious challenges are generated by a given source and the rate in the Poisson process that is assumed to generate the susceptible population. We have proposed a number of candidate models for f in the contact distribution model. The main assumptions are as follows:

$$\begin{aligned}
\textbf{Rayleigh}(\alpha)\textbf{:} \qquad & f(r;\alpha) = 2r\alpha \exp\left(-\alpha r^2\right) \\
\textbf{Exponential}(\alpha)\textbf{:} \qquad & f(r;\alpha) = \alpha \exp\left(-\alpha r\right) && (7.4.1) \\
\textbf{Cauchy}(\alpha)\textbf{:} \qquad & f(r;\alpha) = \frac{2}{\pi\alpha\left(1+\frac{r^2}{\alpha^2}\right)} && (7.4.2)
\end{aligned}$$

in which r is the Euclidean distance between a given pair of premises and measured in km. Note that these $1-$dimensional kernels are isotropic defined on the positive real axis. These models represent scenarios in which the propensity for long-range interactions increases with the Cauchy model exhibiting jumps in the infection process that are very long-range. The kernels defined above ensure that different pattern of the epidemic could be compared directly. While thin-tailed kernels such as Rayleigh and exponential kernels give rise to a slow spreading wave of new infections, the thick-tailed kernels such as Cauchy result in a rapid and long dispersal ahead of the source (Mollison, 1977; Clark et al., 1999). Techniques developed here can readily be adapted to other models in addition to the three variants considered above.

7.4.2 Data and inference methods

The approaches followed can be applied in any setting where spatial locations of infected premises and onset times of infectiousness are recorded (and treated as a proxy for infection time). They can in principle be extended to situations where

infection times are not recorded. The data from Tunisia (see Figure 7.2) are used to illustrate the approach and we denote this data-set by $\mathbf{y}$. Analyses were initially carried out assuming that the infectious period γ was known, but this assumption is relaxed in analyses presented here where γ is inferred.

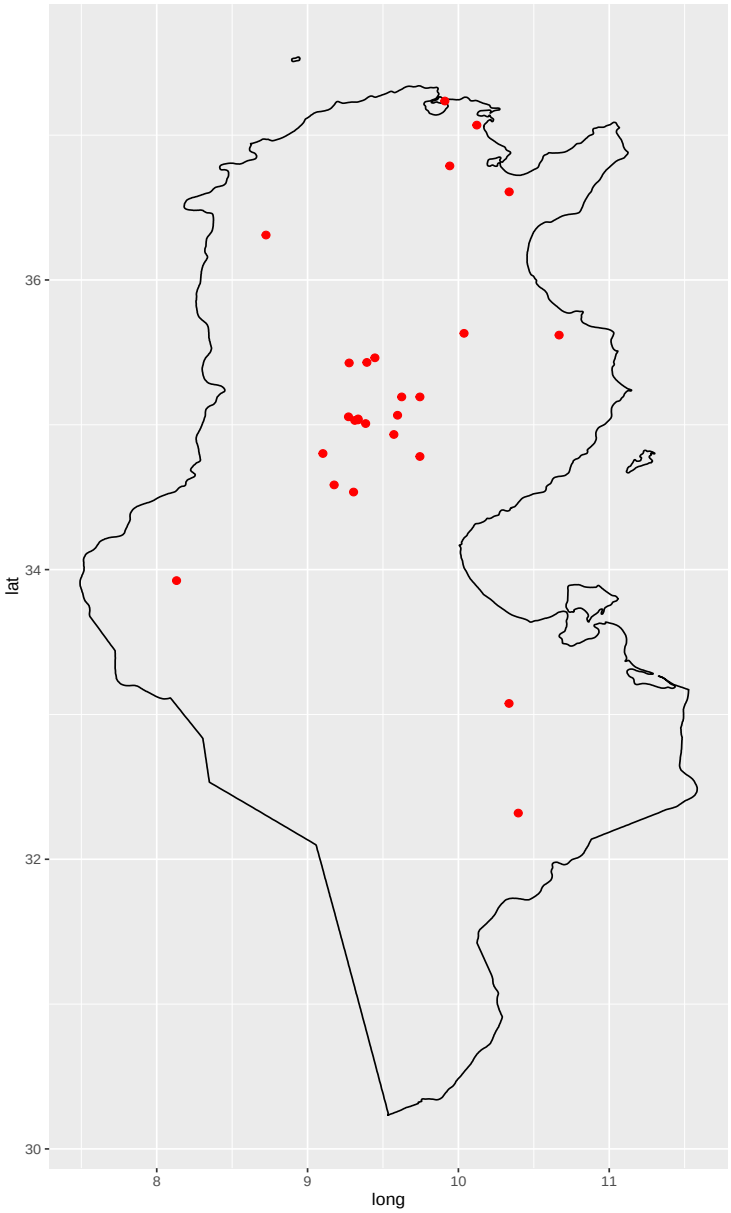


Figure 7.2: Locations of PPR outbreak in Tunisia.

.

7.4.2.1 Bayesian imputation of the network

In the Bayesian approach, a prior distribution $\pi(\alpha, \beta, \gamma)$ is placed on the unknown parameters. The likelihood function $\pi(\mathbf{y}|\alpha, \beta, \gamma)$ is not the most convenient to compute due to the fact that the network of infections (who infected whom) denoted by s here

is not observed. However, using the data-augmented approach we can overcome this problem by treating the infectious network as an additional unknown parameter and investigating the joint posterior $\pi(\alpha, \beta, \gamma, \mathbf{s}|\mathbf{y})$. Markov chain Monte Carlo (MCMC) methods are well suited to solving such problems. We note that the inclusion of $\boldsymbol{s}$ as parameter allows us to use Gibbs steps in some of the updates.

7.4.2.2 Markov chain Monte Carlo Algorithm

As in the previous chapters, key parameters of the process α, β, γ are estimated in the Bayesian framework using MCMC coupled with data augmentation methods. Let $[0, T]$ denotes the observation period and let t_i denotes the infection time of the premises i. We denote by $\mathbf{y} = (t_1, \ldots, t_N)$ the ordered infection times, where $t_1 = 0$ without loss of generality. The network of transmissions is specified by the set of sources for each infection occurring after the first infection and this is denoted by $\mathbf{s} = (s_2, \ldots, s_N)$, where s_i denotes the source of the infection occurring at time t_i. If the transmission network (the set of sources) is treated as a parameter, then the likelihood is given by

$$L(\alpha, \beta, \gamma, \mathbf{s}) \propto \exp\left(-\beta \int_0^T n(t)dt\right) \prod_{i=2}^{N} \beta \frac{f(r_i; \alpha)}{r_i}, \qquad t_i - t_{i-1} < \gamma \forall i = 1, \ldots N. \tag{7.4.3}$$

where $n(t)$ is the number of infectious premises at time t and r_i denotes the distance between the premises infected at time t_i and the source premises as specified by s_i. Had these sources been observed then parameters (α, β, γ) could be estimated using for example the maximum likelihood method. In the absence of this information we can nevertheless investigate the posterior distribution of the augmented parameter $(\alpha, \beta, \gamma, s)$ in the Bayesian framework using MCMC.

To complete the Bayesian framework, we need to assign prior distributions to the transmission rate β, the dispersal parameter α and the infectious period γ. We propose independent vague Gamma priors for parameters $\alpha \sim \text{Gamma}(a, b)$ and $\beta \sim \text{Gamma}(c, d)$, and a uniform distribution for the infectious period $\gamma \sim U[e, f]$. From Bayes' theorem, the joint posterior distribution is then given by,

$$\pi(\alpha, \beta, \gamma, \mathbf{s}|\mathbf{y}) \propto L(\alpha, \beta, \gamma; \mathbf{s}) \beta^{c-1} \alpha^{a-1} \exp(-d\beta - b\alpha), \quad \alpha, \beta > 0,$$
$$e < \gamma < f \tag{7.4.4}$$

and can be sampled using the Markov chain. Inference on individual parameters is made by considering the marginal posterior distribution of the parameter from the simulations from the joint posterior distribution.

7.4.2.3 Construction of the Markov chain

A single iteration of the Markov chain requires making updates to each of the parameters in the augmented parameter vector. Where it is straightforward to sample directly from the posterior density of a parameter conditional on the values of all other parameters then that parameter can be updated using a Gibbs step, whereby the updated value is simply a draw from the respective conditional density. If this cannot be effected for a given parameter, then knowledge of the posterior density up to unknown normalising constants allows Metropolis-Hastings methods to be used.

Conditional distribution of β

The full conditional distribution of the transmission rate density (given all other parameters) is given by:

$$\beta|\gamma, \mathbf{y} \sim Gamma(N + c, \int_0^T n(t)dt + d) \tag{7.4.5}$$

Conditional distribution of α

The full conditional distribution of the dispersal parameter α depends on the particular kernel function f that is chosen.

- For the Rayleigh kernel considered initially, we have

$$\alpha|\mathbf{s}, \mathbf{y} \sim Gamma(N + a, b + \sum_{i=2}^{N} r_i^2) \tag{7.4.6}$$

- For the exponential kernel considered later, we have

$$\alpha|\mathbf{s}, \mathbf{y} \sim Gamma(N + a, b + \sum_{i=1}^{N} r_i) \tag{7.4.7}$$

- For the Cauchy kernel, using a uniform prior $[0, 1000]$ in this case, we have

$$p(\alpha|\mathbf{s}, \mathbf{y}) \propto \prod_{i=1}^{N} \frac{1}{\alpha\left(1 + \frac{r_i^2}{\alpha^2}\right)} \tag{7.4.8}$$

Conditional distribution of the infectious period γ

The full conditional distribution of the infectious period is given by

$$\pi(\gamma|\beta, \mathbf{y}) \propto \exp\left(-\beta \int_0^T n(t)dt)\right) \tag{7.4.9}$$

where the dependence on γ on the right-hand side arises through the dependence of $n(t)$, the number of infected premises at time t, on the infectious period.

Conditional distribution of the infecting sources s.

It is possible to identify formally the conditional distribution of $\mathbf{s}$ given the other parameters as

$$\pi(\mathbf{s}|\alpha,\gamma,\mathbf{y}) \propto \prod_{i=2}^{N} \frac{f(r_i;\alpha)}{r_i} \tag{7.4.10}$$

where the r_i is the distance between the premises infected at time t_i and the infecting source s_i.

7.4.2.4 Parameter updates

Using the conditionals identified above it is possible to use Gibbs steps (i.e. updating a parameter by directly drawing from its conditional distribution) to update β, and to update α for the case of the Rayleigh and Exponential kernels. However it is not straightforward to draw samples directly from the conditional distributions of α in the Cauchy kernel or of γ.

For these parameters we therefore use a Metropolis update with Gaussian and Uniform proposal distribution for α and β respectively, centred on the current value of these parameters. The acceptance probability for a proposed move of this form is given by

$$p_{acc} = \min\left\{1, \frac{\pi(\boldsymbol{\theta}^{new})}{\pi(\boldsymbol{\theta}^{old})}\right\} \tag{7.4.11}$$

where π here denotes the conditional density of the parameter $\boldsymbol{\theta}$ being updated.

To update the transmission network, an infected premises j is selected at random from infections $2,\ldots,N$. A new source s_j is proposed uniformly from the set of premises that are infected at time t_j. Specifically, this is the set of sources with an infection time t_i such that $t_i < t_j < t_i + \gamma$. The proposed source is then accepted with probability

$$p_{acc} = \min\left\{1, \frac{f(r',\alpha)}{f(r_j,\alpha)}\frac{r_j}{r'}\right\} \tag{7.4.12}$$

where r_j and r' denote the distance between site j and the current source and proposed source respectively.

7.4.3 Results on Tunisian Data

The model is applied on the Tunisia data described above. We reflect our vague prior knowledge on the parameters α and β by setting $a = c = 10^{-4}$ and $b = d = 10^{-6}$ but we choose an informative prior belief for the infectious period by setting $e = 0$ and $f = 120$. This choice allows premises to remain infectious for a period up to four months as opposed to three months period considered in EFSA (2015) since there is no evidence showing an external source of infection for the 4^{th} infection which occurred 108 days after the third one.

The MCMC algorithm is iterated for 100000 iterations, discarding the first 10000 iterations (the burn-in period) to ensure that convergence to the stationary distribution is reached. Trace plots in Figure 7.3 demonstrate that the chains are mixing well and there is no sign of non-convergence. The Tunisian data are analysed using the model with the three choices of kernel. For each model, the marginal posterior distributions of α, β, and γ are presented in the table 7.1. Note that to allow comparison of parameters α in the spatial kernel models we re-parameterise the exponential and Rayleigh kernel in terms of new parameters: $\alpha' = \frac{1}{\sqrt{2\alpha}}$ for the Rayleigh and $\alpha' = \frac{1}{\alpha}$ for the Exponential. We remark that owing to the parameterisation chosen, the kernel parameter in each model is km and the rate of infection β is measured in $days^{-1}km$. The original parameterisation of the Rayleigh and Exponential models is of course more convenient for Bayesian analysis due to the conjugacy of the Gamma prior, but we will present posterior distributions by transforming to the new parameterisation.

The posterior distributions and the summary statistics are shown in Figure 7.4 and in Table 7.1 respectively. These results indicate that the posterior distributions of β and γ are stable with respect to the choice of spatial kernel. The parameter α has a posterior distribution that varies with the choice of kernel. Of course the interpretation of this parameter is specific to the kernel choice so that the stability observed for other parameters should not be expected.

In Section 7.5, samples from the joint posterior density are used to estimate rates of spatial spread of PPR for use in the network model considered later for the spread of PPR through a network of countries. As observed in Section 7.5, there is considerable sensitivity in these predictions to the choice of spatial kernel.

All models suggest that the infectious period distribution present several peaks. This is essentially due to the clustering of infection times as shown by the data.

