Spatial survival analysis of infectious animal diseases

submitted by

Trevelyan John McKinley BSc.

to the University of Exeter

as a thesis for the degree of

Doctor of Philosophy in Mathematics.

March 2007

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signature . . . . . .

Abstract

This thesis investigates the feasibility of using spatial survival modelling techniques to develop dynamic space-time predictive models of risk for infectious animal disease epidemics. Examples of diseases with potentially vast socioeconomic impacts include avian influenza, bovine tuberculosis and foot-and-mouth disease (FMD), all of which have received wide coverage in the recent media. The relatively sporadic occurrence of such large scale animal disease outbreaks makes determination of optimal control policies difficult, and policy makers must balance the relative impacts of different response strategies based on little prior information. It is in this situation that the use of mathematical and statistical modelling techniques can provide powerful insights into the future course of an infectious epidemic.

The motivating example for this thesis is the outbreak of FMD in Devon in 2001, however we are interested in developing more general techniques that can be applied to other animal diseases. Many of the models fitted to the 2001 UK FMD data set have focussed on modelling the global spread of the disease across the entire country and then using these models to assess the effects of nationwide response strategies. However it has been shown that the dynamics of the disease are not uniform across the whole of the UK and can vary significantly across different spatial regions. Of interest here is exploring whether modelling at a smaller spatial scale can provide more useful measures of risk and guide the development of more efficient control policies.

We begin by introducing some of the main epidemiological issues and concepts involved

in modelling infectious animal diseases, from the microscopic through to the farm population level. We then discuss the various mathematical modelling techniques that have applied previously and how they relate to various biological principals discussed in the earlier chapters. We then highlight some limitations with these approaches and offer potential ways in which survival analysis techniques could be used to overcome some of these problems.

To this end we formulate a spatial survival model and fit it to the Devon data set with some naive initial covariates that fail to capture the dynamics of the disease. Some work by colleagues at the Veterinary Laboratories Agency, Weybridge (Arnold 2005), produced estimates of viral excretion rates for infected herds of different species type over time, and these form the basis for the development of a dynamic space-time varying viral load covariate that quantifies the viral load acting at any spatial location at any point in time. The novel use of this covariate as a means of censoring the data set via exposure is then introduced, though the models still fail to explain the variation in the epidemic process.

Two potential reasons for this are identified - the possible presence of non-localised infections and/or premise varying susceptibility. We then explore ways in which the survival approach can be extended to model more than one epidemic process through the use of mixture and long-term survivor models. Some simple simulations suggest that resistance to infection is the most likely cause of the poor model fits, and a series of more complex simulation experiments show that both the mixture and long-term survivor models offer various advantages over the conventional approach when resistance is present in the data set. However key to their performance is the ability to correctly capture the mixing, although in the worst case scenario they still replicate the results from the conventional model.

We also use these simulations to explore potential ways in which space-time predictions of the hazard of infection can be used as a means of targeting control policies to areas of 'high-risk' of infection. This shows the importance of ensuring that the scale of the control order matches the scale of the epidemic, and suggests possible dangers when using global

level models to derive response strategies for situations where the dynamics of the disease change at smaller spatial scales. Finally we apply these techniques to the Devon data set and offer some conclusions and future work.

Dedication

To Mum and all my family, who have given me their full and unconditional support. I love you all very much.

To Michelle, without whom I would never have had the courage to even attempt a PhD, and whose love, beauty and humour over the past seven years has provided me with more happiness than I ever thought possible.

And to Dad, whose guidance and wisdom has always been, and will always be invaluable. It gives me some comfort to know that you saw me submit this thesis, even though you are not able to see me graduate. I know that your faith in me, like my faith in you, never wavered.

Acknowledgements

I am privileged to have been surrounded by so many people who have consistently provided me with support and guidance throughout the past three and a half years. Most notably my supervisor Trevor Bailey, for trusting in me, allowing me access to vast quantities of his time and expertise, and for keeping me focussed.

To Peter Durr and Mark Arnold at VLA for providing the data and modelling work that formed the basis of our viral load; all the staff at SECaM for making my time at Exeter so enjoyable; and to EPSRC/VLA (project grant reference CASE 0305) for providing the financial support necessary to complete this work. Also to everyone at CIDC, particularly James Wood, for trusting that I would *eventually* finish my thesis.

In addition I would like to thank the many people that have been complicit in keeping me (at least partially) sane over the years. In particular Sam and Emma, and also Ralph, Tim and Dave, for indulging my morbid sense of humour and helping me to put the world to rights (aided in no small part by the erudite surroundings of the Ram). To Raj for always reminding me that there is more to life than work - though making it easy for me to forget that at least some of it has to be; and also to Idayu Mahat, who though often busy with her own work never failed to find the time to aid me in mine.

I am indebted to Graham, Luke and Bruce for their computing support, and finally to Jez, without whose expertise James' trust would have probably been short-lived.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Progressive advances in medical science over the last few centuries have resulted in a massive decline in mortality from many curable and preventable diseases. This is particularly true in the developed world where previously endemic diseases such as smallpox have been almost completely eradicated. However costs associated with disease epidemics remain extraordinarily high, both in terms of the number of human lives lost and also due to resulting socio-economic impacts. Although medical science has identified and provided cures for many established diseases, the recent past has not only seen large numbers of fatalities from treatable and preventable diseases such as malaria, tuberculosis and cholera, but also the emergence of new deadly pathogens such as the Ebola, SARS (Severe Acute Respiratory Syndrome) and AIDS (Acquired Immune Deficiency Syndrome) viruses.

Of course the effects of disease epidemics on a population of individuals are not solely limited to human populations, nor are they exclusively caused by infectious human disease. Animal disease epidemics are also a major issue, the effects of which are wide-ranging across both human and animal populations. A disease such as Newcastle's disease for example will decimate a population of birds, most of which will die before clinical signs have appeared. The effects of the disease ranges between weight loss and diarrhoea to almost complete paralysis. The horrifying nature of the virus notwithstanding, clearly the ecological impact of an outbreak could be devastating to wild and domestic bird

populations; the latter also constituting a large economic cost with regards the poultry industry.

Furthermore, epizootic diseases such as rabies and avian influenza have the capacity to infect and be transmitted between both animals and humans. In these circumstances such a disease constitutes a major risk to human and animal health. This has been most recently illustrated by the outbreak of the H5N1 strain of avian influenza across Asia and Eastern Europe that has so far (October 2006) seen 256 confirmed cases in humans, but that current evidence suggests has not yet resulted in human-to-human transmission.

In general, diseases tend to belong to one of two main classes: infectious or non-infectious, but many of the most serious problems that arise are due to infectious diseases with their capacity for large-scale spread. Here particular reference will be given to animal disease epidemics (specifically foot-and-mouth disease or FMD).

Note that the terms *infectious* and *contagious* are used interchangeably in this discussion, although in some texts the term *infectious disease* relates to a disease that is caused by an invasion of the host organism by a foreign microbe (usually a virus or bacterium), and a *contagious disease* relates to a disease that is easily transferred between different host organisms. In this context the definition of an infectious disease will be based on that laid down by Bailey (1975), i.e. we are concerned with

> ...diseases that are infectious in the sense of being capable of transmission at some stage in the life-cycle of the appropriate organism from an infected host to an uninfected susceptible, with or without the agency of an intermediate insect or animal vector.

The nature of this type of disease means that there are many difficulties associated with disease management, and expertise is required across a wide range of different disciplines. For example diseases can be caused by many different types of microbe, primarily viruses and bacteria, each of which exhibits different characteristics regarding regeneration and spread. In addition diseases can often be contracted through different strains of the same

basic pathogen. Understanding the biological nature of the disease in question is important not only in researching potential cures but also in the development of robust modelling techniques that can aid understanding and guide the development of control policies to help prevent spread. Indeed the latter issue may depend on many external factors, for example environmental landscape, climate, migration or other movement patterns in the population at risk. Even if treatments and control techniques can be established, the additional financial and logistical constraints involved in producing and implementing such procedures exert a major influence on the overall efficacy of these measures. To overcome these complications each different facet must be dealt with - effective disease control policies rely on a combination of good science *and* good management.

Such complexities are, for example, particularly evident in the ongoing HIV/AIDS epidemic; which 2005 Global Health Council (GHC) estimates suggest affects upward of 40 million people worldwide. In this case even though anti-viral drugs have been developed that can be mass-produced and used to help control and slow the progress of the disease in individual sufferers, the number of infections continues to increase. There are many reasons for this, for example many of the countries worst affected are in the developing world where poor socio-economic conditions mean that the costs of producing and distributing these drugs put massive financial strains on governments and aid agencies. Ignorance of the method of transmission and/or refusal to admit the extent of the problem, perpetual states of conflict, the use of rape as a weapon of war and the sheer geographical scale over which the disease is prevalent are just a few examples of issues that are facing those working to get the epidemic under control.

Examples like this highlight the importance of having properly focussed research across a variety of different areas, each of which can help with different aspects in the development of control and treatment policies. Mathematical modelling offers a variety of different techniques that can help in many of these situations, and when combined with research in other areas of medical science, such as biology and chemistry, provides a useful repertoire of tools for the study of infectious diseases and their associated epidemiology. Advances in information technology have allowed better quality data to be collected and stored and

more complex mathematical models to be developed and fitted. Increased funding into epidemic research has resulted in this becoming a fast moving and exciting field of study.

Within this general framework animal disease epidemics offer their own peculiarities and difficulties. The Office International des Épizooties (OIE) was set up in 1924 by twenty-eight different states with a mandate to use scientific methodology to explore and understand animal disease. The aim was (and still is) to improve veterinary science and protect the welfare of animals, the safety of the food chain and to safeguard world trade in animal products. Until they changed their classification structure in 2006, they split infectious animal diseases into different lists based on their severity. List A contained the 15 most contagious diseases and list B the next 80. List A diseases were classified as:

> Transmissible diseases that have the potential for very serious and rapid spread, irrespective of national borders, that are of serious socio-economic or public health consequence and that are of major importance in the international trade of animals and animal products (OIE 2005).

Top of list A was FMD, which is estimated to have cost the British government upwards of £8 billion and resulted in the slaughter of 6 million (possibly as high as 10 million) animals, predominantly livestock, as a result of the 2001 UK epidemic.

FMD is therefore an extremely contagious and economically dangerous disease that once established is very hard to control. This is particularly true in the UK which holds a 'disease-free without vaccination' status, meaning that livestock have no conferred resistance to the disease (Follett et al. 2002). Other diseases that are similarly dangerous include classical swine fever, Newcastle's disease and avian influenza.

This thesis arises from a CASE studentship developed in collaboration with the Veterinary Laboratories Agency (VLA), Weybridge, in order to develop more sophisticated spatial survival models than had previously been used to model infectious animal diseases. This includes extensions to incorporate spatial heterogeneity and multiple sources of infection, and to explore the feasibility of using these models to predict the future course of an

epidemic. There is little in the literature regarding the use of survival analysis in modelling spatio-temporal spread of animal diseases, with perhaps the exception of Sanson and Morris (1994), who applied survival techniques to model the probability of infection in discrete time periods - with the spatial aspect reflected by premises being classified into two groups based on distance from the point source of infection. Survival modelling has various useful features, such as the ability to predict not only the risk of infection (hazard), but also the time to infection directly. Recently the use of random effects (see Oakes and Jeong 1998) in survival models (Li and Ryan 2002, Henderson et al. 2002) provides a useful starting point for the exploration of new techniques to detect clustering and spatio-temporal correlation structures in animal disease epidemic data. The models will be tested on a real data set provided by the VLA for the Devon subset of the 2001 UK FMD epidemic.

Chapter 2 will discuss ideas and particular issues affecting the modelling of animal disease epidemics in general and some of the biological aspects associated with FMD. A range of potential modelling techniques that have been applied to animal disease data will be discussed in chapter 3, along with their various advantages and disadvantages, with particular reference to survival modelling.

Chapter 4 gives more in-depth mathematical detail about survival modelling and its basic principals. In addition we discuss various alternative model formulations that can potentially deal with a range of different epidemiological issues. In chapter 5 some of these techniques are applied to data from the 2001 FMD epidemic in Devon, and a spatially and temporally varying covariate is derived that attempts to quantify the infectiousness of the disease through measuring the 'viral load' over space and time. This covariate is also used to censor the data set to help target surveillance and focus the model fit to a subset of the total population deemed 'at-risk' of infection at each point in time.

Chapter 6 extends the basic models of the previous chapter to incorporate 'resistance to infection' into the modelling framework. A range of candidate models are discussed along with their various advantages and disadvantages. In addition some simple simulations are

used to investigate their usefulness.