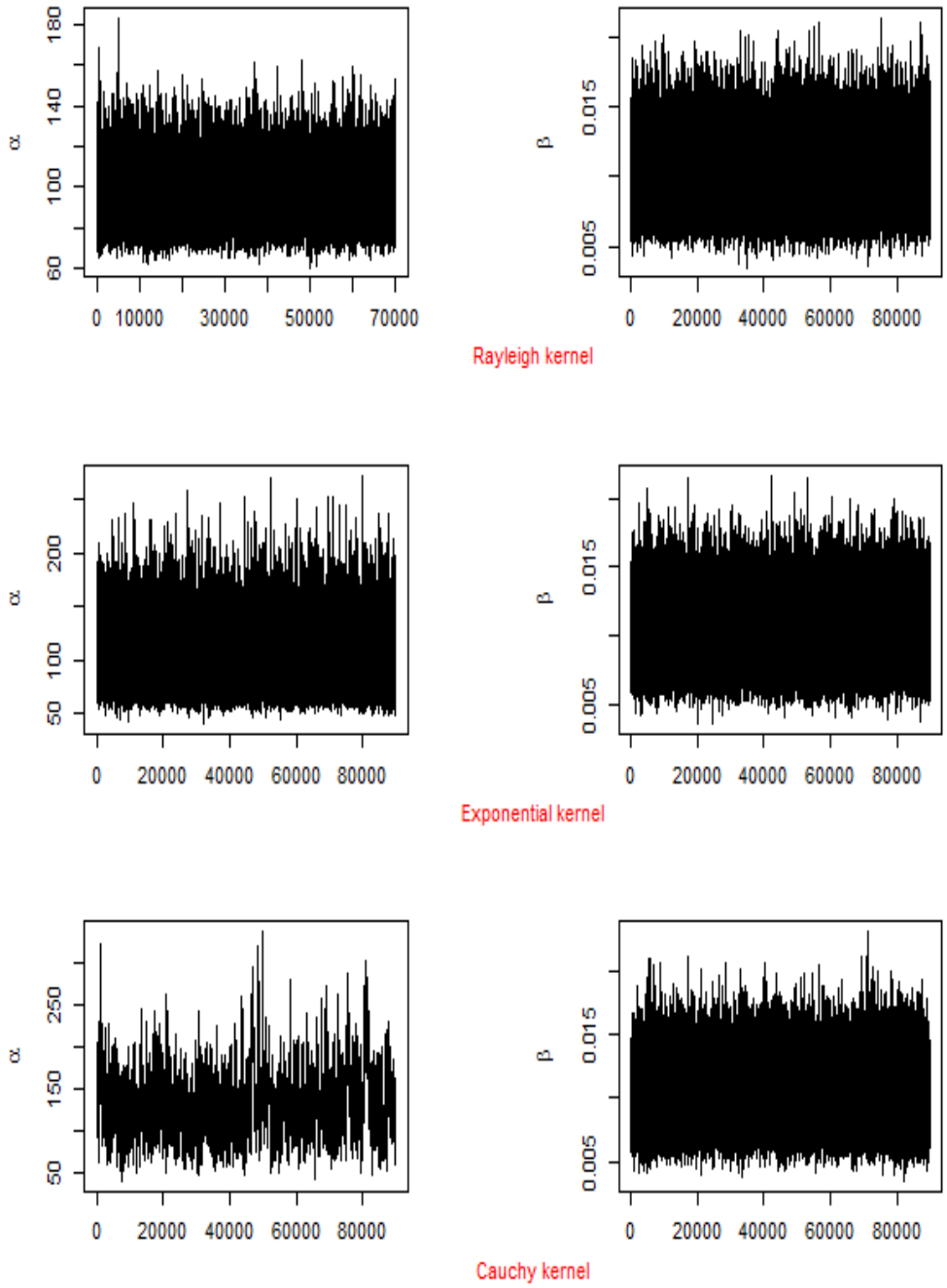
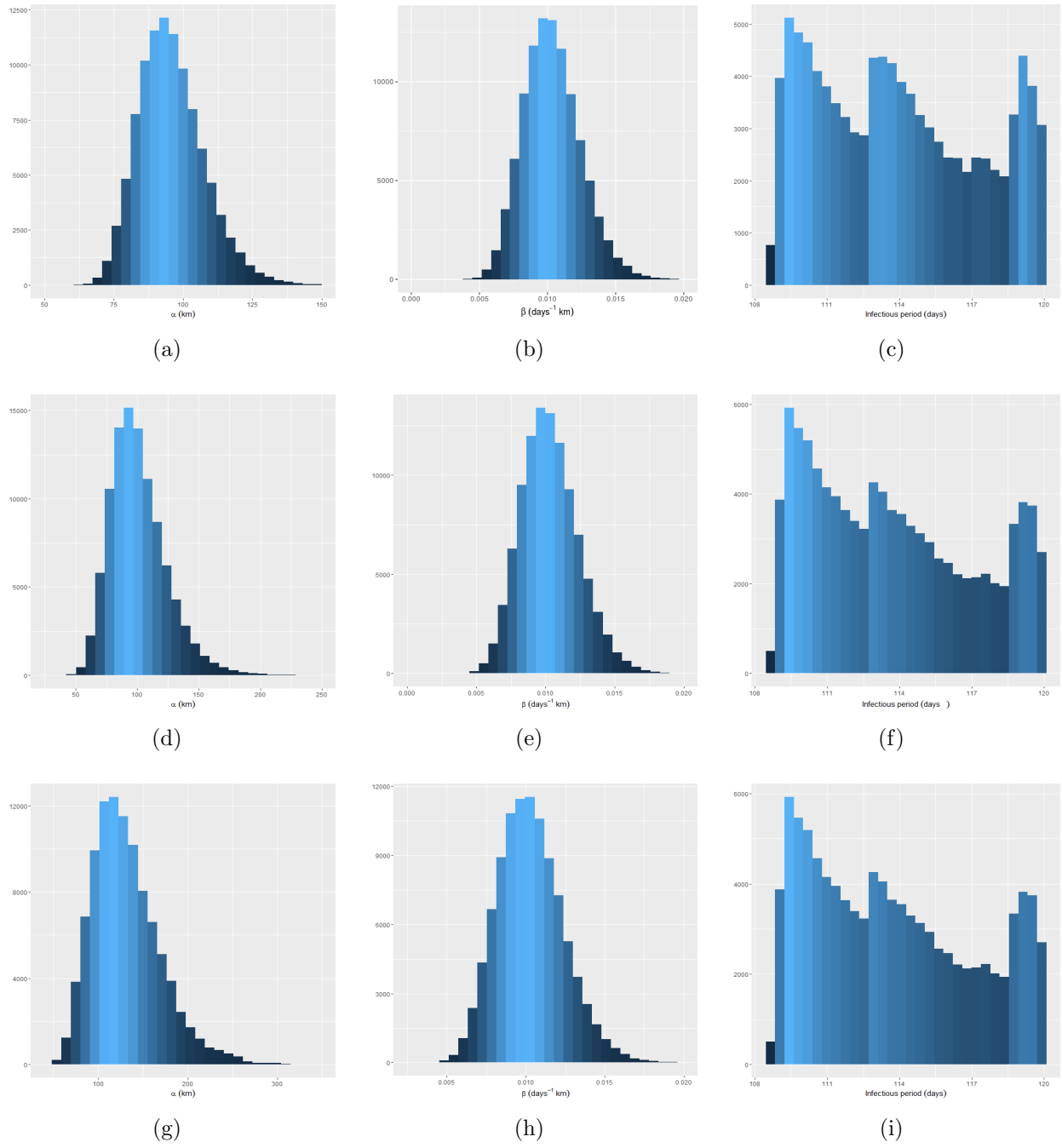


Figure 7.3: Trace plots of the model parameters showing the convergence of the chains



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 7.4: Posterior distribution of the model parameters α, β and γ for different kernels. The first, second and third row correspond to Rayleigh, exponential and Cauchy kernels respectively

7.4.4 Testing methods of inference

A standard approach to testing MCMC methods is used; namely to use the techniques to estimate parameters using data simulated from the fitted model with known parameter values. By benchmarking posterior densities against these known values, it

	Mean	SD	MC Error	0.025	Median	0.975
			Rayleigh			
α	95.78	11.96	0.07	75.72	94.66	122.34
β	0.01	0.006	2.44e-05	6.607e-03	0.01	0.0147
γ	113.92	3.34	0.09	1.09e+02	113.5e	119.8
			Exponential			
α	100.975	22.648	1.126e-01	65.72	97.864	154.248
β	0.01	0.0022	1.24e-05	6.67e-03	0.01	0.0148
γ	113.593	3.293	8.46e-02	1.09e+02	113.157	119.77
			Cauchy			
α	134.244	39.182	2.217	7.442e+01	128.02	231.349
β	0.01	0.0022	1.09e-05	6.642e-03	0.01	0.015
γ	113.65	3.364	9.095e-02	1.09e+02	113.172	119.791

Table 7.1: Mean, standard deviation, Monte Carlo error and quantile function evaluated at $(0.025, 0.5, 0.975)$ for marginal posteriors of model parameters as estimated using MCMC for three different choices of kernel function. Plots of marginal distributions of these parameters are displayed in 7.4.

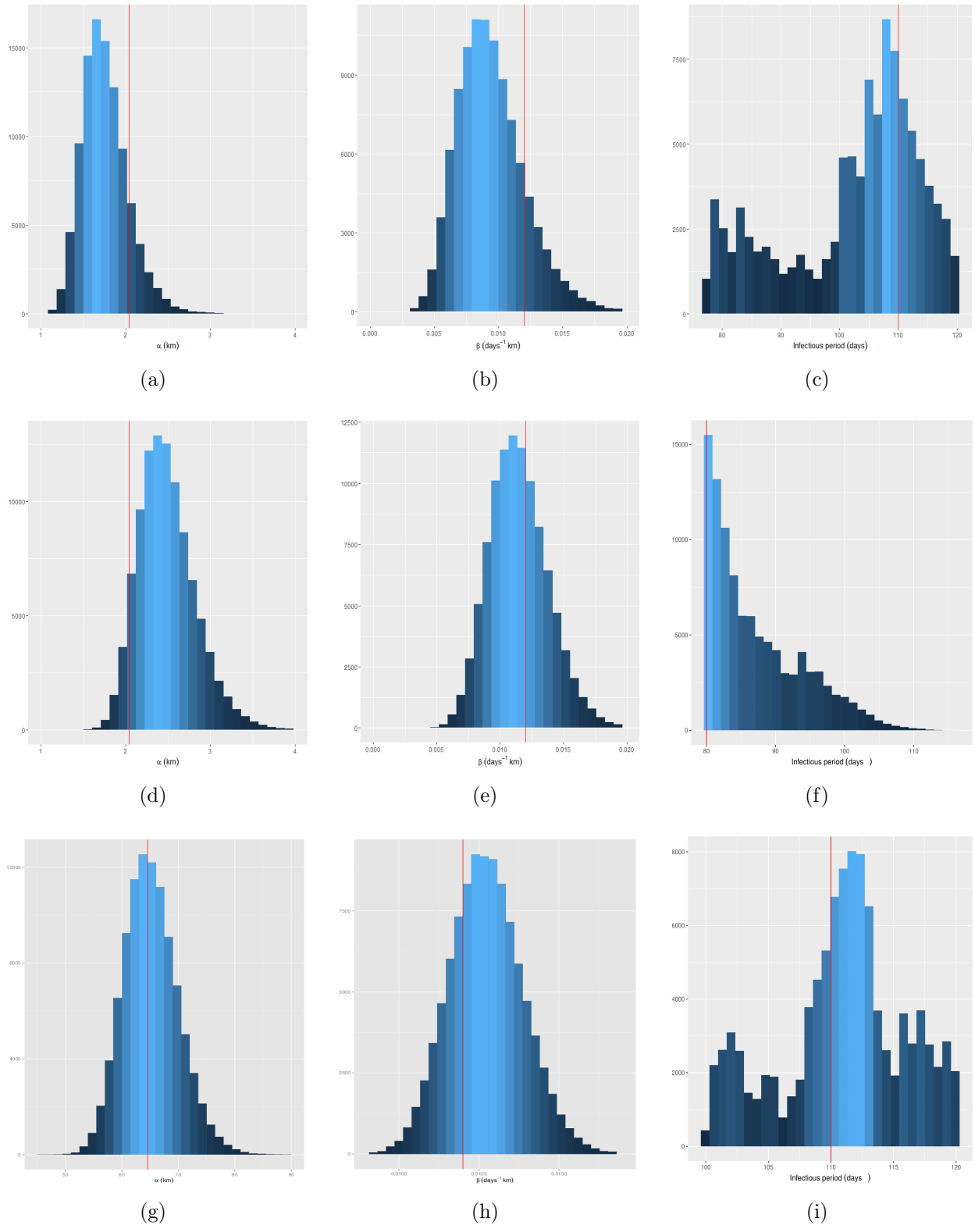
is possible to elicit evidence that the methods are working correctly.

This verification approach is illustrated using simulated data with parameters and total epidemic size shown in the following table. The Rayleigh kernel is used for the simulations.

	α'	β	γ	T_{max}	No. of infections
I	2.04	0.012	110	500	17
II	2.04	0.012	80	500	29
III	64.5	0.02	110	325	61
IV	64.5	0.012	110	500	126

Table 7.2: Parameter values used in simulation study..

In each case, we consider vague gamma priors for the parameters α and β but use an informative uniform prior $U(0, 120)$ in scenarios I and II and $U(100, 120)$ in III and IV for the infectious period duration γ. The Markov chain is run for 100000 iterations and no evidence of lack of convergence is observed. The posterior distributions are presented in Figures 7.5 and 7.6 with the known parameter values indicated by the red line on these diagrams. These results indicate that the MCMC methods appear able to recover the model parameters from data with true parameter values lying in the region supported by the respective marginal posterior densities.

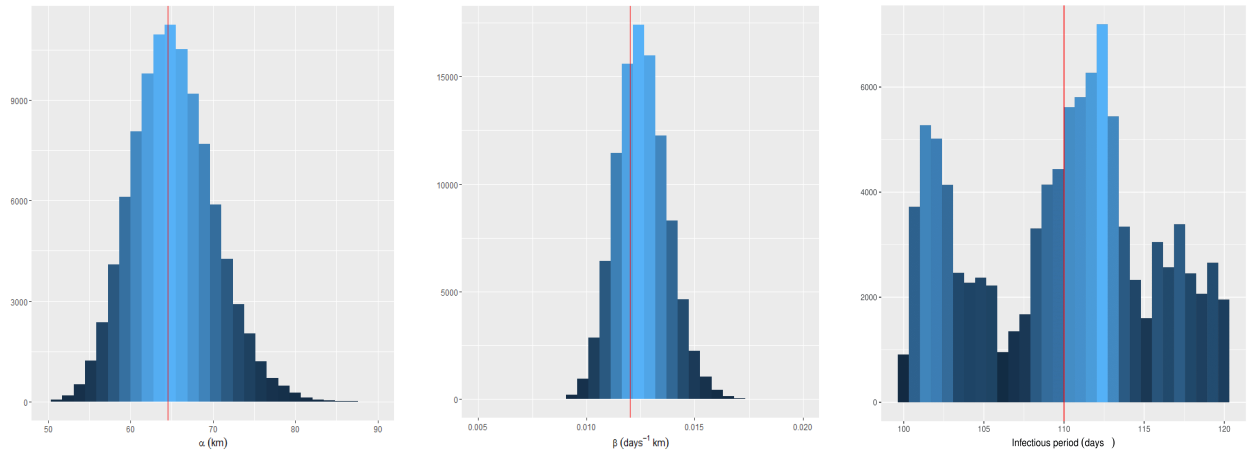


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 7.5: Posterior distribution of the model parameters α, β and γ taking account of the unobserved transmission network for cases I (first row), II (second row), III (third row) as shown in table 7.2. A Rayleigh distribution for $f(r)$ is assumed.



(a) (b) (c)

Figure 7.6: Posterior distribution of the model parameters α, β and γ taking account of the unobserved transmission network for case IV as shown in table 7.2.

7.4.5 Generalisation of the modelling approach

The assumption of constant infectious period many be unrealistic for many real-life epidemics. In reality, individuals respond differently to an epidemic. We therefore relax this assumption (for the case of Rayleigh kernel only) by assuming instead that individuals remain infectious for a time period that is exponentially distributed. In this case, the likelihood is then given by:

$$L(\alpha, \beta, \gamma; \mathbf{s}, R) \propto \exp\left(-\beta \int_0^T n(t)dt\right) \prod_{i=1}^{N} \beta \frac{f(r_i; \alpha)}{r_i}$$

$$\prod_{j=1}^{N_r} \phi\left(r_j - t_j\right) \prod_{s=1}^{N-N_r} \left(1 - \Phi\left(T - t_s\right)\right) \tag{7.4.13}$$

where ϕ and Φ are respectively the density function and the CDF of the exponential distribution, N_r the number of individuals removed, and R the vector of removal times.

Inference in this case is similar to the previous case. In addition to the gamma priors used previously for the parameters α and β, we assume a gamma prior for the parameter γ (considered here as the parameter of the exponential distribution) i.e. $\gamma \sim Gamma(e, f)$. The joint posterior distribution is then given by :

$$\pi(\alpha, \beta, \gamma, \mathbf{s}, R|\mathbf{y}) \propto L(\alpha, \beta, \gamma; \mathbf{s}, R)\beta^{c-1}\alpha^{a-1}\gamma^{e-1} \exp(-d\beta - b\alpha - f\gamma) \tag{7.4.14}$$

The full conditional distributions in Equations 7.4.6, 7.4.15, 7.4.8, 7.4.5, 7.4.10 are

maintained whereas the full conditional distribution of the infectious period parameter is given by:

$$\gamma|\mathbf{y} \sim Gamma\left(N_r + e, f + \sum_{i=1}^{N_r}(R_i - t_i) + \sum_{j=1}^{N-N_r}(T - t_j)\right) \tag{7.4.15}$$

We use Gibbs sampling to sample from the parameters α, β and γ. However updates to the transmission network along with the removal times are done using Metropolis-Hasting algorithm. We use an independent sampler for the proposal distribution of the removal times which uses the likelihood function as $q(R_j - t_j, R'_j - t_j) \sim Exp(\gamma)$. The algorithm is described as follows:

Algorithm to generate sample from the posterior distribution using Metropolis within Gibbs Sampling

i) Initiate the chain with values β^0, α^0, γ^0, s^0 and R^0.

ii) Update β, α, γ using Gibbs Sampling by drawing from their corresponding full conditional distribution.

iii) Update the transmission network s using Metropolis-Hastings scheme:

(a) Randomly choose an infection j in the set of infection with the first infection excluded i.e $2, \ldots, N$.

(b) Update its source by randomly choosing among the potential sources i.e. from the set $\{i : t_i < t_j < R_i\}$.

(c) Accept the new transmission network s^{new} with probability

$$p_{acc} = \min\left\{1, \frac{p(\mathbf{s}^{new}|\alpha, \gamma, \mathbf{y})}{p(\mathbf{s}^{old}|\alpha, \gamma, \mathbf{y})}\right\} \tag{7.4.16}$$

iv) Update the removal times using Metropolis-Hastings steps

Repeat

(a) Select an infected premise j.

(b) If j is not removed, propose to add a removal time R'_j and accept the new vector of removal times with probability

$$p_{acc} = \min\left\{1, \frac{\Phi(T - t_j)}{2\phi\left(R'_j - t_j\right)}\Omega\right\} \tag{7.4.17}$$

where $\Omega = \frac{L(\alpha,\beta,\gamma,\mathbf{s};R^{new})}{L(\alpha,\beta,\gamma;\mathbf{s},R^{old})}$.

(c) If j is removed during the observation period, propose to do one of the following:

- Move its removal time with probability 1/2. To this end, we propose a new removal time R_j' and accept it with probability

$$p_{acc} = \min\left\{1, \frac{L(\alpha, \beta, \gamma; \mathbf{s}, R^{new})}{L(\alpha, \beta, \gamma; \mathbf{s}, R^{old})}\right\} \tag{7.4.18}$$

- or delete its removal time with probability 1/2. We accept the new removal time with probability

$$p_{acc} = \min\left\{1, 2\frac{\phi\left(R_j - t_j\right)}{\Phi(T - t_j)}\Omega\right\} \tag{7.4.19}$$

v) Repeat i)-iv) until convergence.

It is worth noting that from the above description, if host j selected is a source of infection of individual k, we propose to move or add its removal time (depending on its status) R_j' such that $t_k - t_j < R_j' - t_j < T - t_j$.

The method described here is then applied to the Tunisia data. We use the above MCMC routine to sample from the Posterior distribution of the parameters (α, β and γ) using vague priors on α, β as in the previous section. We use however two different priors for the infectious period parameter γ. Specifically, we consider a vague gamma prior by setting $e = 10^{-4}$ and $f = 10^{-6}$; and an informative prior using $e = 27.6$ and $f = 2790$. The MCMC is run for 100000 iterations with the first 10000 discarded.

The trace plots in Figure 7.7 of the model parameters show no evidence of non-convergence though the chain mixes better when using an informative prior for γ. The corresponding posterior densities are shown in Figure 7.8. It can be seen that the choice of prior does not have much impact on the posterior distribution of α and β. However, the mean infectious period or the infectious period parameter γ is sensitive to the choice of its prior which will surely have an impact on the wave speed of propagation.

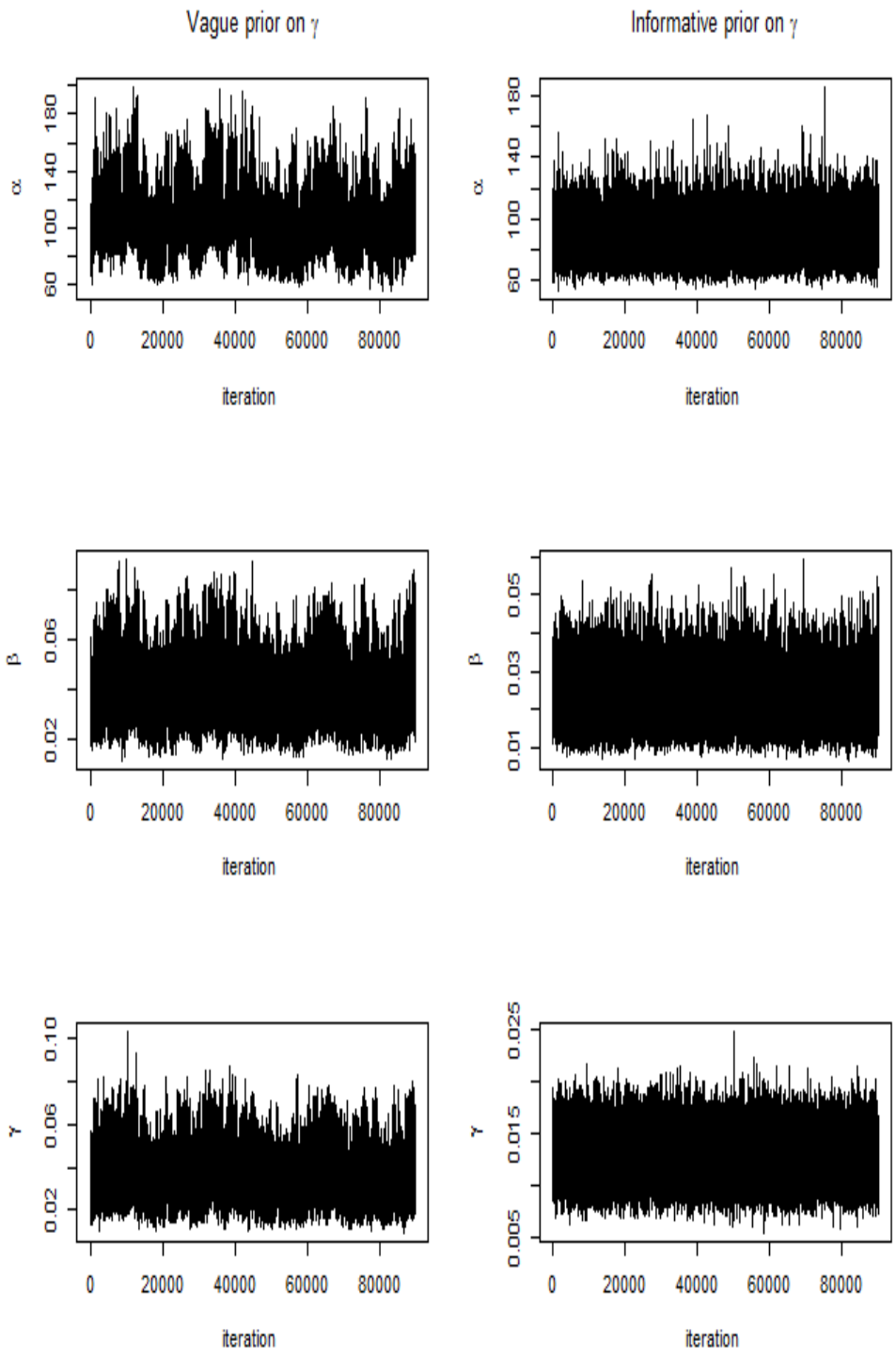


Figure 7.7: Trace plot after a burn-in period of 10000 iterations of α, β and γ using a vague and an informative prior on the infectious parameter γ in the case of the general model considering the Tunisia data.

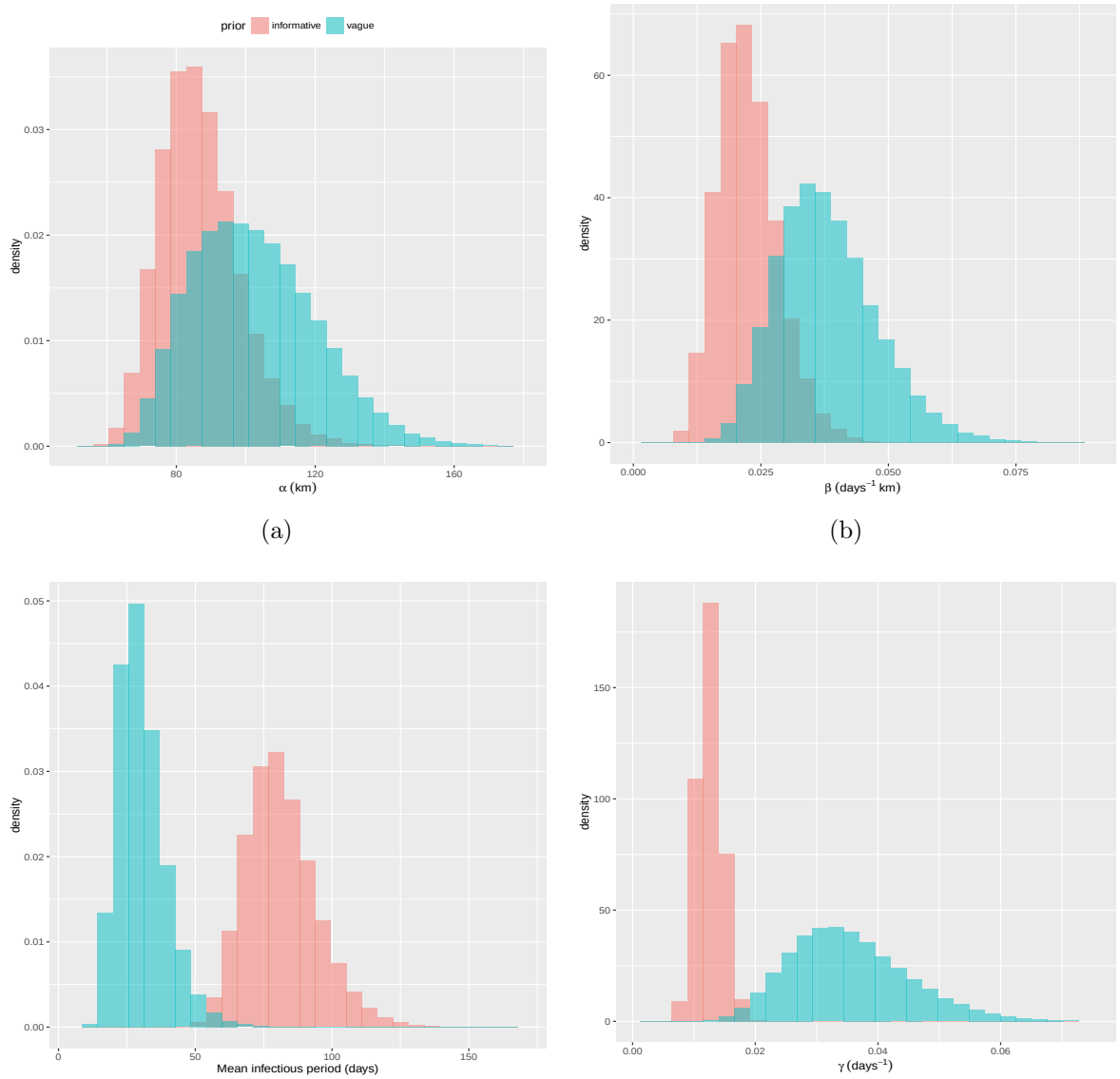


Figure 7.8: Posterior distribution of the model parameters α, β, γ and the mean infectious period taking account of the unobserved transmission network for the exponential infectious period using a vague and an informative prior on γ.

7.5 Generating wave speeds for PPR transmission

In this section the estimation of wave speeds from the analyses of the previous section is described. This is essential to establish the link between the country-specific model and the network model considered later, as the wave speed is related to the time taken for PPR to transit a given country and infect a neighbour. To this end, we simulate the epidemic process using the posterior distribution of the parameters obtained from the inference and measure how far the disease spreads from the initial source after some particular period. The following is the algorithm (essentially a variant of the well-

known Gillespie algorithm (see algorithm 1)) to simulate a realisation of the epidemic, given a set of parameters (α, β, γ), and estimate wave speeds. This algorithm is then repeatedly run using different parameters sets (α, β, γ) drawn from the posterior distribution $\pi(\alpha, \beta, \gamma|\mathbf{y})$ where $\mathbf{y}$ denotes the Tunisian data.

7.5.1 Simulation algorithm

Speed of wave propagation simultion algorithm

1. Initialise $t = t_0$, the initial time, and set location of the initial infection to be $X_0 = (x_0, y_0)$, the number of infectious hosts $n = 1$ and the initial distance measuring how far the disease spread $d_0 = 0$.

2. Add new infectious period, γ (for constant infectious period for example) for the new infection.

3. New infections (or removals) are generated as follows. Suppose that at time t there have been k infections and $n \geq 1$ currently infectious premises and let Δ_R denote the infimum of the remaining infectious periods for active infections at time t. Draw $u \sim U(0, 1)$ and calculate

$$\Delta_t = -\log(u)/(n\beta) \tag{7.5.1}$$

4. If $\Delta_t < \Delta_R$, then select a currently infected premises j uniformly at random from the n infections not removed yet, draw $\theta \sim U[0, 2\pi]$, draw $r \sim f(r; \alpha)$ and place a new infection at location

$$X_{k+1} = X_j + (r\cos(\theta), r\sin(\theta)) \tag{7.5.2}$$

   Assign this new infection to have infectious period γ and reduce remaining infectious periods of all other infections by Δt, setting $t = t + \Delta t$. Calculate $d_{k+1} = d(X_0, X_{k+1})$ where $d()$ denotes Euclidean distance.

5. If $\Delta t > \Delta_R$ then remove the infection with the minimal remaining infection time and reduce all other infectious periods by Δ_R, and set $t = t + \Delta_R$, then return to step 3.

6. Repeat $2 - 5$ until a stopping criterion is reached (e.g. $t \geq T$).

For a given realisation of the process, a set of estimates of wave speed $v_1, \ldots, v_m$ can be generated at times $T_1, \ldots, T_m$, where $v_i = D_i/T_i$ and D_i is the maximum distance the wave has travelled by time T_i and is given by

$$D_i = \max_k \{d(X_0, X_k) | t_k \leq T_i\}, \tag{7.5.3}$$

that is the largest distance between X_0 and any premises whose infection time is less than or equal to T_i.

To account for the uncertainty on the model parameters, we sample from the joint posterior distribution of the model parameters. Therefore each draw of parameters from the posterior distribution provides m estimates of the speed, which we use to build up a population of estimates of speed.

7.5.2 Estimating wave speeds for Tunisian data

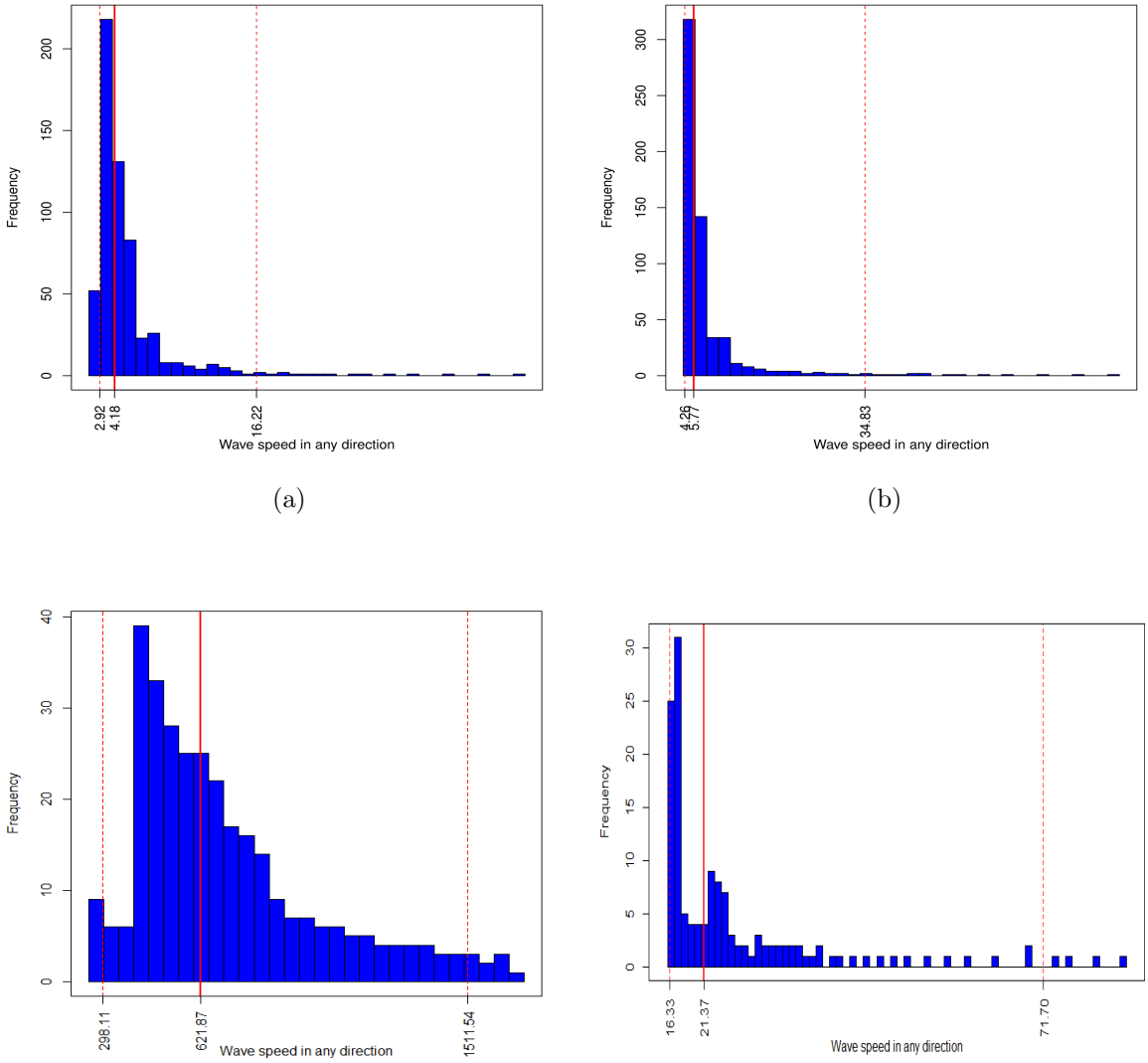
The methodology for estimating wave speeds is illustrated on the data from Tunisia $\mathbf{y}$ assuming a fixed infectious period. Three different kernel functions were used (Rayleigh, Exponential, Cauchy) and distributions of wave speeds were estimated.

Algorithm

1. A total of 1000 samples from the posterior distribution $\pi(\alpha, \beta, \gamma|\mathbf{y})$ were generated using the MCMC algorithm of section 7.4.2.

2. For each sample the process is simulated until time $T = 325$ and speeds $v_1, \ldots, v_{325}$ are estimated at times $i = 1, 2, \ldots, 325$ respectively.

Figure 7.9 shows the predictive distribution of the wave speed in the absence of controls using the 3 different kernels along with the posterior densities of the model parameters and the exponential infectious period described in the previous Section. As should be expected, the model using the Cauchy kernel leads to a posterior distribution of wave speeds that supports very high speeds compared to the Rayleigh and Exponential.