In chapter 7 we test two of the approaches discussed in chapter 6, those of the *long-term survivor* and *mixture* models, using both a simulated epidemic based on the dynamics of the FMD spread in Devon in 2001, and the real data set. In addition the predictive potential of these two approaches in both situations is explored. Chapter 8 gives some conclusions and further considerations in the survival modelling of both FMD and other contagious animal diseases, and highlights potential areas for future work.

Throughout this project the R statistical language (R Development Core Team 2005) has been used for all data analysis except the MCMC model fitting, which was achieved in WinBUGS (Spiegelhalter et al. 2003) using the R2WinBUGS (Sturtz et al. 2005) package. Examples of the R and WinBUGS code used to implement the techniques described in this thesis are provided on an attached CD-ROM.

Chapter 2

Infectious disease and FMD

This chapter is intended to be an introductory discussion concentrating on some particular matters associated with the aetiology and pathogenesis of infectious diseases, and considers some of the main generic differences between infectious animal and human disease. Particular reference is given to FMD. The focus is on biological and epidemiological issues rather than those associated with the mathematical modelling of epidemics, which will instead be discussed in the next chapter.

The importance of these considerations when designing mathematical models is not to be underestimated, and in order to be effective in understanding the aetiology of infectious diseases and their associated dynamics it is vital to consider contributions from a range of different sources. For example biological knowledge of the disease is essential in developing reasonable and intuitive mathematical models for modelling infectious processes. Likewise, the assumptions, requirements and aims used to build such models can greatly influence the choice and design of associated biological experiments. Effective epidemiological research relies on effective working relationships across a range of different scientific disciplines.

Although human disease can be thought of as a species subset of general animal disease, in this context the two will be considered separately, since there are quite marked differences between them, particularly in the development and implementation of control strategies.

2.1 General issues

Clearly the dynamics of an infectious disease vary at different levels. Diseases are caused by pathogenic microbes such as viruses, bacteria or fungi. Once inside the host organism a pathogen needs to be absorbed by cells in order to reproduce and replicate, both within the host cell itself and also in the eventual colonisation of other cells. Once this is achieved a method of excretion and transmission to other potential hosts is required. As an illustrative example of some of the complexities that must be examined at each stage, we consider tuberculosis (TB).

Tuberculosis is a disease that affects a variety of different species; common symptoms include: diarrhoea, vomiting, coughing, weight-loss and eventual death. It is curable but still causes approximately 5000 deaths per day (WHO 2004), predominantly in the developing world. In humans it is principally caused by a bacilli (a rod-shaped bacterium) known as *Mycobacterium tuberculosis*, itself part of a larger tuberculosis complex that includes multiple strains, such as *M. bovis*, *M. africanum*, *M. canetti* and *M. microti*, and also multiple serotypes within those strains.

Each strain causes TB under different circumstances (e.g. in different species or geographical locations), however the risk of infection from some strains is not exclusive to one species and each of the strains is thought to have developed from one common ancestor (Smith et al. 2006). For example *M. bovis* is the principal cause of tuberculosis in cattle but can also affect many other species of animal, including sheep, horses, deer, dogs as well as humans. *M. tuberculosis*, in contrast, is almost entirely limited to humans.

When an individual is exposed to the infectious agent, it is generally accepted that the bacilli are absorbed by *macrophage* cells, produced by the body as part of the immune response system. Their task is to absorb and digest foreign pathogens, though in the case of tuberculosis this can instead result in multiplication of the bacilli within macrophages, and lead to eventual distribution of the infected cells around the body to secondary infection sites were they form calcified lesions known as *granulomas* (or 'tubercles' in the case of

TB).

In many circumstances the bacilli can remain dormant within the cells for long periods of time (latent TB), only to become active under certain conditions (active TB). In humans, up to 90% of infections in healthy individuals are latent infections that the immune system manages to keep under control (WHO 2006). It is estimated that only 10% develop into active infections. Under certain circumstances - often relating to the strength of the immune system of the infected individual (HIV sufferers are particularly susceptible for example) and the virulence of the strain - the probability of a latent infection developing into an active one is much higher. The length of latent periods for TB can therefore vary dramatically between individuals.

After colonisation, infectious bacteria are then excreted from the body (through a variety of means - commonly in the faeces, milk, urine or via respiratory secretion) and transmitted to a new host. This can happen through direct contact, or through some other process, such as wind or water carriage for example (Ayele et al. 2004). Usually this secretion begins before clinical signs appear - which further compounds the issue of preventing spread.

In certain cases organisms can act as carriers of the disease without succumbing to infection themselves. An example of this is badgers, who act as potential carriers for *M. bovis* in cattle. This has been well documented in the UK press in recent months, with many farmers calling for a cull of the badger population in order to protect their herds. Many animal rights groups on the other hand, dispute whether a cull is necessary to reduce the spread of the disease, arguing that mass livestock movements are more likely to be the major cause of problem. It is clear that informed scientific research is just one facet required in order to develop effective disease management strategies.

The cyclical biological process described above is referred to as the life-cycle of the disease. It results in an individual host organism moving through a series of different states over time. It is this general principal that underpins most of the modelling work that has been done on infectious disease transmission. A simple example of a series of transmission

states for an individual is:

$$\text{susceptible} \rightarrow \text{exposed} \rightarrow \text{latent} \rightarrow \text{infectious} \rightarrow \text{recovery and/or death.}$$

Factors such as environmental conditions, the physical constitution of the host, the pathogenesis of the disease and the efficacy of drug treatments can all be important in defining the various groups and the length of time that an individual remains in each group. For example, as mentioned before the susceptibility of an individual to develop active TB is greater if they suffer from HIV/AIDS. Likewise, exposure to TB very much depends upon socio-economic conditions - it is predominantly a disease of poverty, where unhygienic conditions and a lack of sanitation lead to a higher prevalence of the pathogen. In the case of *M. bovis* in cattle, exposure time can vary greatly depending on the environmental conditions; some evidence suggesting that the bacilli can survive in soil for up to two years at a time (Ayele et al. 2004). The latent and infectious periods also vary greatly. Other factors include whether or not a disease will confer immunity after recovery (such as measles), or whether an infected individual will rejoin the susceptible group (e.g. gastroenteritis).

The efficacy of an epidemic control policy depends greatly on knowledge of how the disease progresses at each stage in the life-cycle and at each level. A good example of this is the development of drug treatments for infectious diseases, where a major issue, other than multiple strains, is the capacity of many pathogens to mutate. With TB this has lead to the occurrence of multiple drug-resistant strains of the pathogen that are immune to treatments that are ordinarily very effective in curing the disease. This can happen for a variety of reasons that are often linked with factors occurring at the administrative level of the immunisation strategy - inconsistent treatments, variable drug supplies and wrong diagnoses often result in the pathogen being given enough time to mutate and render subsequent treatments ineffective. This is a common cause-and-effect difficulty when developing and implementing large scale immunisation/eradication strategies for many infectious diseases, and highlights again the importance of effective collaboration between disciplines.

Hopefully we have given a brief illustration of the types of complications facing scientists and epidemiologists in the study of infectious diseases in both humans and animals. However there are a number of problems that are more prevalent in the modelling of animal epidemics that are of particular interest here, especially the implementation and consequences of control policies.

Firstly there are inter-species differences to account for, with some diseases species-specific (such as myxomatosis in rabbits and classical swine fever in pigs) whilst others affect a variety of species (such as foot-and-mouth disease or rabies for example). This compounds difficulties in controlling the populations at risk, particularly for a disease such as rabies that can affect wild and domesticated animals, and for which one of the symptoms of the disease is often uncharacteristic aggressiveness and a tendency to roam over large geographical areas.

The inability to control wild animal populations and their movement further exacerbates these problems. For example myxomatosis has become endemic in the wild rabbit population in the UK since its introduction 50 years ago. The virus that causes the disease has a high capability to mutate and can be passed through an intermediary host such as fleas or mosquitoes, or through direct contact with an infected rabbit. This makes isolation of the disease and eventual eradication extremely difficult.

The converse of this is that for diseases that only affect say, domesticated livestock, then control of population movement should theoretically be much easier. However advances in the ways in which meat and poultry are farmed and marketed mean that the ease of movement of livestock over large areas is much greater than it was a few years ago. Despite the potential to vastly restrict these movements, if response policies are not instigated quickly enough they can result, as the 2001 FMD epidemic showed, in a localised epidemic rapidly evolving into a global one.

Many domesticated animals such as livestock can be vaccinated from many diseases but particularly virulent strains of the microbe can still lead to mass infections. Excessive use of vaccines and antibiotics in itself can lead to mutated pathogens that become resistant

to the treatment, or result in animals that are capable of carrying the disease even though they are not themselves infected (carriers). On top of this, various trading laws exist hindering the sale of vaccinated meat on the open market. With FMD the European Union (EU) has in place legislation that prohibits the use of routine vaccination in order to retain its international trade status of being 'FMD-free without vaccination'. This balance is especially crucial when considering animals that will one day enter the food chain, as many altercations exist regarding the risks involved in eating potentially infected meat against those involved in eating vaccinated meat.

An additional risk is posed by *zoonotic* diseases. The Pan American Health Organisation (PAHO) defines zoonoses to be any communicable disease that is 'transmissible from vertebrate animals to man'. These are in addition to those diseases that are common to both humans and animals (Acha and Szyfres 2003). Particularly relevant cases of this type of disease include rabies, bovine tuberculosis (as previously discussed), bovine spongiform encephalopathy (BSE) with its human form of Creutzfeldt-Jacob disease (CJD), and the current epidemic of the H5N1 strain of avian influenza. The methodology used to model and control certain animal diseases is therefore intrinsically linked with that of associated human diseases.

2.2 Foot-and-Mouth Disease

The 2001 UK FMD epidemic was caused by the *Type O Pan Asia* strain of the virus; one of seven main strains. It was the first major outbreak of the disease in Britain since 1967. The virus itself is highly contagious, and given ideal conditions can survive for long periods outside of the host. It can be transmitted in many ways, commonly through direct (or indirect) contact with infected animals (Samuel and Knowles 2001) or through dispersion by an environmental factor such as wind (Ferguson et al. 2001a, Hugh-Jones and Wright 1970, Donaldson 1983).

Thompson et al. (2002) estimate the total cost of the outbreak on the UK economy to have been approximately 0.2% of the gross domestic product in 2001. However this does not reflect the costs to individual industries such as tourism and food, since expenditure that would previously have gone into these was simply directed elsewhere. It is estimated that the financial cost of the epidemic to the tourism industry was between £2.7 and £3.2 billion, with costs of approximately £3.1 billion to the food and agriculture industry. Much of the latter was covered by compensation from the government, but even then the remaining uncovered losses amounted to almost 20% (£355 million) of the total estimated income from farming in 2001.

The ability of the virus to mutate gives rise to limited cover regarding vaccination policies, and the economic cost of widespread vaccination for all strains of the virus would be huge. Britain currently holds a 'disease-free' status, which affects trading rights and influences FMD control policies (Follett et al. 2002), however countries that hold this status rely heavily on strict import regulations to stop entry of the virus (Samuel and Knowles 2001). Since routine vaccination is not implemented, once the disease gains a foothold it becomes very difficult to control.

The first confirmed case of FMD was an infected pig found on an abattoir in Essex on 20th February 2001, but it is thought that the index case for the outbreak was a farm in Northumberland and that the initial entry of the virus was through an infected food source. Dating the earliest lesions on infected animals suggested that the disease was certainly present on the index premise on the 12th February but could have been present as early as the 26th January (DEFRA 2002b). From here it is thought that the movement of pigs to the abattoir plus airborne movement of the disease to nearby premises led to the initial spread. A variety of factors, such as the delay in clinical signs appearing and frequent animal movements allowed the disease potentially up to a month to get established across a wide area before the initial response orders came into effect. The first set of control policies commenced on the 23rd February and resulted in the culling of infected premises (IPs) and dangerous contacts (DCs) with associated movement restrictions. Once the latter were introduced new infections seemed to primarily come from localised transmission (Ferguson

et al. 2001a).

On 23rd March contiguous premises (CPs) were included in the cull, just prior to the epidemic reaching its peak on 26th March (54 cases in one day). 3km ring culling was introduced in Cumbria on 27th March and on the 29th March the 24/48hr cull policy meant that IPs were culled within 24hrs and CPs and DCs within 48hrs of report (Keeling et al. 2001b). June 20th was the first day since the initial report where there were no reported infections. The last reported case of the disease was on 20th September.

The pathogenesis of FMD is extremely complex, and the infection dynamics at each of the within-host, within-herd and between-premise levels can vary significantly depending on a range of factors. The reader is referred to Alexandersen et al. (2003b) for a comprehensive paper on the pathogenesis and diagnosis of the disease, in which the authors draw on their own research and a host of other sources to provide an in-depth and detailed account of the biological nature of the disease and factors affecting reproduction and transmission of the virus.