(a) (b)

(c) (d)

Figure 7.9: Posterior wave speed in absence of control meausred in $km\,days^{-1}$ using different kernels (a) Rayleigh, (b) exponential and (c)Cauchy kernels and (d) Rayleigh kernels with exponential infectious period . Dots lines show the 95% confidence interval and the solid line represents the median

7.5.3 Comparison with baseline paper

In the EFSA Opinion, an attempt is made to estimate wave speeds from the Tunisian data. This is achieved by considering all possible infecting sources for each infection and for each potential source, estimating a wave speed based on the distance between the respective source and the given infection and the difference in infection times. This led to a population of estimates from which a 95% confidence interval was estimated to be $(0.3, 65.5)$, where speed is measured in km/day, with a median speed of 3.9. Using

the above algorithm to generate a population of estimates, we generate a 95% equal tailed posterior credible interval for the speeds, and estimate the median speed. These are illustrated for the three models in Figure 7.9. It should be noted that credible intervals for the Rayleigh and exponential kernel respectively (with fixed infectious period) are narrower compared to $(0.3, 65.5)$ obtained in the baseline paper. The median speed for the model with the Rayleigh kernel (4.18) is comparable to the median obtained in the EFSA analysis (EFSA, 2015). The wider credible interval in the EFSA opinion may arise due the fact that the approach therein does not use any statistical model that distinguishes between more likely and less likely infection links, while the stochastic spatio-temporal model does.

The Cauchy model, which predicts very high speed, is unlikely to be realistic unless there is strong reason to believe that long-range processes are actually operating in the context of PPR spread, other than the animal movement which is represented in the next section using a separate process. Indeed, the theory of spatio-temporal processes indicates that the analogy of a spreading wave is not appropriate for a kernel like the Cauchy, which is not exponentially bounded, so that the concept of transmission speed is not meaningful.

7.5.4 Model selection

The common tool used in the literature for model comparison is the Deviance Information Criterion (DIC) introduced by Spiegelhalter et al. (2002). The best model is the one with the smallest DIC. In the presence of missing data as often the case in real-world data especially in epidemiology where data are partially observed, the DIC developed by Spiegelhalter et al. (2002) presents some limitations since the DIC cannot be uniquely defined as pointed out by Celeux et al. (2006). Here we use the one named as DIC_4 by Celeux et al. (2006) defined as:

$$DIC = -4\mathbb{E}_{\boldsymbol{\theta},\mathbf{s}}\left[\log\left(\pi(\mathbf{y},\mathbf{s}|\theta)\right)|\mathbf{y}\right] + 2\mathbb{E}_{\mathbf{s}}\left[\log\left(\pi(\mathbf{y},\mathbf{s}|\hat{\boldsymbol{\theta}}(\mathbf{y},\mathbf{s})\right)|\mathbf{y}\right] \tag{7.5.4}$$

where $\hat{\theta}(\mathbf{y}, \boldsymbol{s})$ is a posterior point estimate such as the mean (used here), median etc.. based on the likelihood of both the observed and unobserved model $\pi(\mathbf{y}, \mathbf{s}|\boldsymbol{\theta})$ and $\mathbf{y}$ and $\mathbf{s}$ are the observed data and missing transmission network respectively. The computation of the DIC in Equation 7.5.4 required the computation of the expectation of some quantities which can be easily derived using the MCMC techniques.

We then compare different models used for Tunisia data for fixed infectious period. The results are shown in Table 7.3. It can be seen that the DIC favours the model with the exponential and Cauchy kernel over the Rayleigh one, which seems to contradict our suggestion that the Rayleigh kernel represents adequately the PPR dispersal. This is due to the insufficiency of the data to enable any kind of rigorous model

Model	Exponential	Rayleigh	Cauchy
DIC	658.9445	937.3481	673.1241

Table 7.3: Model comparison with DIC.

choice and the biological understanding of the spread of PPR may be sufficient to inform the choice of model. Nevertheless, the techniques described in the chapter are not restricted to any particular choice of kernel and with minor modification could be applied to fit a model with any desired choice of kernel function.

7.6 Network models for spread at the global scale

In selecting an approach to larger-scale modelling of PPR in this project the practical goals of the work to inform policy regarding risk of incursion of PPR have been borne in mind. In particular, it is important that the approach to modelling a network is capable of:

- quantifying risk of spread from a given source country to any other country within a given timescale;
- taking account of information on spread of PPR from a range of sources;
- propagating uncertainty regarding key system parameters into predicted risk;
- adapting to new information on PPR of various forms as it becomes available;
- representing impact of potential controls on the likelihood of invasion;
- lending itself to some kind of extrapolation for example to estimate parameters for parts of the network for which data are unavailable;
- assessing relative risks of different courses of incursion into any given country;
- integrating the results of other studies at smaller scales.

With these factors in mind, an approach is followed whereby the temporal dynamics of PPR are represented stochastically and in which different mechanisms of spread are combined in an integrated network model. In particular, the system is represented as a network of vertices and directed edges, where each edge corresponds to a mechanism of spread and behaves stochastically in a manner which can reflect current knowledge regarding the respective mechanism.

7.6.1 PPR Network Model: Structure

Specifically, we consider a network with vertex set

$$V = v_1, \ldots, v_n \tag{7.6.1}$$

where each vertex represents a country. The network is further specified by the set of directed edges

$$E = e_1, e_2, \ldots, e_m \tag{7.6.2}$$

where each edge represents a mode of transmission of PPR between two particular countries. Any edge therefore requires 3 pieces of information in order to specify it:

- the donor country v_d,
- the recipient country v_r,
- the mode of transmission, m, where m could be any one of: illegal movement of infected animals, legal movement of infected animals, or spread across national boundaries due to other mechanisms.

The directed nature of edges is important as there is no reason to assume any symmetry, for example, in the patterns of illegal movement from country A to B and from B to A. Clearly, certain aspects of the network will be constrained by geography, with two countries requiring a common border in order for an edge corresponding to transboundary spread to be present between them. Figure 7.10 shows a representation of possible transmission routes from one country to another.

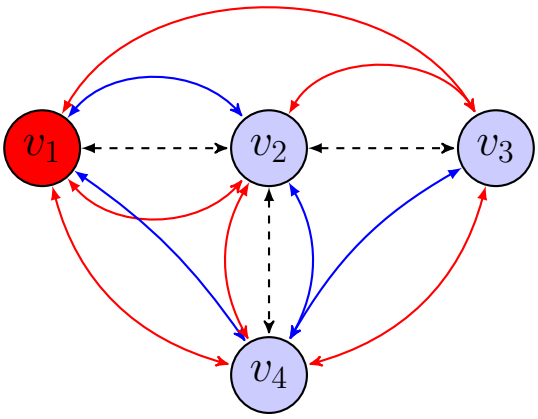


Figure 7.10: Graph showing different routes PPR could be transmitted from one country to another. Dashed line shows that countries share the same border so transmission could occur accross the border. Red and blue conexion show the existence of illegal and legal animal movement between countries. Red node denotes the country that is source of infection.

7.6.2 Stochastic dynamics using first-passage times

To generate a realisation of the stochastic dynamics of the PPR spread through such a network we adopt the following approach. To each edge e_j we assign a random time T_j from a probability distribution ϕ_j. T_j has the interpretation of being the random time (following initial incursion of PPR into $v_d(j)$) that it would take PPR to be transmitted from $v_d(j)$ to $v_r(j)$ via the mode $m(j)$. We assume initially that the times T_j are assigned independently across edges (though not identically distributed - see later).

Once times are allocated to edges, the dynamics of the disease subsequent to initial incursion of PPR into a given country A, say, at time $t = 0$ can be determined. The method for doing this utilises some standard techniques frequently used in operational research and studies of communication networks which exploit the connection between the first-passage times and minimal path lengths through a directed graph. In essence, by identifying the shortest path of directed edges (where path-length corresponds to the sum of the T_j over edges e_j in the path) from A to any given country, generates a realisation from the distribution of time taken for the disease to pass to B from A. We denote such time the first-passage time of the pathogen to to B from A. This calculation can be effected using classical optimisation algorithms such as the Bellman-Ford algorithm (Thippeswamy et al., 2010), or the more specialised Dijkstra's algorithm (Thippeswamy et al., 2010). In this project we use the former algorithm. It is surprising that this connection between shortest-paths in directed graphs and epidemic dynamics has not been more fully exploited in the applied epidemics literature. By repeatedly generating realisations of the edge process and computing first-passage times, the distribution of times for PPR to reach country B from any given incursion to country A can be investigated and risk measures (such as probability of spread within a given time horizon can be computed).

7.6.3 Algorithm for generating first-passage times

In the algorithm, having generated an edge (with associated transmission time) for every potential transmission mechanism between any pair of countries, we first carry out some pre-processing which reduces the directed graph to one where at most a single edge exists between any two countries i and j. For each pair of countries i, j with a potential mechanism of transmission between them, a time T_{ij} is defined to be the infimum of the transition times of all edges corresponding to transmission between i and j. As well as storing this minimum time, the mechanism (transboundary spread, legal movement, illegal movement) is recorded. It is clear that first-passage times in the original network, in which multiple edges from i to j may be represented, are precisely the same as those calculated in the reduced network with at most one edge

from i to j. The Bellman-Ford algorithm operates as follows.

Bellman-Ford algorithm for first passage time

1. Consider a network of n nodes where each node represents a country. Between any two countries i and j where a mechanism of spread from i to j is known to exist, represent this with a single edge of length T_{ij}, this being the minimum transition time over all mechanisms of transmission for i to j. Let M_{ij} denote the mechanism with this minimum time.

2. Next we consider a source country, k, this being the particular country which is assumed to be the initially infected country. We initialise the algorithm by setting $time(k) = 0$ and then for $i \neq k$ we set $time(i) = \infty$. On completion of the algorithm, $time()$ will denote the time at which the disease first reaches i from the source k.

3. For $i = 1, \ldots, n-1$ we repeat the following process. For every edge-length T_{ij},

$$
\begin{aligned}
&\textit{If } time(j) > time(i) + T_{ij}\{ \\
&\quad time(j) = time(i) + T_{ij} \\
&\quad predecessor(j) = i \\
&\quad mode(j) = M_{ij} \\
&\quad \}
\end{aligned}
$$

Note that the scenario of having multiple initial sources $k_1, \ldots, k_r$ can be captured by including an edge of length 0 between every pair of vertices in this set.

7.6.4 Advantages of 1^{st}-passage-time representation

There are several benefits from using this representation of the process.

7.6.4.1 Conditional distribution of infection source

A useful by-product of the approach is that it naturally allows for investigation of the distribution of the source of infection, conditional on a given country being infected.

This can potentially indicate which source is the most hazardous so far as initial incursion into a given country is concerned, and may help identify what the most appropriate control measures may be so far as preventing spread to that country.

7.6.4.2 Integration and assimilation of knowledge

By allocating each mechanism its own edge in the network, we are able to combine information from separate sources in a coherent way. For example, information on Tunisia regarding spatial spread can inform the model used for the transition time over an edge representing trans-boundary spread from Tunisia to a neighbouring country. Information on animal movement rate between one country and another, coupled with knowledge of the sero-prevalence within the donor country can inform the choice of model for the random time for spread due to legal animal movement between countries. Moreover, where there is little knowledge on a mechanism, that uncertainty can nevertheless be represented and propagated into prediction we give further information on the modelling of transition times below. As further information becomes available on a particular mechanism, then this can be used to update assumptions regarding the parameters governing that mechanism.

In addition, this representation is particularly useful in assessing the impact of surveillance and control measures discussed later.

7.6.5 Passage-time models and parameterisation

Many approaches could be taken to assigning distributions to passage-times for individual edges, e_j. Here some straightforward approaches are considered.

- For passage time due to spread across boundaries between neighbouring countries we consider a $Gamma(a_j, b_j)$ where a_j and b_j are estimated from distributions of wave-speed in relation with some measure of Euclidean distance between the countries involved. Specifically the population of wave speeds obtained for the Rayleigh kernel are used to generate a population of passage times for edge j. This population of passage times are then fitted to a Gamma distribution by the method of moments in order to obtain a_j and b_j.

- For passage times related to animal movement the occurrence of infection events is represented as a Poisson process. This fits well with the approach taken in the EFSA opinion (EFSA, 2015) where a $Binomial(n, p)$ distribution was used to represent the numbers of infected animal transferred in a year via the movement with n denoting numbers moved and p denoting prevalence in the source country. The natural continuous-time representation of these assumptions uses a Poisson process with rate np. Thus the time to traverse edge e_j would be modelled as

an $Exp(\lambda_j)$ where λ_j can be related to the numbers of animal movements and probability of each animal being infected in a straightforward way.

Note that uncertainty in these processes could be captured by assigning distributions to the parameters a_j, b_j, or λ_j and propagating this into the distribution of the associated T_j.

7.6.6 Illustrative results with 1st passage-time model

Since extensive data regarding illegal animal movement along with information on transboundary spread are not available, we propose a simulation using arbitrary parameters to illustrate how the Bellman-Ford algorithm can be used. To ensure estimated parameters are as realistic as possible, we first consider the data regarding legal animal movement shown in (EFSA, 2015, P 36) where the average amount of animal movement between some countries during the period of $2009 - 2013$ is reported. To estimate the distance between contiguous countries necessary for identifying the distribution of time taken for transboundary spread we consider the 149 countries (South, 2011)) supplemented by the inclusion of Libya, and we use the corresponding geographical information (locations of the capitals of these countries) provided by Mayer and Zignago (2005) in order to identify a notional distance between contiguous countries. In addition, we assume that the posterior distribution of the wave speed obtained from the Tunisian data can be applied to other countries and use the above approach to parameterise the Gamma distributions used to model passage time via spatial spread between contiguous countries.

The information provided in (EFSA, 2015, P 36) covers legal animal movement between some countries. Where these data are not available we assume that this implies only illegal movement. Initially, it is assumed as in EFSA (2015) that the virus is already present in north Africa and Middle east countries except Libya, and central Asia and south Asia countries (EFSA, 2015). We assume that the value of the prevalence p is identical for all countries (once infected with PPR) along with the rate of illegal animal movement n giving a net rate np in the corresponding Poisson process between any two countries. For the simulations, we assign different values for the prevalence $p = 0.00037, 0.003$ and the number of animals moved illegally $n = 100, 1000/year$. The algorithm is then used to generate 10^5 independent realisations (assignments of passage times to edges) and we present results relating to spread into Europe and into Libya (since the data we use from EFSA (2015) do not consider Libya as infected in 2013 though the virus is now detected in the country). A plausible route of transmission into Europe when coming from Africa is countries such as Greece or Bulgaria, given their proximity to Turkey. Another possibility may be to enter directly to Italy (via illegal movement), given the proximity of Tunisia to the south of Italy.

Therefore for illustration we consider target countries Bulgaria, Greece, Italy and Libya.