The most common way in which the virus can enter a host organism is through airborne transmission, with the virus usually lodging somewhere in the respiratory tract. Other possible ways in which the disease can be transmitted include through oral routes (e.g. through eating contaminated food) or through direct contact with the skin or hooves. It has been shown that the risk of infection through these latter mechanisms is much smaller than through respiratory transmission (Donaldson 1987), although the transmission potential is greatly increased if there is damage present at the site of contact e.g. skin abrasions or mouth ulcers for example.

The transmission mechanism can also have a large effect on the within-host infection dynamics; affecting factors such as the length of the incubation period, the minimal infectious dose required to initiate infection, the rate of viral colonisation and the level of viral excretion. Experimental data (Sellers 1971, Donaldson 1987) has shown that animals are much less susceptible to infection from oral transmission than from airborne transmission.

Once the initial infection has occurred, the virus is transmitted around the body via the lymphatic and circulatory systems to secondary infection sites, predominantly in the skin and mouth. The virus is thought to replicate in the lymph nodes and then cause lesions to appear (most discernibly around the feet and mouth). The incubation period (i.e. the period between infection and the appearance of the first clinical signs) can vary between species and individual. It can range from anywhere between 1 to 14 days under certain conditions - though for within-herd spread it typically ranges between 2-6 days (Alexandersen et al. 2003a,b). In addition there is often a delay between the appearance of suspected clinical signs and actual (laboratory) confirmation of the disease. Typically an infected animal will have begun excreting the virus in the pre-clinical phase and will continue to excrete the virus until recovery (Menach et al. 2005). Though potentially fatal, the mortality rate for adult animals is very low, the result of infection usually being reduced weight gain and milk yield (Ferguson et al. 2001a). However it carries a high probability of fatality in young animals.

The excretion dynamics vary between species, but generally follow a pattern of high initial viral excretion, followed by a phase of reduced excretion in response to antibody production before recovery. For example, in contrast to their susceptibilities, experimental data has shown that once infected pigs excrete far more airborne virus than cattle (Alexandersen et al. 2003a), though in contrast they require much higher concentrations of airborne virus to become infected through this means (Donaldson and Alexandersen 2001, Donaldson et al. 1987). This is partly due to the fact that cattle are more susceptible to respiratory infection (e.g. the minimal infective dose required is much smaller) and partly due to cattle having a bigger lung capacity and inhaling a larger volume of air.

It is worth noting that some animals can remain carriers of the disease after recovery, and under certain environmental conditions the virus can remain active outside of a host for extended periods of time. The potential risk from carrier animals is also a problem when considering vaccination strategies, since it has been shown that some vaccinated animals can still act as carriers even though they don't contract the disease themselves.

Infection potential also changes with the particular strain and size of the virus particles - some strains of the virus (*C Novill*) are capable of causing spread up to 300km away (Gloster et al. 1981, Sorensen et al. 2000), though in the case of the type O strain it is unlikely to cause infections over a distance of greater than 20km (Alexandersen et al. 2003b). Even with the capability of virus carriage over such a distance other factors such as density (Hugh-Jones and Wright 1970) and type of animal (Keeling et al. 2001a, Donaldson et al. 2001), landscape fragmentation (Kao 2001), animal husbandry and welfare conditions, length of exposure, control orders and biosecurity (Ferguson et al. 2001a,b, Keeling et al. 2001a), human (and vehicle) interaction, wild animal movements and the movement of livestock around the UK also have a part to play. The latter point in particular is thought to have been a significant factor in explaining why the 2001 epidemic affected a much larger geographical area than the 1967-68 epidemic (DEFRA 2001).

Other unknown external factors that may affect the propensity of disease spread include temperature, rainfall and wind direction, and often knowledge of these agents is unknown or incomplete. This can make pre-emptive control strategies, in particular vaccination strategies, difficult to implement.

Chapter 3

Mathematical modelling of infectious diseases

The first recorded use of a mathematical model applied to a contagious disease epidemic was fitted to smallpox data by Bernoulli (1760). This approach was deterministic and based upon a series of differential equations, the core principals of which form the basis of many epidemic models used today (see Bailey 1975, Murray 2003). The literature regarding the application of mathematical modelling techniques to infectious epidemic situations is large. Reviews of the history of mathematical epidemiology can be found in Bailey (1975) and Anderson and May (1991).

A mathematical model is by its nature a representation of reality and not reality itself, and in order to develop realistic models for infectious diseases various assumptions must be made about the physical processes that drive epidemics. There is a wealth of methodology, from a range of different mathematical backgrounds that can be used to model epidemic data. In each case it is important that the creation and interpretation of any model is driven by sound mathematical and physical principals. In an epidemiological context this means that often the aetiology of the disease must play an important role in guiding model development.

To quote Murray (2003):

> From a *mathematical* [sic] point of view, the art of good modelling relies on: (i) a sound understanding and appreciation of the biological problem; (ii) a realistic mathematical representation of the important biological phenomena; (iii) finding useful solutions, preferably quantitative; and what is crucially important (iv) a biological interpretation of the mathematical results in terms of insights and predictions. The mathematics is dictated by the biology and not visa-versa.

This chapter will focus on introducing some of the main techniques that have been used in the modelling of infectious disease epidemics, including specific references to animal disease. In particular some more recent developments will be discussed. It is not intended to be a comprehensive account of the subject but rather to highlight similarities and differences between contrasting frameworks and their potential uses when modelling this type of data.

3.1 Compartmental models

Following the discussion in the previous chapter, the most common way to model epidemic data is to consider that at any time point a population of individuals can be classified into a series of groups based upon various stages in the life-cycle of the disease. A set of equations can then be developed that model the rates of transitions between the groups over time.

Consider the model proposed by Kermack and McKendrick (1927), which assumes that each member of the population belongs to one of three states: susceptible, infective or removed, and that individuals can only move between states in that order. The removed group consists of those individuals that have either recovered and been conferred immunity from the disease, have died, or have the disease but are no longer infective.

In a simple deterministic framework, models based on differential equations can be developed for modelling the rates of transitions between the groups. The simplest form of this model assumes that the mixing between the susceptible and infective groups is homogeneous - i.e. that each susceptible individual has an equal probability of coming into contact with each infective individual. The transition rates between groups can then be modelled as:

$$\begin{aligned} \frac{dS(t)}{dt} &= -aS(t)I(t), \\ \frac{dI(t)}{dt} &= aS(t)I(t) - bI(t), \\ \frac{dR(t)}{dt} &= bI(t), \end{aligned} \tag{3.1}$$

where $S(t)$, $I(t)$ and $R(t)$ are the numbers of susceptibles, infectives and removed individuals at time t respectively. The infection rate is given by $a > 0$ and the removal rate of infectives by $b > 0$. In this set-up the total population (N) is assumed constant and the framework ensures that $S(t) + I(t) + R(t) = N$ at each time point. This model also assumes that there is a negligible incubation period for the disease i.e. that an individual becomes infectious immediately after contracting the disease.

This framework may be limiting in reality, and many adaptations and extensions of this basic approach have been developed to deal with different assumptions regarding the disease dynamics. These include the incorporation of latent periods and exposure (SEIR models), recurrent susceptibility (SIS models - that is individuals that are not conferred immunity after recovery), temporary immunity, carriers and host-vectors, models for diseases that only affect subsets of the population (such as many venereal diseases) and heterogeneous mixing of populations. A more detailed introduction to all of these approaches can be found in Murray (2003).

From these models several important quantities can be readily obtained. One example is the epidemic curve, which gives the rate of new infections over time and is obtained by plotting $dI(t)/dt$ against t. Also a key interest when modelling infectious diseases is the ability to predict whether an initial small-scale outbreak will develop into an large-scale

epidemic. The *basic reproductive rate* of the disease, denoted R_0, is an important quantity in this respect. It is defined as,

$$R_0 = \frac{aS(0)}{b}, \tag{3.2}$$

where $S(0)$ is the number of initial susceptibles. It measures the number of secondary infections from each primary infection. A result of this is that if $R_0 > 1$ then an epidemic situation will occur. In order to control or prevent spread R_0 must be reduced to less than one. The idea of a threshold value that determines the ultimate global course of an outbreak is of central importance in the development of control strategies.

In order to gauge any useful information from a model, a method is required to estimate the values of the parameters from the observed data (model fitting). Once this has been done a simple epidemic model such as the one described by (3.1), in a closed population with known initial conditions, can be solved analytically. If not all of the parameters can be estimated from the data, then often the solution must be obtained numerically.

A limiting factor is the quality of the available data, since unreliable data can make the model fitting and parameter estimation difficult. However technological advances in the late twentieth-century have resulted in the capacity to collect more complex epidemiological data than was previously available; though often in practice numerical methods are still required to solve the system of equations, in certain cases (i.e. when the epidemic is small), approximations can be used (see Kermack and McKendrick 1927, Murray 2003).

In addition the system is often *non-dimensionalised*. This removes any dependence from physical units of measurement in the model, making relative inferences more meaningful. A discussion of this technique, including applications, can be found in Segel (1972) or Murray (2003). A key point to note is that the fitting mechanisms and model formulations described above attempt to provide an exact solution to the problem. In this sense the models are deterministic, however in reality this is rarely the case since there are often many other (unknown) factors that can produce variation in the data. The problem here is that although parameter estimates can be obtained they do not account for random variation in the data. One solution could be to increase the complexity of the models;

possibly incorporating some prior knowledge about the spreading mechanism - this will improve the parameter estimates but at the cost of making the model harder to fit.

An alternative is to use a stochastic framework. This regards the numbers of susceptibles, infectives and removed individuals as random variables and models the transmission rates between groups through probability distributions (see Bailey 1975). These have the advantage that they incorporate random variation into the model, which removes some of the rigid assumptions about the data required in the deterministic framework. Point estimates for the parameters can also be obtained, but also information about their variability. In addition predictions produced from stochastic models are often more informative, due to the inclusion of random variation in the generation of the predicted values.

The model formulation for the stochastic framework is slightly different to the deterministic approach, and so for illustrative purposes consider the case when there are only two groups: susceptibles and infectives. Here the total population size is still fixed to be N, but this time $S(t)$ and $I(t)$ are treated as random variables. In the deterministic case the number of new infective cases in a small time period Δt ($\Delta t \ll 1$) is given by $aS(t)I(t)\Delta t$. That is the contact rate, a, multiplied by the number of potential contacts and the length of the time period. In the stochastic case it is now the probability of infection that is measured in this way, i.e. the probability of infection in a time period of length Δt is $aS(t)I(t)\Delta t$.

Assuming that the mixing between the groups is homogeneous, the probability of s susceptibles remaining at time t is given by $p_s(t)$. After non-dimensionalising (by re-scaling time so that one unit equals at), the probability of s susceptibles remaining at time $t+\Delta t$ ($\Delta t \ll 1$) is:

$$p_s(t+\Delta t) = \{(s+1)(n-s-1)\Delta t\}p_{s+1}(t) + \{1 - s(n-s)\Delta t\}p_s(t). \qquad (3.3)$$

This follows from the fact that in order to have s susceptibles at $t+\Delta t$, there must have either been $s+1$ susceptibles at time t followed by one infection in $(t, t+\Delta t)$, or s susceptibles at t and no infections in $(t, t+\Delta t)$. This then gives a set of equations for dp_s/dt and dp_n/dt which can be solved subject to given initial conditions, though again

solving analytically in practice is often tricky, and as such numerical methods are usually used to overcome this.

The models described in this section so far have focussed on modelling the temporal spread of the disease but not the spatial spread. Using the simple SI model described above, Murray (2003) considers a spatially-varying compartmental model where the spatial aspect of the epidemic spread is modelled through a simple diffusion term. For a simple one-dimensional case this corresponds to:

$$\begin{aligned} \frac{\partial S(t)}{\partial t} &= -aS(t)I(t) + D\frac{d^2}{ds^2}S(t), \\ \frac{\partial I(t)}{\partial t} &= aS(t)I(t) - bI(t) + D\frac{d^2}{ds^2}I(t), \end{aligned} \tag{3.4}$$

where $S(s,t)$ and $I(s,t)$ are now distributed over space (s) and time (t). The parameters a and b are as before and the diffusion coefficient $D > 0$ controls the degree of spatial spread. Furthermore the model assumes that both the susceptible and infective populations are uniformly spread over space and gives a wavefront solution for the spatial advance of the disease over time. More complex spatial spreading mechanisms have been developed and applied extensively in the study of diseases such as rabies (e.g. Murray et al. 1986).