In Figures 7.11, 7.12, 7.13 , 7.14, and 7.16, some exemplar outcomes are presented for these countries. For

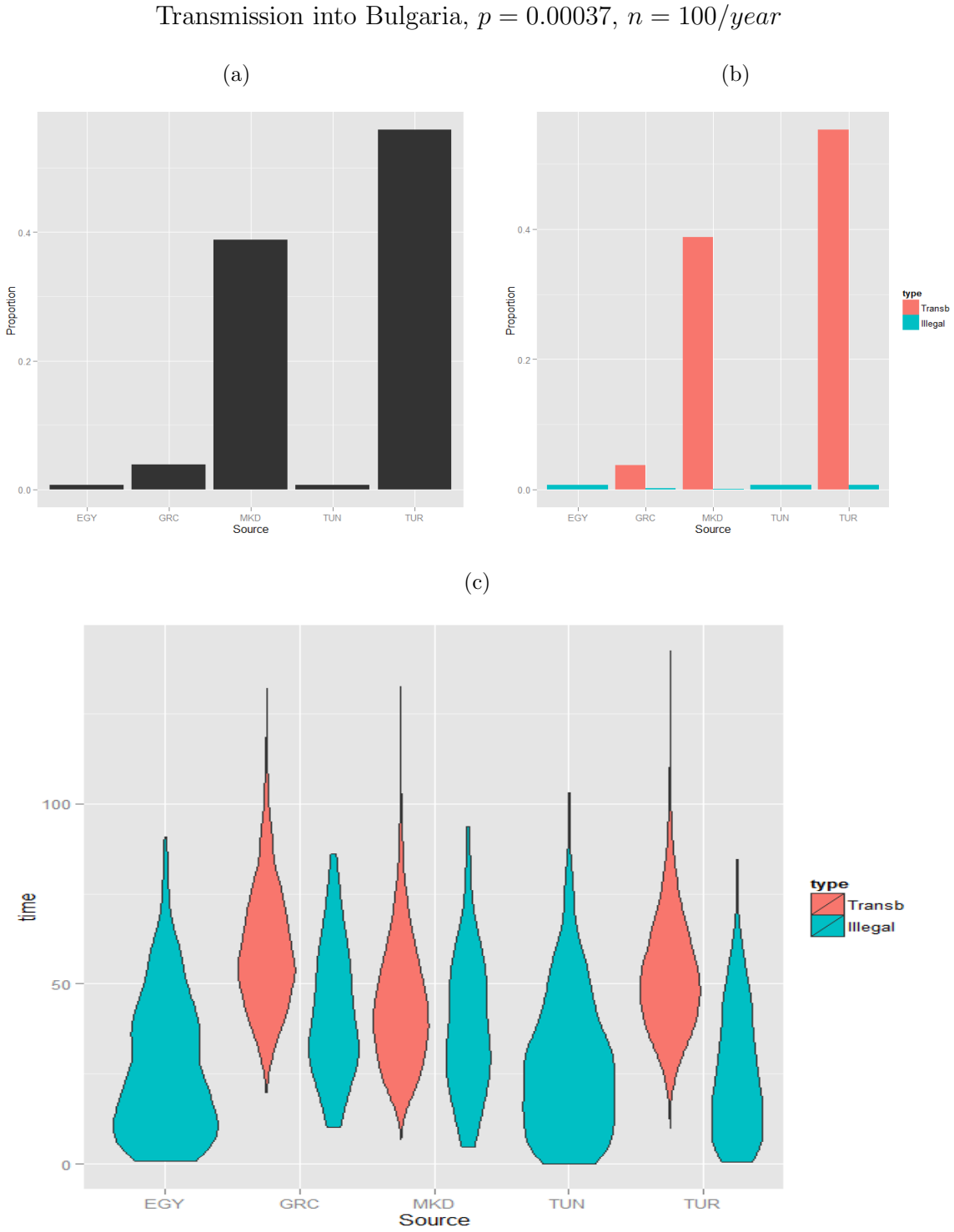


Figure 7.11: Bar plots showing the relative importance of different countries (a) for spread of PPR into Bulgaria and the corresponding type of transmission responsible of the infection (b). The posterior distribution of the first passage time (conditioned on the source of infection and the type of transmission)(c)

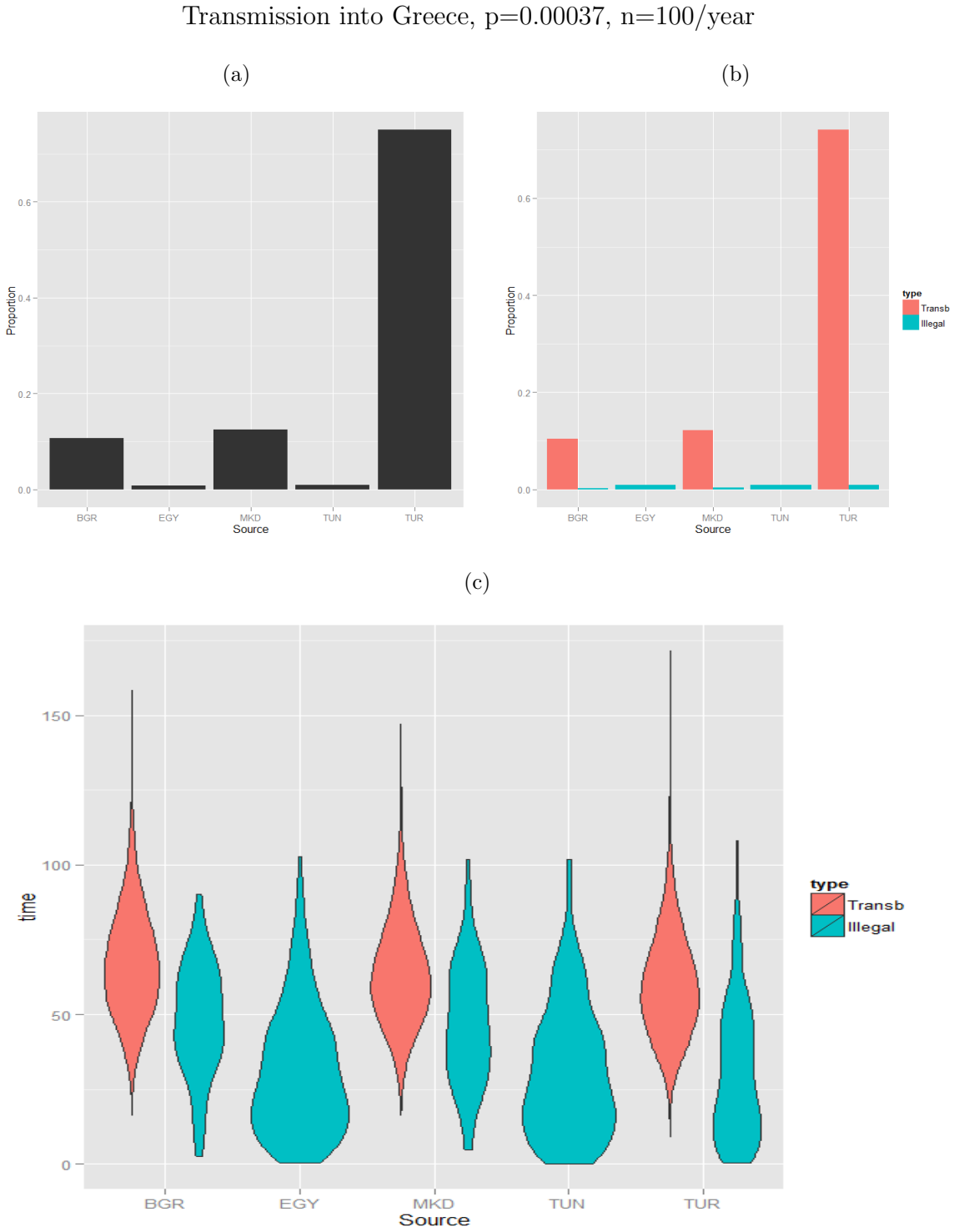


Figure 7.12: Bar plots showing the relative importance of different countries (a) for spread of PPR into Greece and the corresponding type of transmission responsible of the infection (b). The posterior distribution of the first passage time (conditioned on the source of infection and the type of transmission)(c)

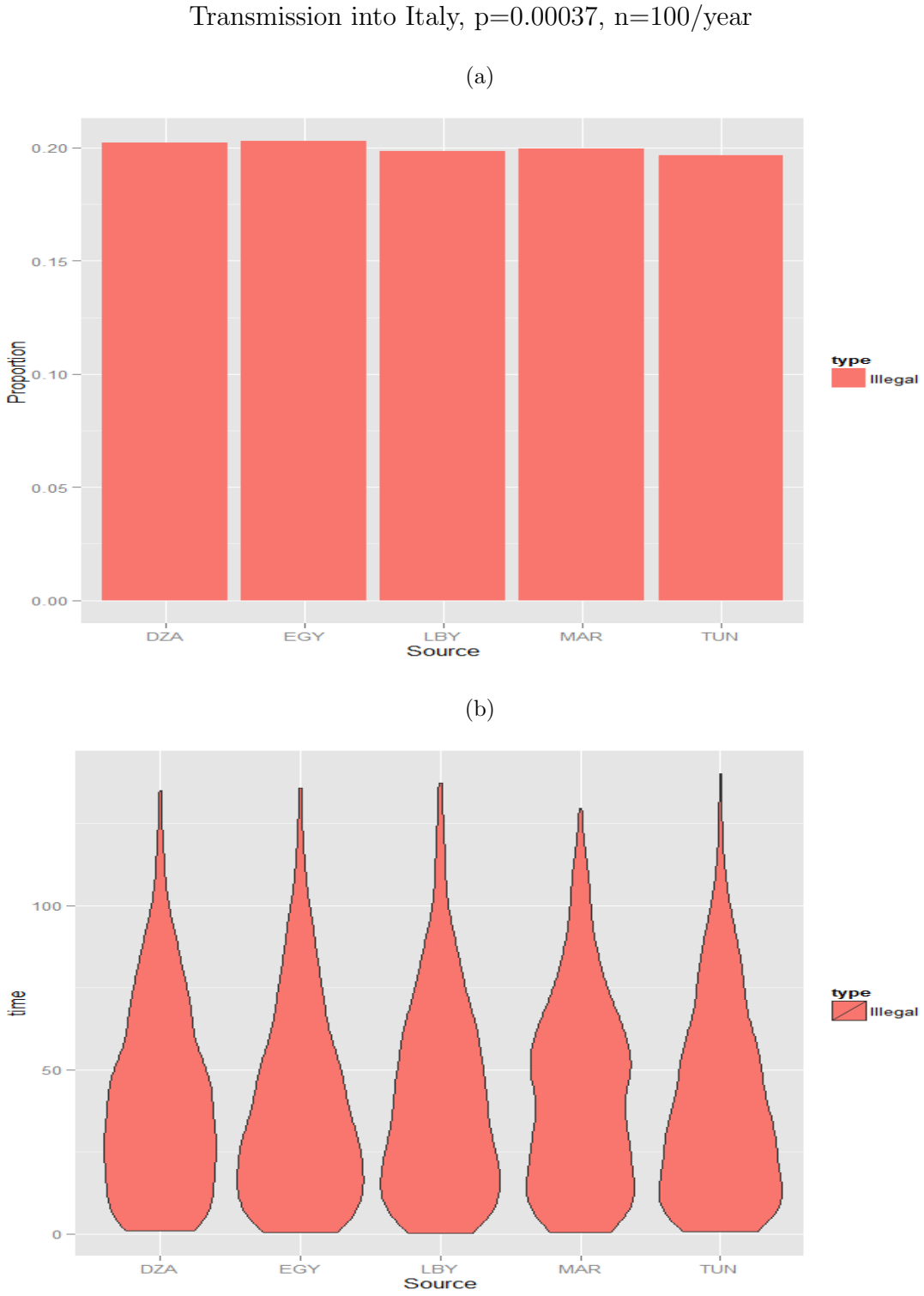


Figure 7.13: Bar plots showing the relative importance of different countries (a) for spread of PPR into Italy. The posterior distribution of the first passage time (conditioned on the source of infection and the type of transmission) (b)

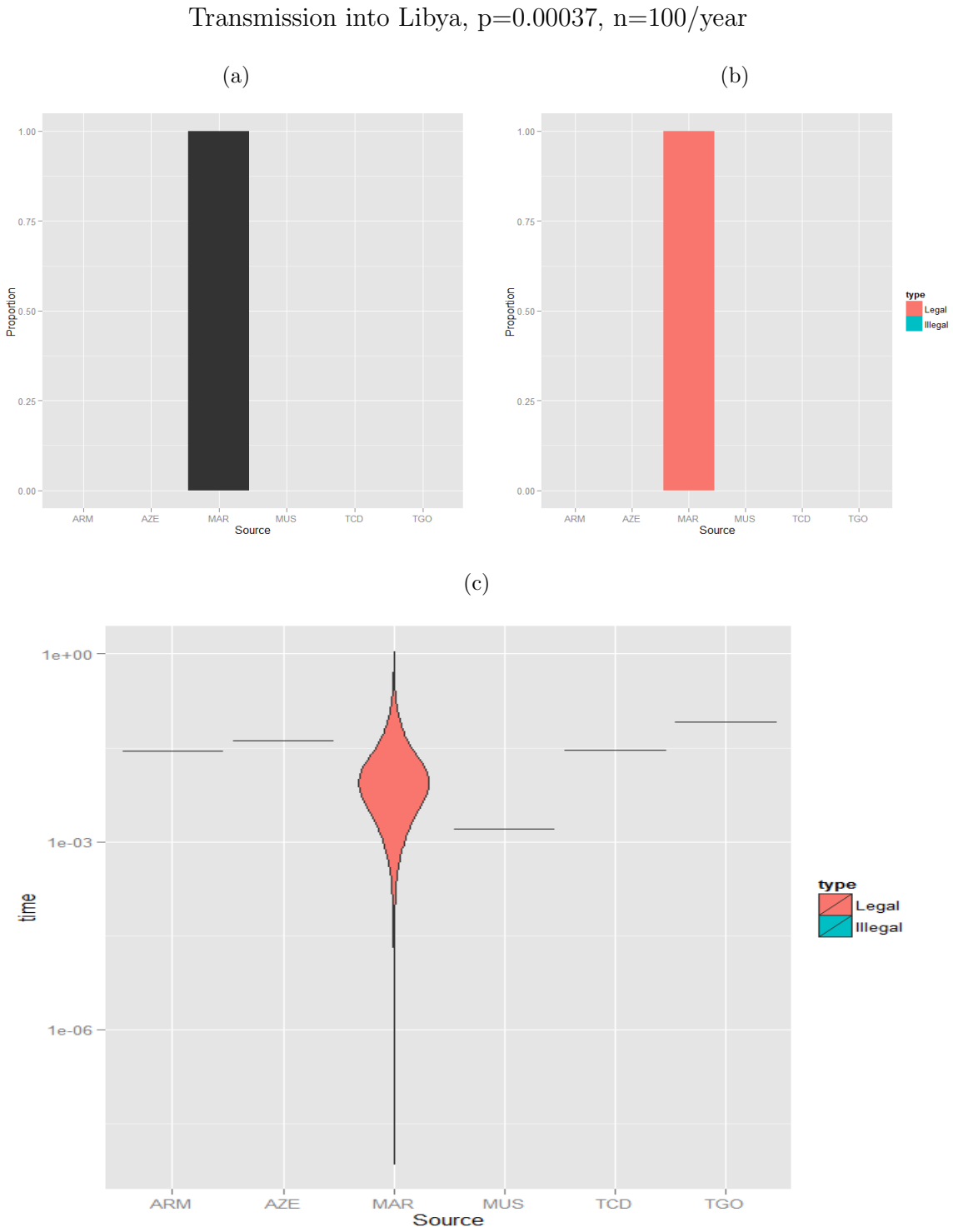
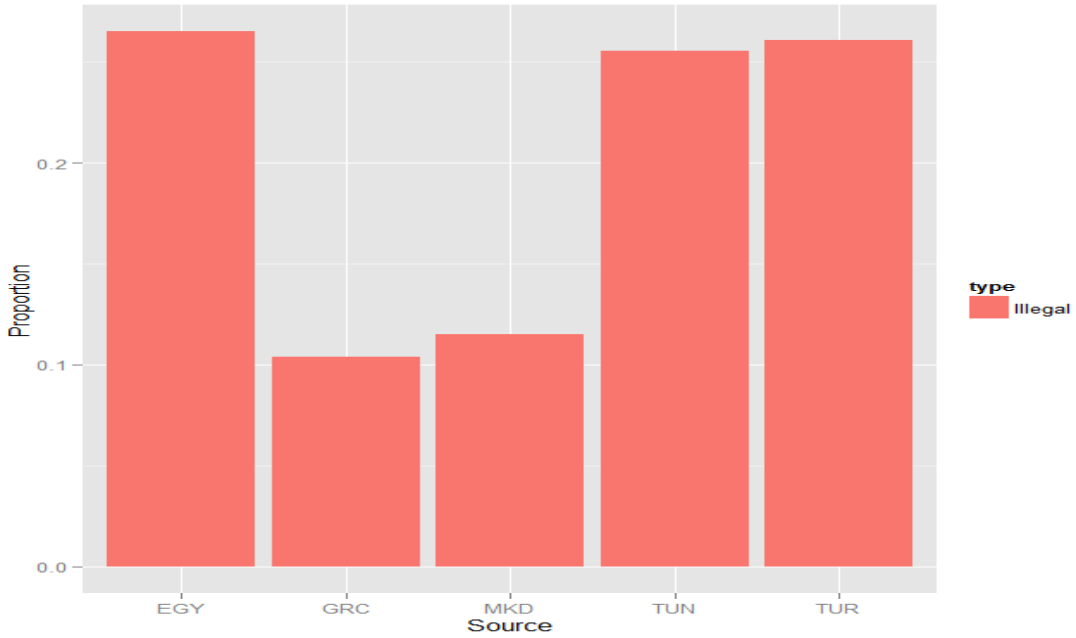


Figure 7.14: Bar plots showing the relative importance of different countries (a) for spread of PPR into Libya and the corresponding type of transmission responsible of the infection (b). The posterior distribution of the first passage time (conditioned on the source of infection and the type of transmission)(c)

Figure 7.15: Transmission into Bulgaria, p=0.003, n=1000/year

(a)



(b)

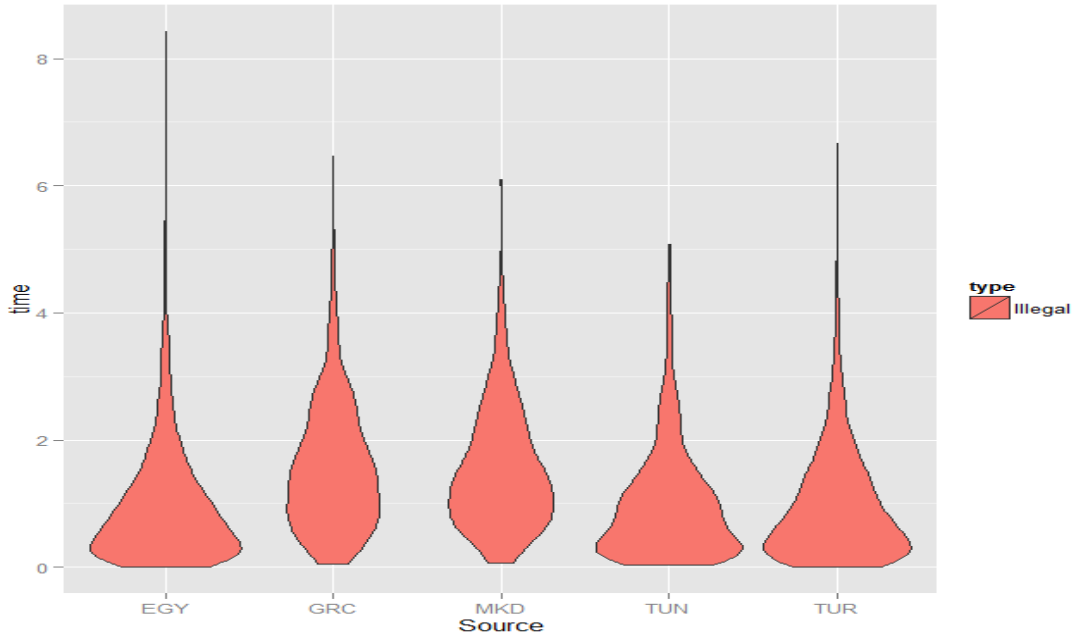


Figure 7.16: Bar plots showing the relative importance of different countries (a) for spread of PPR into Italy. The posterior distribution of the first passage time (conditioned on the source of infection and the type of transmission) (b).

each country (Bulgaria, Greece, Italy and Libya). We show:

- the distribution of the source country of infection (top left histogram);

- the distribution of the source country and infecting mechanism (top right);
- the conditional distribution of the 1st-passage time given the source of the infection (lower histogram).