Turning attention to recent applications of compartmental models to animal disease epidemic studies: Haganaars et al. (2006) use a deterministic compartmental approach to model the spread of scrapie between sheep flocks in the UK; Cox et al. (2005) use a similar approach to model the spread of bovine tuberculosis (TB) in cattle and badgers; Stegeman et al. (1999) use a stochastic framework to model transmission of classical swine fever between herds in the 1997-1998 outbreak in The Netherlands and Meester et al. (2002) extend the model of Stegeman et al. (1999) to include temporal autocorrelation and prediction of the future path of the epidemic by using a discrete-time multitype branching process (Harris 1963, Athreya and Ney 1972). Applications to FMD in particular are discussed in more detail in section 3.4.

3.2 Cellular automata

An alternative approach to modelling epidemics through differential equations is to consider the use of cellular automata (CA). CAs discretise time and space and model the evolution of complex physical systems through local neighbourhood interactions (usually) based on a lattice structure.

In general each cell in the automaton is assigned to a specific state dependent on the application of the model. Transitions between states are governed by a set of rules relating to the state of the local neighbourhood surrounding each cell. In epidemic modelling a very simple example would be to represent space as a 2-dimensional regular lattice, where each cell represents an individual in the process, and takes the values 0 if susceptible or 1 if infected. At each time point t the cells are updated according to a function, $f(\cdot)$, that relates to the number of neighbouring infected cells at the previous time point.

In reality of course the aetiology of the disease can be used to define more realistic update functions and neighbourhood criteria. These can relate to temporal characteristics of the disease (e.g. latency periods or length of infection), covariates or spatial structure and lags in the definition of the local neighbourhood. Using prior knowledge of the spatial distribution of the individuals can also help to generate a more realistic model (e.g. inclusion of 'empty' cells). In addition stochastic extensions to CA models exist in which the transitions between states for the cells over time are governed by a probabilistic process rather than a deterministic one.

Algorithms for models based on CA often have a fast computation time due to the regularity in their structure (Sirakoulis et al. 2000). They can also produce many of the same quantities as the differential equation approach, for example the idea of a threshold for R_0 which determines the conditions under which an epidemic will ensue can be replicated using CA (Ahmed and Agiza 1998).

There are also various disadvantages with the differential equation approach that can be overcome using CA. As Ahmed and Agiza (1998) note, the former neglect the local

character of the spreading process which is modelled through localised interactions in CA. Variable total population sizes and susceptibility, external infections and complex initial and boundary conditions all cause computational problems in a differential equation setting, but can be dealt with relatively straightforwardly through the CA structure (see also Sirakoulis et al. 2000, Fuentes and Kuperman 1999).

The text by Mikler et al. (2005) gives a good introduction to the use of cellular automata in infectious epidemic modelling. It includes discussion on the formulation of neighbourhood structures and the inclusion of information relating to the pathogenesis of the disease (e.g. incubation/latency periods etc.) in the definition of the update functions. A particular issue with using CA in epidemic modelling is the problem of *neighbourhood saturation*, where the localised structure of traditional CA models can result in the rapid removal of the susceptible populations in the model. Mikler et al. (2005) deal with this problem by extending the local neighbourhood structure to a global one, in which all cells are included but are weighted according to a probability process based on intra-cell distance.

Authors that have considered the use of CA in the study of animal epidemiology include Fuks and Lawniczak (2001) who develop a model that can be fitted to generic epidemics in both humans and animals, and Doran and Laffan (2005) who fit a deterministic CA model to the spread of FMD in feral pigs and livestock in Australia. Morley and Chang (2004) apply a stochastic CA model to investigate the consequences of the British government policy in the 2001 UK FMD epidemic.

3.3 Statistical modelling approaches

When modelling any type of epidemic situation a compelling argument for the use of stochastic model formulations is that knowledge of many agents responsible for the dynamics of the outbreak are unknown or incomplete. Stochastic versions of compartmental and CA models work around the notion that mathematical laws governing physical processes are subject to random influence, however the field of statistical modelling is concerned

with modelling, accounting for, and reducing this random variation directly.

Various statistical modelling approaches have been used to model epidemic data; including simple time series, purely spatial models, space-time approaches and survival modelling. Typically in epidemic situations measurements are taken over a series of discrete time points. In this case the order in which the observations are recorded is of principal importance since consecutive observations are often dependent, with the degree of dependence across the time periods known as the *lag*.

A common way to model this type of data is through an ARMA (autoregressive moving average) framework (Box and Jenkins 1976), which models the values at each point in the time series through a combination of two independent processes; the first (*autoregressive*) treats the observed values as a weighted linear sum of their values at previous time points, and the second (*moving average*) corrects for the error in the previous forecasts through a weighted linear sum of past error terms. The number of components in each case is variable and is related to the temporal lag.

For an observed ordered time series Y_t, $t = 1, \ldots, T$, an ARMA model with 2 AR and 2 MA components takes the form:

$$Y_t = \mu + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \epsilon_t - \gamma_1 \epsilon_{t-1} - \gamma_2 \epsilon_{t-2}, \tag{3.5}$$

where μ is a constant intercept. The $\boldsymbol{\beta}$ parameters correspond to the effects of the autoregression and the $\boldsymbol{\gamma}$ parameters to the effects of the moving average. The stochastic part of the model is governed by the error terms ($\boldsymbol{\epsilon}$), which follow independent and identically distributed normal distributions with mean 0 and variance σ^2. The assumption of normality is key to the theory supporting the ARMA framework, though in certain cases transformations (e.g. Box-Cox transformation) can be used to correct when the data is non-normal.

Another key requirement is that the data is stationary (i.e. the probabilistic structure of the process Y_t is unaffected by a shift in the time origin - Diggle 1990) over time.

Non-stationary data can often be made stationary by differencing between successive time points, which attempts to remove the trend from the data. The degree of differencing required acts as an additional parameter in the model, now known as an ARIMA (autoregressive integrated moving average) model. Variance stabilisation transformations can also be used to correct for heteroscedascity (i.e. non-constant variance along the regression line).

Other issues include accounting for seasonal variation (localised trends), cyclical variation (trends over longer time periods) and irregular fluctuations (due to unknown factors). In addition, forecasting the future evolution of the time series involves extrapolating (i.e. predicting outside of the normal data range) the data set, which carries its own set of modelling difficulties.

For a good introductory text on the subject of time series modelling see Chatfield (2004). For more comprehensive accounts see Chatfield (2001), Box and Jenkins (1976) and Anderson (1971).

In epidemic situations, the ARMA/ARIMA framework can be limiting. Epidemic time series are often measurements on the *number* of infected individuals over time, and these are not normally distributed. A natural way to model this type of data would be through the use of generalised linear models (GLMs) (Nelder and Wedderburn 1972); allowing the simple linear model framework to be extended to include non-normal error terms.

In the simplest case the data are assumed to be binomially distributed; since each observation measures the number of 'successes' (infections) from a fixed number of 'trials' (total population of susceptibles). Let I_t, $t = 1, \ldots, T$, be an observed ordered series of the numbers of new infections at each time point i.e. $I_t \sim B(n_t, p_t)$ with the probability of infection, p_t, modelled through a *logistic* link function (see McCullach and Nelder 1989):

$$\log\left(\frac{p_t}{1-p_t}\right) = \eta_t, \quad t = 1, \ldots, T \text{ and } 0 \leqslant p_t \leqslant 1. \tag{3.6}$$

Here η_t is a linear combination of regression terms thought to directly affect the probability

of infection. (Note that other link functions such as the *probit* or *complementary log-log* links could be used instead of the logistic if required.)

If the probability of infection is small and the population 'at-risk' large, then a Poisson approximation to the binomial can be used i.e. $I_t \sim Po(\lambda_t)$, where now the regression terms are included through a *log* link function in the mean, $\lambda_t = n_t p_t = \eta_t$. Note that an advantage of these techniques is that all parameter estimates can be estimated simultaneously in the model framework.

So far this model contains no spatial structure or temporal correlation and is treating each observation as independent of the others. Spatio-temporal structure can be included in various ways, for example through an autoregressive process in the mean, space-time dependent covariates, or perhaps through the inclusion of random effects.

Alternatively, if collected at the individual level, epidemic data can be viewed as a spatio-temporal point process. Here the probabilistic phenomena of interest are the time and locations of infections (Diggle 2003, 2005). Typical aims involve locating clustering or regularity in recorded events over space and time, estimating and mapping relative risk of the event incidence or identifying clustering around a point source of infection. Techniques are developed around the notion that a completely random point pattern will follow a homogeneous Poisson process over both space and time.

A spatial analogy of the time series count model also exists if the data can be aggregated over space into a set of (regular or irregular) areal units. In this case a range of methods exist that allow patterns or trends in the relative risks of event incidence to be modelled (Lawson 2001). These techniques are widely used in spatial epidemiology and useful methods involve the capacity to model autocorrelation between measurements taken at different spatial lags. However accurate prediction is less important than identifying trends and patterns and their possible causes.

Good introductory texts on spatial analysis can be found in Bailey and Gatrell (1995) and Cressie (1993). Diggle (2003) provides a comprehensive account of spatial point pat-

terns analysis and Lawson (2001) applies areal modelling techniques to epidemiological data. For prediction purposes however, effective epidemic modelling involves accounting for the spread of the disease in both time and space, and as such the GLM framework with spatio-temporal extensions is useful.

The use of a generalised mixed model (Breslow and Clayton 1993) to model the underlying disease rates in different areal units for the incidence of insulin-dependent diabetes in military conscripts in Sardinia between 1936 and 1971 was proposed by Bernardinelli et al. (1995). They assume that the counts follow a Poisson distribution over time and space but that for scarce data the typical maximum likelihood formulation of the model does not account for the excess of random variation caused by the Poisson approximation to the binomial. Instead they formulate a mixed model in which the temporal trend and area specific intercept are modelled as random effects. They use a Bayesian specification since this can easily handle the complex model structure resulting from the inclusion of the random effects.

Knorr-Held and Besag (1998) note that the formulation of Bernardinelli et al. (1995) assumes the temporal trend to be linear. In order to relax this assumption they instead extend the dynamic model methodology (West and Harrison 1997), and allow for non-linear temporal trends and spatial variation (Besag et al. 1991) to be modelled non-parametrically. Knorr-Held (2000) extends this approach to include space-time interactions. By using a Bayesian framework they allow spatial and temporal autoregression to be included through the specification of the prior distributions.

Examples of these models applied to animal disease data include: Yoon et al. (2005), who fit a temporal Poisson regression model for the within-farm spread of avian influenza in the Republic of Korea in 2003-2004, and Berke (2005), who applies spatial relative risk mapping techniques to point data for pseudorabies in pig herds, and to count (area) data for small fox tapeworm infections in red foxes. Ducrot et al. (2005) aggregate a spatial point pattern in incidence of BSE in France in areas and test spatial clustering over time using the assumption that a random process should follow a Poisson distribution over

space.

Durr et al. (2005) fit four different Bayesian hierarchical models to count data for bovine fasciolosis in abattoirs in Victoria, Australia. They used logistic regression with spatially varying frailty effects to attempt to identify areas of high risk and possible environmental causes. Diggle et al. (2005) model bovine tuberculosis in Cornwall, UK using non-parametric kernel estimation techniques over discrete time points. Some specific applications to FMD are discussed in section 3.4.

3.3.1 Survival analysis

An alternative to modelling epidemic data through the numbers of infections is to consider modelling the time to infection directly through the use of survival modelling. Mathematical detail of these approaches will be given in chapter 4.

Survival modelling shares many of the advantages of the models described in section 3.3, in that all the parameters of interest (for spatial and temporal effects) can be estimated simultaneously and predictions of future survival times can be obtained for individuals.

The models are usually defined in terms of the *hazard* function. If the survival times are deemed to have come from a continuous distribution then the hazard represents the instantaneous rate of failure (in this case infection) at a point in time, given survival to that point. This is analogous to the transmission rate described in the discussion on compartmental models. The hazard function uniquely determines the distribution of the survival times (see Cox and Oakes 1984, Kalbfleisch and Prentice 2002) and allows many quantities of interest to be extracted - for example the probability that an individual becomes infected in the next week or the corresponding ranks of individuals most likely to become infected in the future. It is also straightforward to incorporate covariate information into the hazard function.

A key issue that affects the interpretation of data in survival analysis is that of *censoring*.

Censored observations are ones that contain incomplete survival information. These are often right-censored; typically individuals who have yet to experience failure at the end of the study period. However they still contribute important information about the underlying survival process that must be incorporated into the formulation of the likelihood. Survival analysis provides a tractable method for doing this, which weights censored values accordingly. There is also methodology to include left- and interval-censoring into models if required.

The specification of censored observations takes on additional importance in epidemic models, where exposure to the disease changes over space as well as time. In spatially varying epidemic situations, there is a reasonable argument to say that some individuals in the data are not representative of the population 'at-risk'; that is that due to their spatial location they are unlikely to ever be exposed to the virus in sufficient quantities to cause infection. Of course this is assuming a localised spreading process not linked to dangerous contacts. However if modelling local spread there is the danger that the information they contribute to the model fit will bias the parameter estimates, since the model will be fitted to an unrepresentative study group. Potential solutions to this problem, and the issue of non-local spread will be addressed chapter 5.

Similar arguments apply when incorporating the spatial aspect of a contagious epidemic into survival models as affected their inclusion into the GLM framework discussed earlier. There are two main facets: firstly there is the issue of modelling space-time dependence in the mean of the process (first-order effects), and secondly through localised 'stickiness' between neighbouring individuals (second-order effects). Space-time covariates are one way to handle the former problem, and the use of spatio-temporally correlated random effects (known as *frailties* in the survival literature) are one way to deal with the latter.

Survival models can be specified in a Bayesian framework, providing full posterior distributions for the parameter estimates and predictions of survival times. When combined with Markov chain Monte Carlo (MCMC) methodology it also provides a tractable method for fitting complex frailty models. In addition these models can provide information not only

on the localised spread of the disease but also, if the predictions are integrated over space and/or time, information on its global evolution.

Extensions to multivariate data and data aggregated by areas are also available (see e.g. Shimakura 2003, Henderson et al. 2002, Li and Ryan 2002), as are modifications that deal with issues such as immunity, multiple survival processes or multiple causes of infection. Some of these will be discussed in more detail in chapter 5.

3.4 Mathematical modelling of FMD

Before the 2001 UK outbreak the best data set available for FMD was from the UK epidemic of 1967-68. This has been studied by a number of authors; early papers including Henderson (1969), Hugh-Jones and Wright (1970), Hugh-Jones (1972), Sanson et al. (1991) and Sanson and Morris (1994). Recent studies have re-visited this epidemic, notably Sanson et al. (2000), Gerbier et al. (2002), Gloster et al. (2005) and Sellers (2006).

A wide variety of different model frameworks have been used to model FMD, though most of these have used the compartmental approach. The 2001 UK epidemic provides researchers with the most comprehensive infectious disease data set currently available (DEFRA 2004), and gives modellers much more information to use when designing mathematical models to capture various features of the spatio-temporal spread. Aspects of this data set will be explored in the next chapter, focussing in particular on data from the Devon sub-epidemic.

Beforehand, it seems sensible to discuss in some more detail the various modelling approaches that have been applied to FMD, in particular to those fitted to the 2001 data set, since they provide a good example of the range of models that can be used and the advantages and disadvantages when modelling different aspects of the epidemic process.

Various papers were published during the 2001 UK epidemic that used mathematical models to evaluate and advise on the potential effects of different control policies on controlling

the spread of the disease. Ferguson et al. (2001a) was published initially online in April 2001, barely two months after the beginning of the outbreak, and examined the potential impact of both movement restrictions and other different control measures on the future path of the disease. The analysis consisted of two parts: firstly they used individual level contact tracing data provided by MAFF (Ministry of Agriculture, Fisheries and Food - the predecessor to DEFRA) to parameterise a density function for the distance between infected premises and likely subsequent infected premises, and secondly to feed this information into a deterministic compartmental model to determine the spatio-temporal dynamics of the disease.

The spatial density function comprised a mixture of two terms: the first representing a localised spreading process acting uniformly over the local neighbourhood surrounding each IP, and the second a kernel function that weights contributions from connected premises based on distance. The parameters were estimated by fitting the density to the distribution of known infectious contacts.

The temporal aspect of their approach depended on modelling the distribution of times between three main events: report, confirmation and culling. These were estimated from the observed data and combined with the spatial kernel function within a differential equation model consisting of five groups: susceptible, infected but not infectious, infectious but not reported, infectious and reported and culled. Simulations were then used to assess the effects of different control policies on the course of the global epidemic (note that this included some 45,000 premises in areas that were currently infected - the actual potential susceptible number of premises in the UK exceeds 130,000 premises in total). The main conclusions were that ring cull or vaccination strategies would be vital in containing the epidemic, and that rapid slaughter of infected premises would help to slow the progress of the disease.

Ferguson et al. (2001b), published in October 2001 extended the approach of Ferguson et al. (2001a) to incorporate a time-dependent transmission coefficient, variable susceptibility and variable infectiousness. The contact tracing data allowed estimates of the spatial

scale of disease transmission to be obtained - identifying both long-range and short-range (localised) infections. Garner and Lack (1995) use a similar state-transition framework to simulate potential outbreaks of the disease in Australia, however their model is non-spatial. Durand and Mahul (2000) extend the models in Garner and Lack (1995) to include more classes and within-herd spread and apply it to simulated outbreaks in France.

Another key paper, also published in October 2001, was that of Keeling et al. (2001a). Here the authors used a stochastic rather than deterministic compartmental model. They formed an individual farm level model for the probability of infection of an uninfected premise based on proximity to nearby IPs. They included variable susceptibility and transmission based on the numbers and species of animal present on susceptible and infected premises respectively. So for a susceptible individual i at time t, the probability of infection in a given day was modelled by:

$$p_i = 1 - \exp\left[-\boldsymbol{S}\boldsymbol{N}_i \sum_{j \in \mathcal{I}_t} \boldsymbol{T}\boldsymbol{N}_j k(d_{ij})\right], \tag{3.7}$$

where $\boldsymbol{S}$ is a susceptibility vector based on species of animal (sheep or cattle), $\boldsymbol{T}$ is a transmission vector also based on species of animal, $\mathcal{I}_t$ is the set of all infectious premises at time t, $\boldsymbol{N}_i$ is a vector of the numbers of animals of different species on premise i, and $k(\cdot)$ is a kernel weighting function based on distance d_{ij} between premises i and j. The specification of the kernel function was estimated using contact tracing data, and the other parameters by using a maximum likelihood approach; forming the likelihood function from the product over the individuals of the probabilities of susceptible premises remaining susceptible and infected premises becoming infected at each day, in a similar way to the method described in section 3.1.

They then used simulations to measure the impact of culling and vaccination strategies. They identified heterogeneities in transmission intensities resulting from different numbers and different types of animal and their results indicate that the models were not able to reproduce the national epidemic as well when these species differences were excluded than they did when included. They also concluded that the rapid implementation of control

policies such as ring culling was essential in controlling the spread of the disease, but noted that the size of the neighbourhood would be situation and disease specific.

As a result of the Royal Society report on *Infectious Diseases In Livestock* (Follett et al. 2002) and the Lessons Learned Inquiry (DEFRA 2002a), there have been various concerns raised about the necessity of aggressive cull policies in controlling the spread of the disease, motivating researchers to explore the effects of alternative strategies. Tildesley et al. (2006) investigate optimal vaccination strategies for the control of FMD, and their results suggest that an optimal reactive ring cull strategy of 35,000 animals per day is more effective than a policy of CP culling in reducing the number of farms lost. However in a letter to the *Veterinary Record*, Wingfield et al. (2006) question these conclusions, arguing the case that CP culling, or a 'stamping-out strategy', should remain the primary method for control in the future. It is clear that there is still much debate about this issue. In a comment on Wingfield et al. (2006), Keeling et al. (2006) reiterate the importance of combining efforts between veterinary knowledge on the ground and mathematical prediction models when designing and instigating effective control orders.

The models discussed so far have been fitted globally, to data from the whole of the UK in 2001. Gerbier et al. (2002) used a similar approach to Keeling et al. (2001a), but instead applied their model to data from the 1967-68 UK epidemic. A slight difference was that they assumed a point process model in which the spatial spread of the disease was given a probability distribution and the parameters were estimated as part of the model. They model the infectious potential of a premise as:

$$\phi_i(t) = \beta_{1t} + \beta_{2t} \sum_j f(d_{ij}) + \cdots + \beta_{kt} z_k, \tag{3.8}$$

where β_{1t} represents a baseline probability of infection which is uniformly distributed over space and β_{2t} to the effect of localised spreading, based on a local spatial decay function specified by $f(\cdot)$ and governed by the distance, d_{ij}, between two premises (in this case a simple inverse distance decay function was used). Other factors z_k can also be included, and the authors use a function of the number of animals between the susceptible

and previously infected farms, weighted by distance. They then model the probability of infection through a logistic link function to the infectious potential, $\phi_i(t)$, and fit the model using maximum likelihood in a similar manner to Keeling et al. (2001a). Other papers that use stochastic compartmental approaches include Menach et al. (2005) and Chowell et al. (2006).

Diggle (2005) develops a partial likelihood approach to fit a point process model based on that of Keeling et al. (2001a). In the paper he models the conditional rate of transmission between farms i and j, λ_{ji}, based on the complete history of the process up to time t. The problem is that resulting log-likelihood is often intractable. Instead he proposes a partial likelihood approach to fitting the model that is an extension of the method proposed by Cox (1972) for fitting survival models. This latter method will be discussed in more detail in section 4.3.2.

The spatial spread is modelled through a transmission kernel with an exponentially decaying part representing localised spread and an extra parameter allowing for long-range transmission. The model is:

$$\lambda_{ji}(t) = \lambda_0(t) A_j B_i f(d_{ji}) I_{ji}(t), \tag{3.9}$$

where λ_0 is a baseline hazard function over time (see chapter 4), A_i measures the relative infectiousness of premise i and B_i the relative susceptibility. These are based on the numbers of cows and sheep on each premise, and the relative infectiousness or susceptibility of each species respectively. The model was fitted to data for the Cumbria sub-epidemic in the UK in 2001. The paper then discusses possible extensions to the model, for example the inclusion of extra farm level covariates in the susceptibility and transmission parameters, though as with any semi-parametric approach prediction is an issue.

This framework uses aspects of survival analysis in the modelling setup. Generally however, the use of survival modelling for infectious disease epidemic data is limited. Sanson and Morris (1994) consider the use of survival analysis to model spatial spread of FMD from one point source of infection in the 1967-68 UK FMD epidemic. Here they divided

the study region into a series of grid squares and treated certain grid squares as farm premises. This was based on overlaying Ordnance Survey maps of the area and assigning premise status to any grid square that did not contain a confounding geographical feature (such as rivers or woods for example). The size of the grid squares was taken as the median farm size from the 1965 census data. The model used distance from the source farm to calculate the probability of survival for each premise between each given time period.

This is a very simple implementation of survival modelling to epidemic data. A different approach was adopted by Lawson and Zhou (2005) to model spatio-temporal spread in Cumbria in 2001. They fitted a descriptive Binomial model to counts of infections over time in aggregated spatial units (parishes). They also developed a series of farm level *marked survival process* models in which the incidence of the disease, λ_i, on farm i is conditional on the survival time d_i. The survival time was modelled through a Weibull distribution (see chapter 4) and was included as a covariate in the intensity. They also used a range of other covariates in both the intensity and in the scale parameter of the survival distribution.

The spread of FMD at the within-herd level has also been examined. Streftaris and Gibson (2004a,b) consider using stochastic compartmental models in a Bayesian framework to model the dynamics of infectious diseases at this level; the latter paper with application to experimental FMD data. This is a similar approach to the one adopted by Keeling et al. (2001a) but with no spatial dependence structure (deemed unnecessary at this geographical scale). An approach centred on modelling transmission probability was used by Arnold (2005), and this particular method will be discussed in chapter 5.

CA models have also been used by a number of authors, including Doran and Laffan (2005), who use a stochastic CA model on foot-and-mouth disease in feral pigs and livestock in Australia, and Morley and Chang (2004) who use CA to investigate the consequences of the control policies implemented in the 2001 UK FMD epidemic. An alternative approach was used by Wilesmith et al. (2003), who investigated the use of space-time K-functions (Diggle et al. 1995) to explain spatio-temporal interactions in the risk of infection of FMD across

two counties of the UK (Cumbria and Devon) in the 2001 epidemic. These latter functions are used to measure clustering in point pattern data based on inter-event distances. They have the advantage that they can be applied across different spatial and temporal scales, but are predominantly used for exploratory purposes and cannot be used to directly predict future incidence.

3.5 Conclusions

This chapter has explored different frameworks for modelling contagious epidemics, each having its own advantages and disadvantages when dealing with particular aspects of the disease process.

Most of the work done on FMD so far (particularly for the 2001 epidemic) has focussed on estimating the effects of control policies on the global scale, however much less has been done on spatial and temporal prediction on a smaller scale. A problem when modelling large-scale epidemics is that global assumptions are made about the epidemic process, and it has been shown that often these assumptions will vary depending on factors such as location, topography, climate, density of individuals and changes in biosecurity for example. By modelling spatio-temporal spread over smaller areas these differences can potentially be accounted for. Also this raises the question as to whether optimal control strategies could be targeted to particular areas or individuals deemed 'high-risk' by the model, having the potential to greatly reduce the economic and welfare costs involved in disease management. Of course this may come with a greater computational burden, since more analyses must be undertaken, however each individual model will be much smaller and so this trade-off may be reasonable.