Note that countries are designated by their ISO3 code Wikipedia (2016)

7.6.7 Discussion of 1st-passage time model

Given the uncertainty regarding parameter values, and the lack of data on specific interactions between countries, these results of Section 7.6.6 should be interpreted as exemplifying the kinds of analyses and predictions that could be achieved using our modelling framework, rather than representing definitive predictions. The latter can only be achieved if more extensive data can be obtained and model testing and comparison can be carried out. Nevertheless, as with all the methods presented in this case study, the model is generally applicable and we believe it can have considerable use as a decision support tool if it can be properly informed by data.

The kinds of conclusions that could be drawn from the models are illustrated by consideration of Figure 7.12, transmission into Greece. It is seen that the most likely source of transmission is transboundary spread from Turkey, with transboundary spread via Macedonia or Bulgaria also having a significant probability under the model. Although the probability of animal movement being the source is small, conditional on movement from currently infected countries (Turkey, Egypt, Tunisia) being the source then the distribution of the time to reach Greece is small. Such analyses can be useful in identifying the most important mechanisms to focus on for the construction of control strategies, which have not been thus far considered in the work.

7.7 Assessment of impact of control with the network model

The representation selected naturally lends itself to assessment of impact of surveillance and control measures. For example, the impact of eradicating the disease in a country within a given time from initial incursion can be seen by removing edges (Haydon et al., 2003) from that country of length exceeding the eradication time. Specifically, one could investigate the impact of controls by simulating times for every directed edge and then looking to see the impact of controls which, eradicated the disease within a country within a given time horizon T from first infection. The impact of the control could be investigated by removing all edges out from that country with time more than T. The benefit of this approach in terms of first-passage times

to nodes is that, one would obtain a monotonic relationship between the value of T and the first-passage time. Thus, for a simulated population set of edge lengths on which we apply different control strategies, we should expect highly correlated outcomes therefore, a reduction in variance in the difference of control strategies. It is worth noting that the smaller T, the more edges removed and therefore the longer the first-passage time.

Similarly, the impact of controls to reduce animal movement can be assessed. The approach described here can lead to a coupling of realisations under alternative control strategies that can reduce the amount of simulation required to compare competing control strategies in terms of their expected outcomes.

7.8 Conclusion

In this chapter, we have presented a novel methodology for modelling PPR spread within a given country and the risk of subsequent introduction to other countries. To that end, we initially develop a contact model that only accounts for the infected premisses as the information of susceptible population is unavailable. Inference on model parameters is drawn from their posterior distributions generated *via* MCMC algorithm. The most practical implication of this modelling is the estimation of the distribution of wave speed. We then develop an algorithm that can enable the estimation of the posterior distribution of wave speed of propagation. Results from the application on the Tunisia data assuming a constant infectious period show consistency of our approach using the Rayleigh kernel with results in EFSA (2015).

We represent the spread of PPR in a large-scale network of countries in which different modes of transmissions are considered. We borrow techniques from graph theory and operational research such as Bellman-Ford algorithm to introduce the notion of first-passage time from a source country to a given country. The exemplar outcomes presented in this framework show that this representation can be used as a tool to guide decision makers as more information on animal movement becomes available.

Chapter 8

Conclusion and future work

8.1 Summary and conclusions

In the major part of the thesis, we designed experiments to compare the effect of alternatives strategies on real time epidemics. To that end, we used a mechanism that enables us to couple epidemics with different control strategies by matching their latent processes which we assumed are unaffected by controls. We used the Sellke thresholds as the latent processes to induce dependence between the distributions of the effect of alternative controls. The method was applied in two ways: First, we consider simple and accessible temporal epidemic models. We made simplifying assumptions, assuming no available data, homogeneous mixing (O'Neill and Becker, 2001), fixed and known model parameters. We then moved to more general cases where a spatio-temporal model was considered with the assumption that data are available but are partially observed. In this case, the thresholds are then readily recovered as part of the parameter estimation using Bayesian computation methods namely MCMC techniques.

The latter part focussed on PPR case study. A range of flexible modelling tools have been deployed to inform understanding of risk of introduction of PPR and its spread through a network of countries. Little in the way of earlier modelling work on PPR was available to inform the construction of the models, but the approaches taken have been chosen to take account of analyses carried out by EFSA (EFSA, 2015). Given the limited data on PPR available, a main aim has been to develop modelling frameworks that are readily adaptable to the inclusion of new data and improve understanding of PPR as these become available.

8.1.1 Coupling epidemics using Sellke construction

One of the most significant findings to emerge from this study is that, coupled epidemics with different control strategies using Sellke thresholds reduces the variability

between their outcomes, as they are highly positively correlated. This evidence suggests a more accurate estimation of the expected difference in the effect of control strategies. As a consequence, the sample size of simulated realisations needed to compare competing control strategies is reduced. The current findings add to a growing body of literature on experimental designs for epidemic controls, and add substantially to our understanding of causal effect between different interventions.

Based on the idea of coupling epidemics, various controls strategies were constructed and compared using both non-spatial and spatio-temporal models. This provides new tools for future studies in identifying optimal strategies for epidemics control. The approaches developed in this thesis are flexible and could be applied on a wide range of epidemics.

In Chapter 2, we reviewed different types of stochastic models often used to represent epidemics and recall two main numerical approaches namely Gillespie algorithm (Gillespie, 1976) and Sellke construction (Sellke, 1983) often used in the literature, though the former is the most popular. We showed that conditioned on individual Sellke thresholds, the outcome of the epidemic is a deterministic function of the model parameters.

In chapter 4, we modeled epidemic controls assuming that there are no data observed and considering two non-spatial models using the idea of Sellke thresholds introduced in Chapter 2. We first considered an SIR model and we introduced the notion of optimal control strategy using a stochastic optimisation algorithm namely Simulated Annealing (Busetti, 2003; Demon et al., 2011) when the impact function is considered to be the number of infections by a future time. A key feature to that is to increment the number of realisations rather than fixing it from the start. It emerges that SA converges to the optimum distribution of the resources that gives the highest reduction in the epidemic size. Also, results showed highly positive correlation between the outcomes of epidemics when controls are deployed following the optimal strategy and suboptimal strategies. This indicates that the variability of the outcomes of coupled epidemics in these conditions is reduced, therefore providing a better estimate of their expected difference. Second, we designed various control strategies on a simple non-spatial SIS (not typical SIS as the transitions from I to S occur through interventions) model assuming no observations available and that the hosts are not infectious once they contract the disease. The controls include controls based on prioritising hosts with respect to their expected infection time, the risk of infection, random selection and randomly assigned an order. Results showed that if the controls consist of replacement of hosts once detected through diagnostics tests, the prioritisation of hosts based on their expected infection time is more cost-effective in terms of minimising the cumulative time hosts remain infected for.

Chapter 5 generalised the concept of Chapter 4 to spatio-temporal epidemic mod-

els assuming a presence of data. The data considered consisted of a sequence of spatial snapshots of locations of susceptibles and infected hosts (trees) at fixed period of time. We developed reversible jump MCMC (RJMCMC) algorithms that allow not only the imputation of missing information but also the future events. Similar to the non-spatial model, results showed that the posterior distributions of the model parameters do not change if we impute future events. We then moved on to construct various prioritisation measures including the posterior expectation of the risk $\mathcal{R}^*$, the challenge $\mathcal{C}^*$, the expectation of the product of the risk and challenge (threat) $\mathcal{T}^*$ and the product of the expectation of the risk and the expectation of the challenge $\mathcal{U}^*$ at both intervention day and the assessment day. The comparison of different control strategies on three designs space including strategy 1 (equal allocation), strategy 2 (probability-based allocation) and strategy 3 (adaptation of Neyman allocation) respectively reveals three important results on two different datasets (randomly generated locations and structured data of the well known 1995 Citrus canker epidemic from Florida (Neri et al., 2014)). First, the threat measure outperformed other measures. Second it was observed that there is no evidence to prefer using measure constructed on the intervention day over the one of the assessment day, results looked similar. Finally, when considering subregions, probability-based allocation proves to be more cost-effective compared to the adaptation of Neyman allocation. The evidence of this study that made these comparisons feasible is that the correlation between the outcomes of the epidemics with different controls including no control is positively strong. This implies a reduction in the variability between outcomes, thus giving a more accurate estimation of their expected difference which could be seen as their average causal effect (Holland, 1986).

Chapter 6 is divided into two parts where the same approach as in Chapter 5 is used. The first part was devoted to the application of the methodology of Chapter 5 to real life epidemic data namely the 1995 Citrus canker data from Florida (Gottwald et al., 2001a; Neri et al., 2014; Cunniffe et al., 2015). Results are similar to those of the previous chapter with two exceptions mainly due to the introduction of the removal process: The prioritisation is more cost-effective if resources are distributed following the adaptation of Neyman allocation and the prioritisation map is constructed at the intervention time. The second part is an extension to a more general framework where individuals in the population have different level of infectivities and susceptibilities. Results are qualitatively similar to those of Chapter 5 where there is no removal process included accentuating the superiority of the threat map as a tool to guide the design of control.

8.1.2 Modelling PPR

The main points of the case study in chapter 7 are summarised as follows:

8.1.2.1 Modelling and Bayesian analysis of spatio-temporal spread of PPR

Using contact distribution models, extended to include the effect a finite infectious period, a model for representing the spread of PPR within a country has been developed. A key feature is the representation of the susceptible population as a Poisson process modelled simultaneously with the infection process. This allows the models to be fitted to data on the time of location of infections without the need to know the locations of the susceptible units in the population. Statistical algorithms using MCMC have been developed to fit the models and verified using simulated data generated using a range of transmission kernels. The methods have been applied to the Tunisian data to estimate key parameters.

8.1.2.2 Modelling wave speeds

Methods for estimating the posterior distribution of a wave speed for a given spatio-temporal model have been developed (using a variant of the Gillespie algorithm to generate realisations from the posterior predictive distribution of the process). These have been applied to generate credible intervals for wave speeds using models with Rayleigh, Exponential and Cauchy kernels. Results using the Rayleigh kernel are consistent with the analysis carried out by EFSA (2015).

8.1.2.3 First-passage time models for spread of PPR through a network of countries

Using techniques from graph theory and operational research, an efficient representation was developed for the spread of PPR through a network of countries in which multiple modes of transmission are operating. The key feature is the modular approach to representing spread between any two countries in terms of the distribution times for the respective edge in the network, and the use of the Bellman-Ford algorithm to reduce the full network of edge-passage times to determine the time of first passage from any given source country and any given target. This structure means that the model can be adapted to new information on particular mechanisms simply by modifying the corresponding edge transmission time distributions.

8.2 Assumptions and caveats

Departing from temporal data in Chapter 4 to spatio-temporal data in Chapter 5 added more complexity and uncertainty to the estimation. Nevertheless, the models were considered with simplified assumptions. In fact, when assuming that the period it takes for symptoms to appear is fixed, we also assumed that such period is known.

Future research might explore the extent to which we can estimate that period and therefore investigate its sensitivity on the prioritisation. The approach to adopt can be motivated from the work of Cunniffe et al. (2015) where in their SCIR model the transition from cryptic class to infection class is modelled by exponential distribution. A more general distribution such as Weibull or Gamma distributions (O'Neill and Roberts, 1999; Stacey et al., 2004) could be applied.

Although results are presented in chapter 7, it should be noted that further investigations on the sensitivity of conclusions to assumptions indicate that there are currently major gaps in the understanding of PPR which means that predictions from the models here should not be treated as definitive.

Some assumptions have already been described in Section 7.6.7. It should be noted that posterior distributions of wave speeds are also sensitive to the assumption that the infectious period is constant over infectious premises. Relaxing this assumption leads to greater uncertainty in other parameters with consequently greater variance in predictive distributions. It is also assumed that the Tunisian data are complete in the sense that there are no unobserved infections. Again this assumption can be relaxed by imputing unobserved infections with a consequent inflation of the uncertainty in predictions.

8.3 Future work

8.3.1 Extensions of the control model using Sellke construction

The most pressing extension of this work is to consider more realistic designs. For example, removal or replacement of hosts within a certain distance of the detected infective (Cook et al., 2008; Neri et al., 2014; Cunniffe et al., 2015) and incorporate economic factors such as intervention costs in order to identify economically optimal strategies (Forster and Gilligan, 2007; Neri et al., 2014) or cost of detection (Dybiec et al., 2004; Dybiec and Gilligan, 2005; Dybiec et al., 2009). This may be readily effected within our approach by suitably adapting the impact function.

For simplicity, the diagnostic test considered throughout this thesis is assumed to be perfect. Our methodology can easily be extended to incorporate the imperfection of the tests (Cook et al., 2008). The controls considered in Chapter5 and 6 assumed that all resources are deployed at a particular time. However, the impact of the disease could be optimised by deploying control sequentially. For example, if we assume that the control will occur sequentially at $t_{obs} < t^1, \ldots, t^n \leq T$, then the designs will involve updating the measures at every t^j, $j = 1, \ldots, n$ while keeping track of the history of the controls (locations of hosts removed). This might be computationally

intense and very cumbersome. Research is needed to examine the effectiveness to varying the timing and intensity of controls.

We note that a very nice aspect of the approach of Lau et al. (2015), their integration of genetic information into epidemiological data, could easily and consistently be incorporated within our framework. By using Sellke thresholds as latent processes, our approach also allow assessments to be made of retrospective actions as argued by Cook et al. (2008) on historical epidemics data in which intervention for control was applied.

It would be interesting to assess the effects of the model parameters on the control strategies especially in the case where hosts in the population vary in term of their infectivities and susceptibilities since results (in Figure 6.17 for example) show a significant difference in the outcome of the control strategies compared to the single type host.

8.3.2 Extensions and modifications to first passage-time model

There are some further extensions to the network model that could be considered to improve its realism.

- One key assumption is that, once a given country A is infected, the rates of transmission from A to other countries via illegal or legal animal movement are immediately attained. This in effect assumes that steady-state prevalences are immediately attained and this is unrealistic. It would be possible to model the rate of transmission via movement as being time-varying, replacing λ with a function of the form $\lambda(t) = \lambda(1-exp(-t/\mu))$, where t represents the time elapsed since incursion into A and μ represents an additional parameter controlling how quickly asymptotic prevalence is attained. The simulation of the passage time in the model with this modification is effected by drawing t' from $Exp(\lambda)$ and then identifying the actual passage time, t, by solving numerically the equation

$$t' = t + \mu \exp(-t/\mu) - \mu \tag{8.3.1}$$

- A second assumption that also may serve to lead to short transmission times is the assumption that transboundary spread occurs as a consequence of the wave of spread that would occur within a country. In reality it may be the case that short-range local spread within a country, which is manifest in the wave speed may be partly due to local animal or human movement which does not occur to the same extent over national boundaries. The first-passage time model may be further adapted to accommodate this effect by having the distribution of time taken to spread from A to a neighbouring country B by transboundary

spread as being formed from two independent components $T_1 + T_2$. Here T_1 represents the time taken for the wave to reach the region of A which has a boundary with B and could be drawn from the Gamma distribution and T_2 represents the additional random time to traverse the boundary, perhaps drawn from an Exponential distribution. Such a modification to the model would be straightforward to implement, illustrating the adaptability of the general approach.

Bibliography

Addy, C., Longini, I., and Haber, M. (1991). A generalized stochastic model for the analysis of infectious disease final size data. *Biometrics*, 47:961–974.