In the literature the use of survival modelling in these situations is rare compared to the compartmental framework, yet the survival approach has the capacity to incorporate many of the features of these approaches. It has the potential to deal with aspects such as changes in the state of the disease over time, censoring, immunity and multiple sources of

infection, as well as being able to investigate the effects of explanatory variables directly on survival time. It provides a tractable method to obtain predictions for the future course of the epidemic in both time and space, both on a local and global level. The ability to directly predict, not only the risk but also the time at which a premise may become infected is another attractive feature.

Statistical approaches also model random variation directly, and such a stochastic framework is biologically more reasonable than assuming a fixed deterministic process driving the epidemic. The GLM framework could be used but it tends to average the temporal aspect over time across individuals, whereas the survival framework has the capacity to incorporate explicit information on the shape of the epidemic curves over time. This thesis intends to explore the use of spatial extensions to survival modelling to develop space-time predictions for the path of a contagious epidemic.

Chapter 4

Survival Modelling

Survival modelling is used to model the time from the start of follow-up of an individual until some pre-defined event occurs. Typically this event is associated with failure of some kind, for example the death of a cancer patient in a clinical trial or the failure of a machine part during a safety test. As such the time to the event is usually referred to as *survival* or *failure* time.

From a statistical modelling perspective the survival times tend to follow some form of skewed distribution such as the exponential or Weibull for example. Additionally there is the issue of *censoring* i.e. those observations that have not been observed to fail during the study period. These observations still hold important survival information since they record the time over which an individual was 'at-risk' but did not succumb to failure. To prevent bias the modelling framework must be adapted to incorporate these observations in some way.

There are various ways in which these issues can be addressed. The generalised linear modelling (GLM) framework (Nelder and Wedderburn 1972) offers extensions to the classical linear model that allows non-normally distributed error terms to be used. Traditional survival models can be viewed as an extension of GLMs that incorporate censoring (Aitken and Clayton 1980, Whitehead 1980, McCullach and Nelder 1989) and may be fitted using

standard techniques such as iterative re-weighted least squares. Since many of the survival techniques currently used were developed independently of the GLM framework (see e.g. Kalbfleisch and Prentice 2002, Collett 2003, Therneau and Grambsch 2000), there exist a range of alternative fitting mechanisms, such as direct use of maximum likelihood, or in the case of a Bayesian model by an iterative sampling algorithm such as Markov chain Monte Carlo (MCMC).

The basic mathematical background to survival analysis will be discussed in section 4.1, and issues associated with censoring in section 4.2. A range of different modelling strategies exist that depend on the choice of distributional assumptions made about the data. Some of these will be discussed in section 4.3, including non-parametric, semi-parametric and fully parametric approaches and their potential uses when exploring the relationship between explanatory variables and survival time. If a parametric form is chosen, then predictive estimates of future failure times can also be obtained.

Section 4.4 explores Bayesian methodology and fitting mechanisms, principally Markov chain Monte Carlo (MCMC), and includes some discussion on Metropolis-Hastings, Gibbs sampling and Bayesian identifiability.

Finally, due to the complexity and range of different situations to which the application of survival techniques may be appropriate, various extensions to the conventional model have also been developed. Some common examples that may be applicable to disease modelling include mixture models, competing risks models, long-term survivor models, state-space models and change point models, and these will be covered in section 4.5.

Survival analysis is widely documented not only in the statistical literature but also in fields as diverse as engineering, social science and epidemiology. Since it has such a broad scope this section will focus on some of the basic concepts and techniques that are most relevant to epidemiological modelling. It is by no means an exhaustive account of the subject, and for a more comprehensive discussion on the ideas and methodology considered here the reader is referred to excellent texts by Kalbfleisch and Prentice (2002), Collett (2003), Therneau and Grambsch (2000), Klein and Moeschberger (1997), Cox and Oakes (1984)

or Lee and Wang (2003). The application of these techniques to the problem at hand will be discussed in later chapters.

4.1 Basic ideas

In general, survival techniques can be applied to a wide range of different situations, subject to three necessary requirements as stated by Cox and Oakes (1984): firstly a well-defined time origin must be determined, then the scale for measuring the progress of time must be decided upon, and finally the exact definition of failure must be clear.

To begin this discussion, consider first the case for homogeneous data where T is a positive random variable representing failure time. The *survivor* function, $S(t)$, is defined for both discrete and continuous distributions as the probability that an individual survives beyond time t, i.e.

$$S(t) = P(T \geqslant t) \qquad 0 < t < \infty. \tag{4.1}$$

Here $0 < S(t) \leqslant 1$ since $S(0) = 1$ and $\lim_{t\to\infty} S(t) = 0$. The distribution of T can be uniquely determined by the survivor function, or, as is commonly the case, by one of two other related quantities: the *hazard* function or the probability density function.

For a continuous random variable T, the density function, $f(t)$, is given by

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}, \tag{4.2}$$

where the (cumulative) distribution function $F(t) = P(T < t) = 1 - S(t)$, so that $S(t) = \int_t^\infty f(u)du$. The hazard function, $h(t)$, is defined as the instantaneous potential of failure at time t, given survival to t, i.e.

$$h(t) = \lim_{\Delta t\to 0} \frac{P(t \leqslant T < t + \Delta t \mid T \geqslant t)}{\Delta t} \quad \Delta t \ll 1. \tag{4.3}$$

This is a positive measure and is sometimes referred to as the *conditional* or *time-specific*

failure rate.

Following the fundamental theorem of calculus, it can be seen that (4.2) can be written as:

$$\begin{aligned} f(t) = \frac{dF(t)}{dt} &= \lim_{\Delta t \to 0} \frac{F(t+\Delta t) - F(t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(T < t+\Delta t) - P(T < t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(t \leqslant T < t+\Delta t)}{\Delta t}. \end{aligned} \tag{4.4}$$

Using (4.4) and the definition of conditional probability, the hazard (4.3) can be written as:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t \leqslant T < t+\Delta t)}{\Delta t[P(T \geqslant t)]} \\ &= \frac{f(t)}{S(t)}, \end{aligned} \tag{4.5}$$

and from (4.2) it follows that

$$\begin{aligned} h(t) &= -\left[\frac{\frac{dS(t)}{dt}}{S(t)}\right] \\ \Rightarrow S(t) &= \exp(-\int_0^t h(u)du). \end{aligned} \tag{4.6}$$

The quantity $H(t) = \int_0^t h(u)du$ is known as the *cumulative hazard* function.

If T is a discrete random variable then the probability function $f(t) = P(T = t)$ determines the exact probability of failure at time t. Likewise, the hazard function, $h(t)$, is the conditional probability of failure at time t given survival to t, i.e.

$$h(t) = P(T = t \mid T \geqslant t) = \frac{P(T = t)}{P(T \geqslant t)} = \frac{P(T = t)}{\sum_{j|t_j \geqslant t} P(T = t_j)}. \tag{4.7}$$

It is straightforward to define $P(T = t)$ and $P(T \geqslant t)$ in terms of the hazard function by considering that $1 - h(t)$ is the conditional probability of *survival* at t given survival to t.

So for ordered survival times $t_1 < \cdots < t_n$,

$$P(T = t_i) = h(t_i) \prod_{j=1}^{i-1} (1 - h(t_j)) \tag{4.8}$$

and

$$P(T \geqslant t) = \prod_{j|t_j \leqslant t} (1 - h(t_j)). \tag{4.9}$$

It is also possible to mix continuous and discrete distributional forms in one framework if required (see Kalbfleisch and Prentice 2002, chapter 1).

4.2 Censoring

The most common form of censoring, *right-censoring*, occurs when an individual joins a study at the beginning of the study period but does not experience failure. This happens generally either because the study period ends before failure, or because they were lost to follow up or withdraw from the study for some reason. If the time origin is taken to be at time 0, then an individual right-censored at time t is known to have survived the period $[0, t)$.

Left-censoring occurs when an individual fails at some point prior to time t but the exact failure time is unknown. In this case it is known that the individual failed in the period $[0, t)$. If the exact failure time for an individual is unknown, but is known to lie between two points a and b, where $0 < a < b < t$, then the individual is said to be *interval-censored*.

Also important when modelling incomplete (censored) survival data is the type of censoring mechanism being used. In the simplest case this is random and independent of the failure process - known as type III censoring (Lee and Wang 2003). Type I censoring occurs when observations are censored after a pre-defined length of time, and type II censoring when individuals are censored after a pre-determined number of failures have been observed. The type of censoring mechanism affects the form of the likelihood. Kalbfleisch and Prentice (2002) discuss modifications to incorporate type I and II censoring and deal with situations

where the censoring mechanism is not independent of the failure mechanism.

In many cases the assumption of random censoring is reasonable, however it is worth noting that there are certain circumstances in real-life epidemic modelling where this becomes invalid - principally when dealing with individuals who are removed from the study when considered at high- or low-risk of infection (as in the case of culled premises during the 2001 UK FMD epidemic). This issue will be discussed further in chapter 5 - here it will be assumed that all censored observations are randomly censored unless otherwise stated.

4.3 Modelling techniques

Traditionally full parametric forms were assumed for the distribution of survival times, and although this approach has a number of advantages (that will be reviewed in section 4.3.3), it seems sensible initially to discuss alternative approaches to survival model formulation in a wider context. Oakes (2001) cites two key papers responsible for extending the boundaries of survival methodology beyond those offered purely by parametric frameworks. The more recent of the two, Cox's seminal paper on proportional hazards (Cox 1972) will be discussed in section 4.3.2, whilst section 4.3.1 will focus on some non-parametric techniques, most notably that of Kaplan and Meier (1958) and the use of their product-limit estimator to estimate the survivor function for censored data.

4.3.1 Non-parametric methods for estimating the survivor function

Non-parametric techniques are useful in particular for exploratory analysis of survival data, since they are not restricted by the assumption that the data must follow a particular distributional form. Estimates and comparisons of survivor (and hazard) functions can be readily obtained, as well as corresponding summary values such as the mean, median, quartiles and confidence intervals.

In the case where there are no censored observations, the *empirical survivor function* can

be used to estimate the survivor function at a time t. This states that the probability of survival beyond a point t is the proportion of the total number of individuals in the study still alive after t, and is given by:

$$\tilde{S}(t) = \frac{\text{No. of individuals with survival times} > t}{\text{No. of individuals in data set}}. \tag{4.10}$$

If the data contains censored observations then the empirical survivor function in (4.10) is no longer valid. In this case there are various alternative techniques that can be used that work by dividing the study period into a set of discrete time intervals. The survival estimates are then based on the proportions of the total number of individuals deemed 'at-risk' in each interval.

Examples of some of these types of approaches, such as the *Actuarial* estimator, *Nelson-Aalen* estimator and the *Kaplan-Meier* estimator can be found in more detail in Collett (2003). The most well-known of these, the Kaplan-Meier (KM) or product-limit estimator, was first developed in Kaplan and Meier (1958), and for illustrative purposes only this framework will be discussed here.

In a sample of n individuals, consider initially just those that experienced failure. Adopting the convention that the failure time is taken to occur at the beginning of each interval, then a series of time intervals are formed such that each contains just one failure time. If there are $r \leqslant n$ failures let $t_{(j)},\ j = 1, \ldots, r$, be the ordered failure times such that the first interval $[t_{(0)}, t_{(1)})$ contains no failure time (i.e. $t_{(0)}$ is the time origin). In the case of tied observations censoring is taken to occur after failure.

Denote the number in the *risk set* just prior to $t_{(j)}$ as n_j and the number of failures at $t_{(j)}$ as d_j. Assuming failures are independent then an estimate of the probability of survival between $t_{(j)}$ and $t_{(j+1)}$ can be given by $\frac{n_j - d_j}{n_j}$, with the corresponding survival estimate for $t_{(j)} \leqslant t < t_{(j+1)}$ given by

$$\hat{S}(t) = \prod_{k=1}^{j} \left(\frac{n_j - d_j}{n_j} \right), \tag{4.11}$$

i.e. the probability of surviving through $t_{(j)}$ to $t_{(j+1)}$ and all the preceding intervals. This is known as the Kaplan-Meier estimate of the survivor function.

It can be seen that (4.11) returns a decreasing step-function with $\hat{S}(0) = 1$ and $\hat{S}(t)$ constant over each discrete time interval $t(j) \leqslant t < t_{(j+1)}$, $j = 0, \ldots, r$, where $t_{(r+1)} = \infty$. From this a range of useful quantities can be extracted such as the median, mean, quartiles and associated standard errors (using e.g. Greenwood's formula) and confidence intervals for the survivor estimate, as well as equivalent estimates for the hazard and cumulative hazard functions. Plots of the estimated survival and hazard curves against time can provide useful inferences into the underlying form of the survival distribution. Tests also exist to compare survival curves from different groups, such as the *log-rank* and *Wilcoxon* tests. Collett (2003) provides a detailed and lucid account of the derivation and application of these various quantities. See also Lee and Wang (2003).