Alabi, O., Kumar, P., and Naidu, R. (2011). Cassava mosaic disease: A curse to food security in sub-saharan Africa. *APS J.*

Anderson, R. (1982). Fox rabies. In Anderson, R., editor, *The population dynamics of infectious diseases: Theory and applications*, pages 242–261. Chapman and Hall, London.

Anderson, R. and May, R. (1981). The population dynamics of microparasites and their invertebrate hosts. *Philos. T. Roy. Soc. B*, 291:451–524.

Anderson, R. and May, R. (1991). *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press.

Andersson, H. and Britton, T. (2000). *Stochastic Epidemic Models and Their Statistical Analysis*. Springer.

Ault, S. (1994). Environmental management: a re-emerging vector control strategy. *Am. J. Trop. Med. Hyg.*, 50:35–49.

Bailey, N. (1953). The use of chain-binomials with a veritable chance of infection for analysis of intra-household epidemics. *Biometrika*, 40:279–286.

Bailey, N. (1975). *The Mathematical theory of infectious diseases and its applications*. Griffin, London.

Ball, F. (1986). A unified approach to the distribution of total size and total area under the trajectory of infectives in epidemic models. *Adv. Appl. Probab.*, 18:289–310.

Ball, F. and Britton, T. (2005). An epidemic model with exposure-dependent severities. *J. Appl. Probab.*, 42:932–949.

Ball, F. and Lyne, O. (2002). Optimal vaccination policies for stochastic epidemics among a population of households. *Math. Biosci.*, 177-178:333–354.

Ball, F., Mollison, D., and Scalia-Tomba, G. (1997a). Epidemics with two levels of mixing. *Ann. Appl. Probab.*, 7:46–89.

Ball, F., Mollison, D., and Scalia-Tomba, G. (1997b). Epidemics with two levels of mixing. *Ann. Appl. Probab.*, 7:46–89.

Ball, F. and O'Neill, P. (1999). The distribution of general final state random variables for stochastic epidemic models. *J Appl Probab*, 36:473–491.

Barlett, M. (1949). Some evolutionary stochastic processes. *J. R. Stat. Soc.*, 11:211–229.

Bassanezi, R., Montesino, L., Gasparoto, M., Filho, A., and Amorim, L. (2011). Yield loss caused by huanglongbing in different sweet orange cultivars in são paulo brasil. *Eur. J. Plant. Pathol.*, 130:577–586.

Becker, N., Britton, T., and O'Neill, P. (2003). Estimating vaccine effects on transmission of infection from household outbreak data. *Biometrics*, 3:467–475.

Becker, N., Britton, T., and O'Neill, P. (2005). Bayesian inference for epidemics with two levels of mixing. *Scand. J. Stat.*, 32:265–280.

Becker, N. and Dietz, K. (1995). The effect of the household distribution on transmission and control fo highly infectious diseases. *Math. Biosci.*, 127:207–219.

Becker, N. and Dietz, K. (1996). In athens conference on applied probability and time series (springer lecture notes in statistics). In Heyde, C., Prohorov, Y. V., Pyke, R., and Rachev, S., editors, *Plant Pathology: An Advance Treatise*, volume 3, pages 267–276. Springer-Verlag, New York.

Bernoulli, D. (1760). Essaie d'une nouvelle analyse de la mortalité causée par la petite vérole et des avantages de l'inoculation pour la prévenir. *French Royal Academy of Sciences in Paris*, pages 1–45.

Bertsimas, D. and Tsitsiklis, J. (1998). Simulated Annealing. *Stat. Sci.*, 8:10–15. `http://links.jstor.org/sici?sici=0883-4237%28199302%298%3A1%3C10%3ASA%3E2.0.CO%3B2-23`.

Besag, J. (1997). Discussion of "Spatial contact models for ecological and epidemic spread" by D. Mollison. *J. R. Stat. Soc.*, 39:315–316.

Bessell, P. R. (2009). The spatial epidemiology of foot and mouth disease in great britain. PhD Thesis.

Billard, L., Medley, G. F., and Anderson, R. M. (1990). The incubation period for AIDS virus. Lecture Notes in Bioinformatics.

Boender, G. J., Hagenaars, T. J., Bouma, A., Nodelijk, G., Elbers, A. R., de Jong, M. C., and van Boven, M. (2007). Risk maps for the spread of highly pathogenic avian influenza in poultry. *Plos Comput. Biol.*, 3:704–712.

Bootsma, M. C. and Ferguson, N. M. (2007). The effect of public health measures on the 1918 influenza pandemic in U.S. cities. *Proc. Natl. Acad. Sci. USA*, 104:7588–7593.

Boudreau, M. and Mundt, C. (1982). Ecological approaches to disease control. In Rechcigl, N. and Rechcigl, J., editors, *Environmentally Safe Approaches to Crop Disease Control*, pages 33–62. CRC Publications, Boca Raton.

Bové, J. (2006). Huanglongbing: A destructive, newly-emerging, century-old disease of citrus. *J. Plant. Pathol.*, 88:7–37.

Brent, R. (1977). *Algorithms for minimization without derivatives*. Englewood Cliffs, NJ: Prentice-Hall.

Britton, T. (2010). Stochastic epidemic models: a survey. *Math Biosci*, 225:24–35.

Britton, T. and Becker, N. (2000). Estimating the immunity coverage required to prevent epidemics in a community of households. *Biostatistics*, 1:389–402.

Britton, T., Kypraios, T., and O'Neill, P. (2011). Inference for epidemics with three levels of mixing: Methodology and application to a measles outbreak. *Scand. J. Stat.*, 38:578–599.

Britton, T. and O'Neill, P. (2002). Bayesian inference for stochastic epidemics in populations with random social structure. *Scan. J. Stat.*, 29:375–390.

Burdon, J. and Chilvers, G. (1982). Host density as a factor in plant disease ecology. *Annu. Rev. Phytopathol.*, 20:143–166.

Busetti, F. (2003). Simulated Annealing overview. `http://www.aiinfinance.com/saweb.pdf`.

Cabi (2016). Xanthomonas citri (citrus canker). `http://www.cabi.org/isc/datasheet/56921`. Invasive Species Compendium Datasheets, maps, images, abstracts and full text on invasive species of the world.

Carlin, B. P. and Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis*. Chapman & Hall, second edition.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Cambridge University Press, 2nd edition.

Castilho, C. (2006). Optimal control of an epidemic through educational campaingns. *Proc. Natl. Acad. Sci. USA*, 2006(125):1–11.

CDC (2014). Ebola response roadmap. `http://www.cdc.gov/vhf/ebola/travelers/ebola-screening-factsheet.html`.

CDC (2015). Ebola virus disease. `http://www.cdc.gov/vhf/ebola/travelers/ebola-screening-factsheet.html`.

Celeux, G., Forbes, F., Robert, C., and Titterington, M. (2006). Deviance information criteria for missing data models. *Bayesian Anal.*, 1:651–673.

CFSPH and IICAB (2008). Peste des petits ruminants. `http://www.cfsph.iastate.edu/Factsheets/pdfs/peste_des_petits_ruminants.pdf`.

Chis-Ster, I., Singh, B. K., Ferguson, M. N., and Christi, A. (2009). Epidemicological inference for partially observed epidemics: The example of the 2001 foot and mouth epidemic in Great Britain. *Epidemics*, 1:21–34.

Civerolo, E. (1984). Bacterial canker disease of citrus. *Journal Rio Grande Valley Horticultural Society*, 37:127–146.

Clark, J. S., Silman, M., Kern, R., Macklin, E., and Hillerislambers, J. (1999). Seed dispersal near and far: Patterns across temperate and tropical forests. *Ecology*, 80:1475–1494.

Cook, A., Marion, G., Butler, A., and Gibson, G. (2007). Bayesian inference for the spatio-temporal invasion of alien species. *B. Math. Biol.*, 69:2005–2025.

Cook, A. R., Gibson, G. J., Gottwald, T., and Gilligan, C. A. (2008). Construction the effect of alternative intervention strategies on historic epidemics. *J. R. Soc. Interface*, 5:1203–1213.

Corana, A., Marchesi, M., Martuni, C., and Ridella, S. (1987). Minimizing multimodal functions of continuous variables with the simulated annealing algorithm. *ACM T. Math. Software*, 13:262–280. `http://amcg.ese.ic.ac.uk/jgomes/lasme/SA-corana.pdf`.

Cunniffe, N. J., Laranjeira, F. F., Neri, F. M., DeSimone, R. E., and Gilligan, C. A. (2014). Cost-effective control of plant disease when epidemiological knowledge is incomplete: Modelling bahia bark scaling of citrus. *Plos Comput. Biol.*, 10.

Cunniffe, N. J., Stutt, R. O., DeSimone, R. E., Gottwald, T. R., and Gilligan, C. A. (2015). Optimising and communicating options for the control of invasive plant disease when there is epidemiological uncertainty. *New Phytologist*, 4:11.

Dagum, L. and Menon, R. (1998). OpenMP: an industry standard API for shared-memory programming. *IEEE Comput. Sci. Eng.*, 5:46–55.

Daley, D. and Gani, J. (1999). *Epidemic Modelling: An introduction*. Cambridge University Press.

Dawid, A. and Stone, M. (1982). The functional model basis of fiducial inference. *Ann. Stat.*, 10:1054–1067.

Deardon, R., Brooks, S. P., Grenfell, B. T., Keeling, M. J., Tildesley, M. J., Savill, N. J., Shaw, D. J., and Woolhouse, M. E. (2010). Inference for individual-level models of infectious diseases in large populations. *Stat. Sinica.*, 20:239–261.

DEFRA (2013). Chalara management plan. `http://www.defra.gov.uk/publications/`.

Demon, I., Cunniffe, N., Marchant, B., Gilligan, C., and van den Bosch, F. (2011). Spatial sampling to detect an invasive pathogen outside of an eradication zone. *Phytopathology*, 101:725–731.

Dietz, K. (1981). The evaluation of rubella vaccination strategies. In Hions, R. and Cooke, D., editors, *The Mathematical Theory of the Dynamic of Biological Populations II*, pages 81–97. Academic Press, London.

Dietz, K. and Schenzle, D. (1985). Mathematical models for infectious disease statistics. In Atkinson, A. C. and Fienberg, S. E., editors, *A Celebration of Statistics*, pages 167–204. Springer, New York.

Donnelly, C. A., Woodroffe, R., Cox, D., Bourne, F. J., Cheeseman, C., Clifton-Hadley, R. S., Wei, G., Gettinby, G., Gilks, P., Jenkins, H., Johnston, W. T., Fevre, A. M. L., McInerney, J. P., and Morrison, W. I. (2006). Positive and negative effects of widespread badger culling on tuberculosis in cattle. *Nature*, 439:843–846.

Dybiec, B. and Gilligan, C. A. (2005). Opimising control of disease spread on networks. *Acta. Phys. Pol. B*, 36:1509–1526.

Dybiec, B., Kleczkowski, A., and Gilligan, C. A. (2004). Controlling disease spread on networks with incomplete knowledge. *Phys. Rev. E*, 70:066145.

Dybiec, B., Kleczkowski, A., and Gilligan, C. A. (2009). Modelling control of epidemics spreading by long-range interactions. *J. R. Soc. Interface*, 39:941–950.

Dzhini-Obiatey, H., Ameyaw, G. A., and Ollennu, L. (2006). Control of cocoa swollen shoot disease by eradicating infected trees in Ghana: A survey of treated and replanted areas. *Crop Prot.*, 25:647–652.

EFSA (2015). Scientific opinion on peste des petits ruminants. *EFSA Journal*, 13:3985.

Elton, C. S. (1958). *The Ecology of Invasions by Animals and Plants*. Springer US.

FAD PReP and USDA (2013). Peste des Petits Ruminants Standard Operating Procedures: Overview of Etiology and Ecology. `https://www.aphis.usda.gov/animal_health/emergency_management/downloads/sop/sop_ppr_e_and_e.pdf`. SOP Manual; draft November 2013.

FAO (1999). Recognizing peste des petits ruminants. `http://www.cfsph.iastate.edu/Factsheets/pdfs/peste_des_petits_ruminants.pdf`.

Fenner, F., Henderson, D. A., Arita, I., JeZek, Z., and Ladnyi, I. D. (1988). *Smallpox and its Eradication*. World Health Organization Geneva.

Ferguson, N. M., Donnelly, C. A., and Anderson, R. M. (2001). The foot and mouth epidemic in Great Britain: Pattern of spread and impact of interventions. *Science*, 292:1155–1160.

Finckh, M. and Wolfe, M. (2006). Diversification strategies. In et al., B. C., editor, *The Epidemiology of Plant Diseases*, pages 269–308. Springer.

Finckh, M. and Wolfe, M. S. (1997). The use of biodiversity to restrict plant diseases and some consequences for farmers and society. In Jackson, L. E., editor, *Ecology in Agriculture*. Cambridge University Press.

Forster, G. and Gilligan, C. (2007). Optimizing the control of disease infestations at the landscape scale. *Proc. Natl. Acad. Sci. USA*, 104:4984–4989.

Fraser, C., Donnelly, C., Cauchemez, S., Hanage, W., Kerkhove, M. V., Hollingsworth, T., Griffin, J., Baggaley, R., Jenkins, H., Lyons, E., Jombart, T., Hinsley, W., Grassly, N., Balloux, F., Ghani, A., Ferguson, N., Rambaut, A., Pybus, O., Lopez-Gatell, H., Alpuche-Aranda, C., Chapela, I., Zavala, E. P., Guevara, D. E., Checchi, F., Garcia, E., Hugonnet, S., and Roth, C. (2009). Pandemic potential of a strain of influenza A H1N1: Early findings. *Science*, 324:1557–1561.

Gambley, C., Miles, A., Ramsden, M., Doogan, V., Thomas, J., Parmenter, K., and Whittle, P. (2009). The distribution and spread of citrus canker in Emerald, Australia. *Australas. Plant. Path.*, 38:547–557.

Gani, J., Yakowitz, S., Blount, M., and Blount, M. (1997). The spread and quarantine of hiv infection in a prison system. *SIAM J. Appl. Math.*, 57:1510–1530.

Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequence. *Stat. Sci.*, 7:457–511.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distribution and bayesian restoration of images. *IEEE T. Pattern Anal.*, 6:721–741.

Geweke, J. (1991). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Research Department Staff Report 148.

Gibson, G. (1997). Investigating mechanisms of spatiotemporal epidemic spread using stochastic models. *APS Symp. Ser.*, 87:139.

Gibson, G. and Austin, E. (1996). Fitting and testing spatio-temporal stochastic models with application in plant epidemiology. *J. R. Stat. Soc.*, 45:172–184.

Gibson, G. J. (1996). Markov chain monte carlo methods for fitting spatio-temporal sotchastic models in plant epidemiology. *J. R. Stat. Soc.*, 46:215–233.

Gibson, G. J. and Renshaw, E. (1998). Estimating parameters in stochastic compartemental model using Markov chain methods. *IMA J. Math. Appl. Med.*, 15:19–40.

Gibson, R. and Aritua, R. (2002). The perspective of sweet potato chorotic stunt virus in sweet potato production in africa: A review. *African Crop Science Journal*, 10(4):1021–9730.

Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.*, 22:403–434.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81:2340–2361.

Gilligan, C. A., Truscott, J. E., and Stacey, A. J. (2007). Impact of scale on the effectiveness of disease control strategies for epidemics with cryptic infection in a dynamical landscape: an example for a crop disease. *Acta. Phys. Pol.*, 4:925–934.

Gottwald, T. (2010). Current epidemiological understanding of citrus huanglongbing. *Annu. Rev. Phytopathol.*, 48:119–139.

Gottwald, T., Cambra, M., Moreno, P., Camarasa, E., and Piquer, J. (1996). Spatial and temporal analyses of citrus tristeza virus in eastern spain. *Phytopathology*, 86:45–55.

Gottwald, T., Hughes, J., and Schubert, T. (1997a). Citrus canker in urban Miami: An analysis of spread and prognosis for the future. *Citrus Ind.*, 78:72–78.

Gottwald, T., Hughes, J., and Schubert, T. (1997b). An epidemiological analysis of the spred of citrus canker in urban miami, florida, and synergistic interaction with the asian citrus leafminer. *Fruits*, 52:371–378.

Gottwald, T., McGuire, R., and Garran, S. (1988). Asiactic citrus canker: spatial and temporal spread in simulated new planting situations in argentina. *Phytopathology*, 78:739–745.

Gottwald, T., Reynolds, K., Campbell, C., and Timmer, L. (1992). Spatial and spatiotemporal autocorrelation anlysis of citrus anker epidemics in citrus nurseries and groves in argentina. *Phytopathology*, 82:843–851.

Gottwald, T. and Timmer, L. (1995). The efficacy of windbreaks in reducing the spread of citrus canker caused by xanthomonas compestris pv. citri. *Trop. Agric.*, 72:194–201.