4.3.2 Semi-parametric models

The types of non-parametric methods discussed in section 4.3.1 provide useful and tractable ways of estimating and comparing survivor and hazard functions for survival data, including extensions to incorporate censored information. However a key focus in survival modelling is to investigate the effect of covariates on survival time, and a different approach is required to do this.

As shown in section 4.1 survival models can be specified in terms of the survivor, hazard or density function. Since each of these will uniquely determine the corresponding survival distribution, Cox (1972) proposed specifying a model through the hazard function in such a way that for an individual with a vector of covariates $\boldsymbol{x}$, the hazard at time t is made up from two parts: the first modelling the hazard in the absence of covariate information (the *baseline* hazard function), and the second a (usually) parametric function representing the effect of covariates on failure time, over and above the baseline hazard (see Cox and Oakes 1984, chapter 5).

Cox (1972) first introduced his proportional hazards approach as a way to incorporate covariate information into a survival model without having to assume an underlying distributional form for the data. The model is defined in terms of the hazard function as:

$$h(t, \boldsymbol{x}) = h_0(t)\psi(\boldsymbol{\beta}; \boldsymbol{x}), \tag{4.12}$$

where $\boldsymbol{x}$ is a m-vector of explanatory variables, $\psi(\cdot)$ is a parametric function of $\boldsymbol{x}$ and $h_0(t)$ is the unspecified baseline hazard function (i.e. when $\boldsymbol{x} = \mathbf{0}$). Here $\boldsymbol{\beta}$ is a m-vector of parameters. A common way of specifying $\psi(\cdot)$ is to use a log-link to the covariates i.e. $\psi(\boldsymbol{\beta}; \boldsymbol{x}) = \exp(\boldsymbol{\beta}^T \boldsymbol{x})$. Model (4.12) is referred to as *semi-parametric*, since the baseline hazard function is left arbitrary.

There are many reasons for its popularity - Cox and Oakes (1984) offer various arguments. With regard to the model formulation, the idea that the effect of a covariate is to multiply the hazard by a constant factor is not unreasonable, and they argue a weight of empirical evidence in some fields supporting this. Also, censoring and the occurrence of several types of failure can be easily included in the model and furthermore adaptations to the fitting mechanism are straightforward in these cases even though the underlying survival distribution is left unspecified.

In order to fit the proportional hazards (PH) model (4.12), Cox (1972) developed a *partial likelihood* approach, so-called because it does not make use of actual censored and uncensored survival times. Consider n individuals with $r \leqslant n$ ordered failure times $t_{(j)}$, $j = 1, \ldots, r$. In a standard formulation, an (uncensored) individual i with failure time t_i and covariate vector $\boldsymbol{x}_i$ contributes $f(t_i, \boldsymbol{x}_i)$ to the likelihood; however since the form of $f(\cdot)$ is unknown, an alternative likelihood is derived using the conditional probability that an individual i fails at $t_{(j)}$ given survival to $t_{(j)}$, and the additional notion of 'risk-sets'.

The technique works on the assumption that intervals between successive failure times cannot contribute any information to the likelihood, since conceptually $h_0(\cdot)$ in those intervals could be zero. The likelihood is then constructed on the basis of information given by individuals across the whole set of observed failure times.

Using the rules of conditional probability and the fact that the failure times are assumed to be independent of each other, the following statement holds:

$$\begin{aligned} &P(\text{individual } i \text{ fails at } t_{(j)} \mid \text{one failure at } t_{(j)}) \\ &= \frac{P(\text{individual } i \text{ fails at } t_{(j)} \text{ and no one else fails})}{P(\text{one failure at } t_{(j)})} \\ &= \frac{P(\text{individual } i \text{ fails at } t_{(j)} \text{ and no one else fails})}{\sum_{k \in R(t_{(j)})} P(\text{individual } k \text{ fails at } t_{(j)} \text{ and no one else fails})}. \end{aligned} \tag{4.13}$$

However (4.13) can be thought of as the limit as $\Delta t \to 0$ of

$$\frac{P(\text{individual } i \text{ fails in } [t_{(j)}, t_{(j)} + \Delta t))/\Delta t}{\sum_{k \in R(t_{(j)})} P(\text{individual } k \text{ fails in } [t_{(j)}, t_{(j)} + \Delta t))/\Delta t}. \tag{4.14}$$

Therefore if individual i has covariate vector $\boldsymbol{x}_{(j)}$, then (4.14) can be written as

$$\frac{h(t_{(j)} \mid \boldsymbol{x}_j)}{\sum_{k \in R(t_{(j)})} h(t_{(j)} \mid \boldsymbol{x}_k)} = \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_{(j)})}{\sum_{k \in R(t_{(j)})} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_{(k)})}, \tag{4.15}$$

using the definition in (4.3) with $h(t_{(j)} \mid \boldsymbol{x}_j) = h_0(t_j) \exp(\boldsymbol{\beta}^T \boldsymbol{x}_{(j)})$.

The partial likelihood for the r failure times is therefore:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_{(j)})}{\sum_{k \in R(t_{(j)})} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_{(k)})}. \tag{4.16}$$

The effect of the covariates on the survival time is modelled through the $\boldsymbol{\beta}$ parameters, which can now be estimated even though the baseline hazard in model (4.12) is left arbitrary. Furthermore, methods also exist to estimate the shape of the baseline survivor, hazard and cumulative hazard functions, although due to the way in which the PH model is defined they can only be estimated up to the most recent observed failure time .

The presence of tied data points further complicates the derivation of the partial likelihood. Kalbfleisch and Prentice (1973) derive an exact partial likelihood for survival data with tied observations, but computationally simpler approximations have been proposed by Cox (1972), Peto (1972), Breslow (1974) and Efron (1977).

An alternative to the assumption of proportional hazards is to consider that the effect of the covariates is to directly speed up or slow down failure time. An *accelerated life* model does this by modelling the logarithm of the survival time as a linear combination of covariates i.e.

$$\log(T) = \boldsymbol{\beta}^T \boldsymbol{x}. \tag{4.17}$$

Hence the covariates directly accelerate or decelerate failure time, in contrast to the PH approach that assumes a multiplicative effect of the covariates on the baseline hazard function that is independent of time. For more detailed analysis of both of these approaches see Cox and Oakes (1984), Therneau and Grambsch (2000) or Kalbfleisch and Prentice (2002).

Here all covariates are assumed to be fixed over time - the survival models described in this section will not have the same interpretation if the covariates are time-dependent. Extensions to incorporate these will be discussed in section 4.3.4.

4.3.3 Parametric methods for modelling survival data

The Cox proportional hazards model is a powerful tool in the analysis of survival data since it does not require the assumption of a parametric form for the baseline hazard in order to estimate the effect of covariates on the survival time. There may be situations however, when either the survival distribution is known or that it is not unreasonable to assume that it has a certain parametric specification, perhaps due to the results of some exploratory analyses such as those described in section 4.3.1. In this case there are various distributions that are commonly used and a selection will be discussed here.

In addition there are also various advantages to fitting parametric survival models, particularly when it comes to predicting future survival times. In this case the Cox proportional hazards approach can only estimate the shape of the baseline hazard up to the most recent failure time, and without additional knowledge predicted estimates cannot be obtained. Parametric models have fully specified hazard functions dependent on a set of parameters

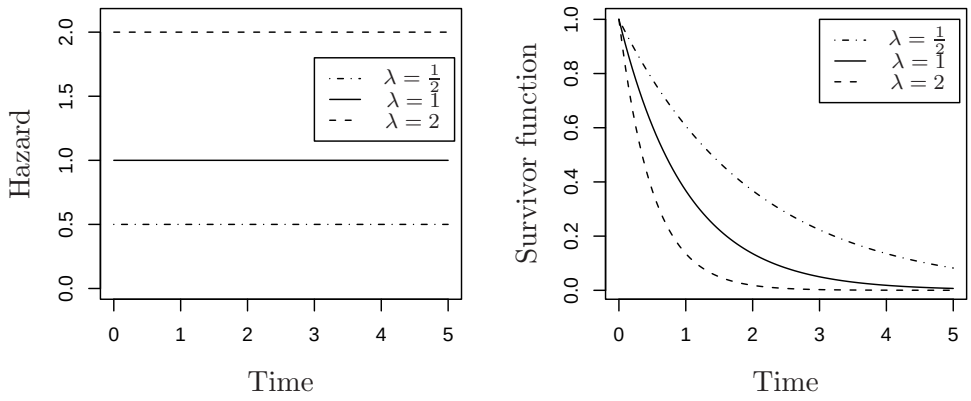


Figure 4.1: Examples of exponential hazard and survivor functions

that determine the overall distributional form governing survival times. These can be estimated, and for a model fitted at any point in time the current estimates can be used to predict the hazard at future time points.

Another advantage is that many parametric models still retain the proportional hazards or accelerated life structures described in section 4.3.2. Consider first some examples of common survival models for continuous homogeneous populations.

a) *Exponential survival model*

If the hazard function $h(t) = \lambda$ where scale parameter λ is a positive constant, then the survival times follow an exponential distribution. The survivor function is given by $S(t) = \exp(-\lambda t)$ and the density function by $f(t) = \lambda \exp(-\lambda t)$. Examples of exponential survivor and hazard functions are shown in figure 4.1.

b) *Weibull survival model*

The Weibull survival model has a monotonic hazard function of the form $h(t) = \alpha\lambda t^{\alpha-1}$ where scale parameter λ and shape parameter α are both positive. The survivor function is $S(t) = \exp(-\lambda t^{\alpha})$ and the density function is $f(t) = \alpha\lambda t^{\alpha-1} \exp(-\lambda t^{\alpha})$. Examples of Weibull survivor and hazard functions are shown in figure 4.2. It can be seen that the exponential distribution is a special case of the Weibull when the

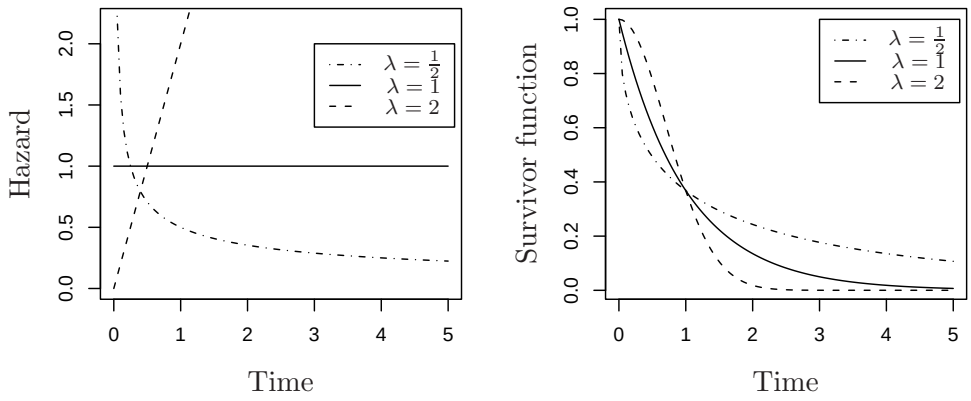


Figure 4.2: Examples of Weibull hazard and survivor functions

scale parameter $\alpha = 1$.

Since it has both a shape and scale parameter the Weibull distribution is very flexible and the hazard and density functions can take a variety of different forms. The inclusion of covariates through a log-link in the scale parameter λ also results in the model having both a proportional hazards and an accelerated life structure. Indeed the Weibull is the only parametric distribution to have this property.

Both the exponential and Weibull models have closed forms for the survivor and hazard functions and are straightforward to work with. Other possible distributions that can be used include: gamma, log-normal, log-logistic, generalised gamma, generalised F and extreme value distributions.

As noted in Kalbfleisch and Prentice (2002) any continuous survival distribution can be discretised by considering a discrete random variable T such that

$$P(T = t) = P(t \leqslant U < t + 1), \tag{4.18}$$

where U is continuous random variable with a fully specified distributional form. For example if U has a Weibull distribution with shape parameter α and scale parameter λ,

then

$$\begin{aligned} P(T=t) &= P(t \leqslant U < t+1) \\ &= P(U < t+1) - P(U < t) \\ &= F(t+1) - F(t) \\ &= S(t) - S(t+1) \\ &= \exp(-\lambda t^{\alpha}) - \exp(-\lambda(t+1)^{\alpha}). \end{aligned} \tag{4.19}$$

(Note that here we are discretising over periods of unit length 1 - this can be altered if required.)