Gottwald, T. R., Graham, J. H., and Schubert, T. S. (2002). Citrus Canker: The Pathogen and its Impact. *APS J.*

Gottwald, T. R., Hughes, G., Graham, J. H., Sun, X., and Riley, T. (2001a). The citrus canker epidemic in florida: The scientific basis of regulatory eradication policy for an invasive species. *Phytopathology*, 91:30–34.

Gottwald, T. R. and Irey, M. (2007). Post-hurricane analysis of citrus canker ii: Predictive model estimation of disease spread and area potentially impacted by various eradication protocols following catastrophic weather events. *APS J.*

Gottwald, T. R., Sun, X., Ripley, T., Graham, J. H., Ferrandino, F., and Taylor, E. L. (2001b). Geo-Reference Spatiotemporal Analysis of the Urban Canker Epidemic in Florida. *Phytopathology*, 92:361–377.

Graham, J. H., Gottwald, T. R., Cubero, J., and Achor, D. S. (2004). Xanthomonas axonodis pv. citri: factors affecting successful eradiation of citrus canker. *Mol. Plant Pathol.*, 5.

Green, P. J. (Dec 1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, 82:711–732.

Greenwood, M. (1946). The statistcal study of infectious diseases, part II. *J. R. Stat. Soc.*, 109:85–103.

Greenwood, M. (1949). The infectiousness of measles. *Biometrika*, 36:1–8.

Grimmett, G. and Stirzaker, D. (2001). *Probability and Random Proesses*. Oxford University Press: Oxford, 3rd edition.

Handel, I. G., Barend, M., Bronsvoort, C., Forbes, J. F., and Woolhouse, M. E. J. (2011). Risk-targeted selection of agricultural holdings for post-epidemic surveillance: Estimation of efficiency gains. *New Phytologist*, 6:5.

Hastings, W. (1970). Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika*, 57:97–109.

Haydon, D., Chase-Topping, M., Shaw, D., Matthews, L., Friar, J., Wilesmith, J., and Woolhouse, M. (2003). The construction and analysis of epidemic trees with reference to the 2001 UK foot and mouth outbreak. *Proc. Biol. Sci.*, 270:121–127.

Heesterbeek, J. and Dietz, K. (1996). The concept of R0 in epidemic theory. *Stat. Neerl.*, 50:89–110.

Hillocks, R. and Thresh, J. (2000). Cassava mosaic and cassava brown streak virus diseases in Africa: A comparative guide to symptoms and aetiologies. *Roots*, 7.

Holland, P. W. (1986). Statistics and causal inference. *J. Am. Stat. Assoc.*, 81:945–960.

Iftimia, A., Montesa, F., Santiynb, A. M., and Martnez-Ruizc, F. (2015). Space-time airborne disease mapping applied to detect specific behaviour of varicella in valencia, spain. *Spat. Spatiotemporal Epidemiol.*, 14-15:33–34.

Isham, V. (2005). Stochastic models for epidemics. In Davison, A. C., Dodge, Y., and Wermuth, N., editors, *Celebrating Statistics: Papers in Honour of Sir David Cox on his 80th Birthday*. Oxford University Press.

Jewell, C. P., Kypraios, T., Neal, P., and Roberts, G. O. (2009). Bayesian analysis for emerging infectious diseases. *Bayesian Analysis*, 4:239–261.

Kao, R. (2003). The impact of local heterogeneity on alternative control strategies for foot and mouth disease. *J. R. Stat. Soc.*, 270:2557–2564.

Keeling, M. J., Woolhouse, M. E., Shaw, D. J., Matthews, L., Chase-Topping, M., Haydon, D. T., Cornell, S. J., Kappey, J., Wilesmith, J., and Grenfell, B. T. (2001). Dynamics of the 2001 uk foot and mouth epidemic: Stochastic dispersal in a heterogeneous landscape. *Science*, 294:813–817.

Kermack, W. and McKendrick, A. (1927). A contribution to the mathematical theory of epidemics. *P. R. Soc. London*, 115:1–45.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by Simulated Annealing. *Am. Assoc. Adv. Sci.*, 220:671–680. http://links.jstor.org/sici?sici=0036-8075%2819830513%293%3A220%3A4598%3C671%3AOBSA3E2.0.CO%3B2-8.

Klinkenberg, D., der Wind, A. E.-V., Graat, E. M., and de Jong, M. (2003). Quantification of the effect of control strategies on classical swine fever epidemics. *Math. Biosci.*, 186:145–173.

Knox, E. (1980). Strategy for rubella vaccination. *Int. J. Epidemiol.*, 9:13–23.

Koeijer, A. D., Diekmann, O., and Reijnders, P. (1998). Modelling the spread of phocine distemper virus among harbour seals. *Bul. Math. Biol.*, 60:585–596.

Koopman, J. S., Simon, C. P., Jacquez, J. A., and Park, T. S. (1989). Selective contact within structured mixing with an application to HIV transmission risk from oral and anal sex. Lecture Notes in Bioinformatics.

Kuhara, S. (1978). Present epidemic status and control of the citrus canker disease [xanthomonas citri (hasse) dowson] in japan. *Review of Plant Protection Research*, 11:132–142.

Lau, M., Marion, G., Streftaris, G., and Gibson, G. J. (2014). New model diagnostics for spatio-temporal systems in epidemiology and ecology. *J R Soc Interface*, 11.

Lau, M. S. Y., Marion, G., Streftaris, G., and Gibson, G. J. (2015). A systematic bayesian integration of epidemiological and genetic data. *Plos Comput. Biol.*, 11.

Lindvall, T. (1992). *Lectures on the coupling method*. Wiley, New York.

Lipsitch, M., Cohen, T., Cooper, B., Robins, J., Ma, S., James, L., Gopalakrishna, G., Chew, S., Tan, C., Samore, M., Fisman, D., and Murray, M. (2003). Transmission dynamics and control of severe acute respiratory syndrome. *Science*, 300:1966–1970.

Longini, I. and Koopman, J. (1982). Household and community transmission parameters from final distributions of infections in households. *Biometrics*, 38:115–126.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). Winbugs- a bayesian modelling framework: Concepts, structure, and extensibility. *Stat. Comput.*, 10:325–337.

Macdonald, G. (1957). *The epidemiology and control of malaria*. Oxford University Press, London.

Marchant, B. and Lark, R. (2007). Optimized sample schemes for geostatistical surveys. *Math. Geol.*, 39:113–134.

Mayer, T. and Zignago, S. (2005). Notes on CEPIIs distances measures: The geodist database. Document de travail.

Mayr, E. (1982). *The Growth of Biological Thought: diversity, Evolution, and Inheritance*. The Belknap Press of Harvard University Press Cambridge, Massachusetts London, England.

McCullach, P. (1980). Regression models for ordinal data. *J. R. Stat. Soc.*, 42:109–142.

McKendrick, L.-C. A. G. (1926). Applications of Mathematics to Medical Problems. *P. Edinburgh Math. Soc.*, 14:98–130.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, H. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1091.

Metropolis, N. and Ulam, S. (1949). The Monte Carlo method. *J. Am. Stat. Assoc.*, 44:335–341.

Mollison, D. (1977). Spatial contact models for ecological and epidemic spread. *J. R. Stat. Soc.*, 39:283–326.

Morris, C. (1983). Parametric empirical bayes inference: Theory and applicatio. *J. Am. Stat. Assoc.*, 78:45–55.

Morris, R., Wilesmith, J., Stern, M., Sanson, R., and Stevenson, M. (2001). Predictive spatial modelling of alternative control strategies for the foot and mouth disease epidemic in great britain. *Vet. Rec.*, 149:137–145.

Müller, P. (1999). Simulation-Based Optimal Design. *Bayesian Statistics*, 6:459–474.

Müller, P. and Parmigiani, G. (1995). Optimal design via curve fitting of Monte Carlo experiments. *J. Am. Stat. Assoc.*, 90:1322–1330.

Mundt, C. (2002). Use of multiline cultivars and cultivar mixtures for disease management. *Annu. Rev Phytopathol.*, 40:381–410.

Neal, P. and Roberts, G. (2005). A case study in non-centering for data augmentation: stochastic epidemics. *Stat. Comput.*, 15:315–327.

Nelson, S. (1995). Spatiotemporal distance class analysis of plant disease epidemics. *Phytopathology*, 85:37–43.

Neri, F. M., Cook, A. R., Gibson, G. J., Gottwald, T., and Galligan, C. A. (2014). Bayesian analysis for inference of an emerging epidemic: Citrus canker in urban landscapes. *Plos Comput. biol.*, 10.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *J. R. Stat. Soc.*, 97:558–606.

OIE (2015). Peste des petits ruminants (PPR). General Disease Inforamtion Sheets.

O'Neill, P., Balding, D., Becker, N., Eerola, M., and Mollison, D. (2000). Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. *J. R. Stat. Soc.*, 49:517–542.

O'Neill, P. D. (2010). Introduction and snapshot review: Relating infectious disease transmission models to data. *Stat. Med.*, 29:2069–2077.

O'Neill, P. D. and Becker, N. (2001). Inference for an epidemic where susceptibility varies. *Biostatistics*, 2:99–108.

O'Neill, P. D. and Roberts, G. O. (1999). Bayesian inference for partially observed stochastic epidemics. *J. R. Stat. Soc.*, 162:121–129.

Papaspiliopoulos, O. and Roberts, G. O. (2013). Non-centered parameterisations for hierarchical models and data augemtation. *Bayesisan Statistics*, 7:307–326.

Parnell, S., Gottwald, T., van den Bosch, F., and Galligan, C. (2009). Optimal strategies for the eradication of asiatic citrus canker in heterogeneous host landscapes. *Phytopathology*.

Parry, M., Gibson, G. J., Parnell, S., Gottwald, T. R., Irey, M. S., Gast, T. C., and Gilligan, C. A. (2014). Bayesian inference for an emerging arboreal epidemic in the presence of control. *J. Appl. Probab.*, 111:6258–6262.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6.

Raftery, A. E. and Lewis, S. (1992). How many iterations in the Gibbs sampler. In Bernardo, J., Berger, J., Dawidand, A., and Smith, A., editors, *Bayesian Statistics 4*, volume 4, pages 763–773. Oxford University Press.

Rajasekaran, S. (1990). On the Convergence Time of Simulated Annealing. Technical Report MS-CIS-90-89, University of Pennsylvania, Department of Computer & Information Science.

Renshaw, E. (1991). *Modelling biological populations in space and time*. Cambridge University Press, Departement of Statistics and Modelling Science University of Strathclyde.

Reynolds, K. and Madden, L. (1988a). Analysis of epidemics using spatio-temporal autocorrelation. *Phytopathology*, 78:240–246.

Reynolds, K. and Madden, L. (1988b). Spatio-temporal analysis of epidemic development of leather rot of strawberry. *Phytopathology*, 78:246–252.

Ripley et al. (2003). Transmission dynamics of the etiological agent sars in hong kong: impact of public health interventions. *Science*, 300:1961–1966.

Roberts, G. (2012). ST911 Fundamentals of Statistical Inference Part III. Course notes.

Ross, R. (1911). *The prevention of Malaria*. London: Murray, second edition.

Schubert, T. S., Rixvi, S. A., Sun, X., Gottwald, T. R., Graham, J. H., and Dixon, W. N. (2001). Meeting the challenge of eradicatiing citrus canker in florida again. *APS J.*, 85:340–356.

Schumann, G. L. (1991). *Plant Diseases: Their Biology and Social Impact*. Aps Symp Ser.

Sellke, T. (1983). On the asymptotic distribution of the size of a stochastic epidemic. *J. Appl. Probab.*, 20:390–394.

Selvey, L. A., ao, C. A., and Hall, R. (2015). Evaluation of border entry screening for infectious diseases in humans. *Emerging Infectious Disease*, 21:197–201.

Serizawa, S. (1981). Recent studies on the behaviour of the causal bacterium of the citrus canker. *Proceedings of the International Society of Citriculture*, 1:395–397.

Service, M. (1991). Agricultural development and arthropod-borne diseases: a review. *Rev. Saude Pblica*, 25:167–178.

Shaddick, G. (2013). Pumps, maps and pea soup: Spatio-temporal methods in environmental epidemiology. 2012-13 van Eeden lecture.

Shibli, M., Gooch, S., Lewis, H., and Tyrell, D. (1971). Common colds on tristan da cunha. *J. Hyg.*, 69:255–265.

Smith, B. (2007). Boa: An R package for mcmc output convergence assessment and posterior inference. *J. Stat. Software*, 21(11).

Smith, G. C. (1964). Factors in the transmission of virus infectius from animals to man. In *The scientific basis of Medicine annual reviews*, pages 125–150. Athlone Press, London.

South, A. (2011). rworldmap: A new R package for mapping global data. *The R Journal*, 3:35–43.

Spiegelhalter, D., Best, N., Carlin, B., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. Roy. Stat. Soc. B*, 64:583–639.

Stacey, A. J., Truscott, J. E., Asher, M. J. C., and Gilligan, C. A. (2004). A model for the invasion and spread of rhizomania in the united kingdom: Implications for disease control strategies. *Phytopathology*, 94:209–215.

Stall, R., Miller, J., Marco, G., and de Echenique, B. C. (1980). Population dynamics of xanthomonas citri causing cancrosis of citrus in argentina. *Proc. Fl. State Hortic.*, 93:10–14.

Streftaris, G. and Gibson, G. (2004). Bayesian inference for stochastic epidemics in closed populations. *Stat. Model.*, 4:63–75.

Streftaris, G. and Gibson, G. (2012). Non-exponential tolerance to infection in epidemic systems-modeling, inference, and assessment. *Biostatistics*, 13:580–593.

te Beest, D. E., Agenaars, T. J., Stegeman, J. A., Koopmans, M. P., and van Boven, M. (2011). Risk based culling for highly infectious diseases of livestock. *Vet. Res.*, pages 42–81.

Thippeswamy, K., Hanumanthappa, J., and ManjaiahD, H. (2010). A study on contrast and comparison between Bellman-Ford algorithm and Dijkstras algorithm. `file:///H:/.windows_settings/My%20Documents/Downloads/A_Study_on_Contrast_and_Comparison_between_Bellman-Ford_algorithm_and_Dijkstras_Algorithms.pdf`.

Thompson, D., Murriel, P., Russell, D., Osborne, P., Bromley, M., Creigh-Tyte, S., and Brown, C. (2004). Economic costs of the foot and mouth disease outbreak in the United Kingdom in 2002. *Rev. Sci. Tech.*, 21:675–687.

Tildesley, M. J., Bessell, P. R., Keeling, M. J., and Woolhouse, M. E. (2009). The role of pre-emptive culling in the control of foot and mouth disease. *J. R. Stat. Soc.*, 279:3239–3248.

USDA/APHIS, in consultation with the Florida citrus industry, F., and other stakeholders (2006). Citrus health response program (chrp) minimum standards for citrus health in florida. `http://www.aphis.usda.gov/plant_health/plant_pest_info/citrus/downloads/chrp.pdf`. Version 1.0.

van Groenigen, J. (1999). Constrained optimisation ofspatial sampling using continuous Simulated Annealing. *J. Environ. Qual.*, 27:1078–1086.

VanderPlank, J. (1960). Analysis of epidemics. In Horsfall, J. G. and Dimond, A. E., editors, *Plant Pathology III*, pages 229–289. Academic Press, New York USA.

VanderPlank, J. (1963). *Plant Diseases: Epidemics and Control*. Academic Press, New York.

Wallinga, J. and Teunis, P. (2004). Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am. J. Epidemiol.*, 160:509–516.

Whittle, P. (1955). The outcome of a stchastic epidemic-a note on Bailey's paper. *Biometrika*, 42:112–122.

WHO (1982). Manual on environmental management for mosquito control with special emphasis on mosquito vectors. *W.H.O offset publication*, 66. Geneva, Sw: W.H.O.

WHO (2003). Global alert and response. update 11WHO recommends new measures to prevent travel-related spread of sars. `http://www.who.int/csr/sars/archive/2003_03_27/en/`.

Wikipedia (2016). ISO 3166-1 alpha-3 — Wikipedia, the free Encyclopedia. `https://en.wikipedia.org/wiki/ISO_3166-1_alpha-3`. Online; accessed 31-March-2016.

Williams, T. (1971). Algebraic proof of the treshold theorem for the general stochastic epidemic. *Adv. Appl. Probab.*, 3:223.

Wilson, E. and Burke, M. (1942). The epidemic curve. *Proc. Natl. Acad. Sci. USA*, 28:361–367.

Wilson, E. and Burke, M. (1943). The epidemic curve, II. *Proc. Natl. Acad. Sci. USA*, 29:43–48.

Wolfe, M. (1988). The use of variety mixtures to control disease and stabilize yield. In Simmonds, N. and Rajaram, S., editors, *Breeding strategies for resistance to the rust of wheat*, pages 91–100. CIMMYT, Mexico.

Woolhouse, M. (2003). Foot and mouth disease in the UK: What should we do next time? *J. Appl. Microbiol.*, 94:126S–130S.

Woolhouse, M. (2011). How to make prediction about future infectious disease risks. *Philos T. Roy. Soc. B*, 366:2045–2054.