Assuming random censoring, for n observed individuals the likelihood takes the form

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[f(t_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})\right]^{\delta_i} \left[S(t_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})\right]^{1-\delta_i}, \tag{4.20}$$

where δ_i, $i = 1, \ldots, n$, is a binary variable that takes the value 1 if individual i failed or 0 if right-censored (alternative formulations exist for left- and interval-censored data). In this way censored observations contribute $P(T \geqslant t)$ to the likelihood i.e. it is known that they survived the period $[0, t)$.

4.3.4 Time-dependent covariates

The models discussed so far can also be adapted to incorporate time-dependent covariates, though care must be taken since this can change their interpretation. Consider a covariate $x_i(t)$ for an individual i that varies over time. Let $X_i(t) = \{x_i(u); 0 \leqslant u \leqslant t\}$ denote the covariate history up to time t. The hazard function for individual i at time t is dependent on the covariate history at t and is defined to be

$$h_i(t) = \lim_{\Delta t \to 0} \frac{P\{t \leqslant T < t + \Delta t \mid T \geqslant t, X_i(t)\}}{\Delta t}. \tag{4.21}$$

Kalbfleisch and Prentice (2002) discuss two types of time-dependent covariate. The first type, so-called *external* covariates, are defined as those where the future path of the covariate up to any time $t > u$ is not affected by the occurrence of a failure at time u. A possible example of this could be air temperature in an influenza study.

A covariate that is not external is said to be *internal*. Internal covariates are typically measurements taken on an individual study subject, leading to a common property of requiring the survival of the individual for its existence. For example, in a study of survival time from a heart bypass operation to death, an internal covariate could be a patient's blood pressure. In this case the survivor function for an individual with observed covariate $x_i(t)$ at time t is $S(t_i \mid x_i(t)) = 1$.

4.4 Bayesian model fitting

A range of model fitting techniques exist to fit the sort of survival models discussed in this chapter, each having its own advantages and disadvantages. Here we will focus on Bayesian model structures and fitting mechanisms. The Bayesian approach has a number of useful properties, for example it yields not only full posterior distributions for the parameters of interest but also full posterior (predictive) distributions for predicted values. It also provides a tractable method to fit more complex models - particularly of interest are those incorporating *random effects* (or *frailties*) that attempt to account for unobserved heterogeneity in the data set.

The reader is referred to Gelman et al. (2004), Congdon (2001) and Congdon (2003) for more detail on general Bayesian methodology; and Ibrahim et al. (2001) and Hougaard (2000) for Bayesian methods in survival analysis. With regard to alternative approaches to model formulation and fitting, by far the most common is the use of maximum likelihood, and Collett (2003) gives a clear and simple account of these techniques in the context of survival analysis.

From a frequentist perspective, the unknown parameters $\boldsymbol{\theta}$ are treated as fixed values that

must be estimated from the data. In contrast the Bayesian approach instead treats the parameters as random variables that are generated from some probabilistic distribution. A standard Bayesian model takes the form:

$$p(\boldsymbol{\theta} \mid \boldsymbol{D}) = \frac{p(\boldsymbol{D} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\boldsymbol{D} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}. \tag{4.22}$$

That is the conditional *posterior* distribution for the parameters $\boldsymbol{\theta}$ given the data $\boldsymbol{D}$ is equal to the likelihood (the distribution of $\boldsymbol{D}$ given $\boldsymbol{\theta}$) multiplied by a *prior* distribution for $\boldsymbol{\theta}$, up to some normalising constant. Hence the unknown posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{D})$ is expressed in terms of a known likelihood $p(\boldsymbol{D} \mid \boldsymbol{\theta})$ and a specified prior distribution $p(\boldsymbol{\theta})$.

For simple models this can be calculated explicitly, however since the denominator involves integrating across the whole of the parameter space this becomes mathematically intractable when the number of parameters is large. Therefore a different fitting mechanism is required, the most widely used of which is that of *Markov chain Monte Carlo* (MCMC) iterative sampling.

Monte Carlo integration involves sampling a large number of observations from a target distribution, and then using these samples to estimate various expected values. The law of large numbers ensures that the estimate can be made more accurate simply by increasing the sample size. Therefore if large numbers of samples can be obtained from the posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{D})$ then Monte Carlo integration offers a method to extract the required quantities of interest from these values. All that is required is a tractable method to sample from the posterior, and this can be done using a Markov chain.

A Markov chain is a sequence of numbers where each number depends only on the previous value in the chain. It can be shown that under certain regularity conditions a Markov chain will converge to a so-called *stationary* distribution. If a Markov chain can be constructed such that its stationary distribution is identical to the posterior distribution of interest, then the required sample values can be obtained. MCMC combines these two techniques and has the advantage that it can produce estimates from the posterior distribution without requiring knowledge of the normalising constant. For more detailed

information see Gilks et al. (1996).

4.4.1 Metropolis-Hastings algorithm

This is an extension by Hastings (1970) of an algorithm proposed by Metropolis et al. (1953) that can be used to construct a Markov chain with a stationary distribution identical to the posterior distribution of interest $p(\cdot)$. Consider a m-vector of random variables $\boldsymbol{\theta}$ from (multivariate) distribution $p(\cdot)$. Then:

1. Set $t = 0$ and $\boldsymbol{\theta_0}$ to some initial value.
2. Sample a *candidate* point $\boldsymbol{\theta}_{\text{cand}}$ from a proposal distribution $q(\cdot \mid \boldsymbol{\theta}_t)$.
3. Accept $\boldsymbol{\theta}_{\text{cand}}$ with probability $\alpha(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{\text{cand}})$ where

$$\alpha(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{\text{cand}}) = \min\left(1, \frac{p(\boldsymbol{\theta}_{\text{cand}})q(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{\text{cand}})}{p(\boldsymbol{\theta}_t)q(\boldsymbol{\theta}_{\text{cand}} \mid \boldsymbol{\theta}_t)}\right). \tag{4.23}$$

4. If $\boldsymbol{\theta}_{\text{cand}}$ accepted set $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_{\text{cand}}$, else set $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t$.
5. Set $t = t + 1$ and return to step 2.

Subject to regularity conditions the proposal distribution $q(\cdot \mid \cdot)$ can take any form and the chain will still converge to the distribution of interest. The choice of proposal distribution is important however as it affects the rate of convergence of the chain. Preferably the proposal distribution should be chosen to be similar to the target distribution $p(\cdot)$. Note that knowledge of the normalising constant is not required, since it cancels out in the ratio of (4.23).

Since MCMC uses large samples to estimate the characteristics of the posterior distribution there are obviously important modelling issues such as choice of initial values for the parameters, the length of the burn-in period, the length of the chain and the value of the acceptance ratio. The length of the burn-in period (i.e. the period before the chain converges) should not be too short that samples are taken before the chain has converged to

the stationary distribution, but neither should it be over long since this can unnecessarily increase computation time.

Once the chain has converged, the number of samples returned must be enough to ensure a reasonable degree of accuracy without over burdening the computation time. Careful monitoring of the acceptance ratio helps to control both the rate of convergence and, along with posterior thinning, ensures independence of the samples. Many techniques have been proposed to assess convergence and some of them will be discussed in later chapters (see Gilks et al. 1996).

4.4.2 Gibbs sampling

The parameters $\boldsymbol{\theta}$ do not have to be updated as a block, but can be updated separately if preferred, with corresponding changes to the proposal distributions. In this circumstance a special case of the Metropolis-Hastings algorithm occurs when knowledge of the full conditional distributions for individual parameters θ_i, $i = 1, \ldots, m$, given θ_{i-}, that is $p(\theta_i \mid \boldsymbol{x}, \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_m)$ are known. Hence the proposal distribution $q(\theta_{\text{cand}_i} \mid \theta_i, \theta_{i-}) = p(\theta_{\text{cand}_i} \mid \theta_{i-})$, and as a result the acceptance probability in (4.23) is always equal to one. This technique is known as Gibbs sampling (Geman and Geman 1984, Gelfand and Smith 1990).

Combinations of Metropolis-Hastings and Gibbs sampling can be used if required, and the adaptive-rejection sampling method proposed by Gilks and Wild (1992) means that as long as the conditional distributions of the parameters are log-concave, then Gibbs sampling can be used even if the distribution is complicated and is not specified explicitly. These techniques are implemented in WinBUGS (Bayesian inference Using Gibbs Sampling), an open source package developed by the MRC Biostatistics Unit in Cambridge and Imperial College School of Medicine at St Mary's, London. It can be downloaded from http://www.mrc-bsu.cam.ac.uk/bugs/ and will be used to fit all of the models in this thesis.

4.4.3 Identifiability

Consider a set of probability densities $\{f(\boldsymbol{y} \mid \boldsymbol{\Psi}) : \boldsymbol{\Psi} \in \boldsymbol{\Omega}\}$ where $\boldsymbol{\Omega}$ is the parameter space. In order for the set of densities to be *identifiable* each set of parameters $\boldsymbol{\Psi}$ must uniquely determine a corresponding member density. If this is not the case then parameter estimates derived from the data are not meaningful.

A more rigorous definition of identifiability is provided by Basu (1983):

> Let U be an observable random variable with distribution function F_θ and let F_θ belong to a family $\mathcal{F}\{F_\theta : \theta \in \Omega\}$ of distribution functions indexed by a parameter θ. Here θ could be scalar or vector valued. We shall say that θ is non-identifiable by U if there is at least one pair (θ, θ'), $\theta \neq \theta'$ where θ and θ' both belong to Ω such that $F_\theta(u) = F_{\theta'}(u)$ for all u. In the contrary case we will say θ is identifiable.

4.5 Extensions to conventional survival models

This section covers some extensions to the conventional model formulation that deal with different situations. Again, detailed analysis will not be given here - rather this is to give a flavour of the scope and potential of survival modelling.

4.5.1 Long-term survivor or cure rate models

Consider a disease in which individuals who recover from infection are conferred immunity from future infection. For a set of survival data it is reasonable to assume in this case that there is a proportion, p, of the total population that are considered 'immune' or 'cured' of the disease. A conventional survival approach would treat these individuals as censored at the end of the study period, such that although they were not observed to become infected they still have the potential to become so. Clearly this is intuitively unreasonable

and will result in biased parameter estimates since the likelihood contributions from these individuals will be incorrect.

Boag (1949) was the first to publish a paper discussing a survival model incorporating a so-called 'cure' proportion; though this was later extended in Berkson and Gage (1952) who noted that the hazard function for 'cured' individuals should reduce to the baseline hazard for the population. The standard model proposed by Berkson and Gage (1952) is modelled through the survivor function as

$$S(t) = p + (1 - p)S^*(t), \tag{4.24}$$

where $S^*(t)$ is the survivor function for the susceptible proportion, and p is the proportion of 'cured' individuals in the population.

This model is usually referred to as a *cure rate* or *long-term survivor* model, and can be used whenever it is believed that there is a proportion of the population 'immune' to failure in some way. For a detailed introduction of the field see Maller and Zhou (1996).

4.5.2 Mixture models

The standard long-term survivor model, (4.24), supposes that the 'cured' proportion can never experience failure. Consider instead a simple generalisation where the population consists of two groups, each subject to a different survival process i.e.

$$S(t) = pS_0(t) + (1 - p)S_1(t), \tag{4.25}$$

where the mixing parameter p is the proportion of individuals in the population from group 0, with corresponding survivor function $S_0(t)$, and $(1 - p)$ is the proportion of individuals from group 1 with corresponding survivor function $S_1(t)$. This is a standard two-group mixture model - though the methodology can be generalised to three or more groups if required.

Mixture models are widely documented in the statistical literature and are used in many different contexts. For a detailed overview of mixture models see McLachlan and Peel (2000).

4.5.3 Competing risks models

Often there is more than one way that an individual can experience failure. Traditional survival analysis techniques do not attempt to differentiate between multiple causes of failure. Competing risks analysis is an extension of survival analysis that incorporates this extra information. Good overviews of competing risks can be found in Crowder (2001), McLachlan and Peel (2000) and Congdon (2001).

A straightforward way of modelling competing risks is to use a framework analogous to the mixture model in (4.25) i.e.

$$S(t) = pS_0(t) + (1-p)S_1(t), \tag{4.26}$$

where now p is probability of failure from cause 0 and $(1-p)$ is the probability of failure from cause 1. In contrast to the mixture model the causes of failure are assumed to be observed and independent (though techniques exist to incorporate certain amounts of missing data). Various extensions exist for the competing risks model, including generalisations to more than two causes of failure, and also as part of a long-term survivor model (Ng and McLachlan 1998).

4.5.4 Multi-state models

All the models discussed so far assume that the underlying survival process for an individual remains the same over time. That is that although the risk may change over time, the underlying process will not. A multi-state framework models a stochastic process by allowing individuals to belong to one of a (finite) number of discrete states at any time

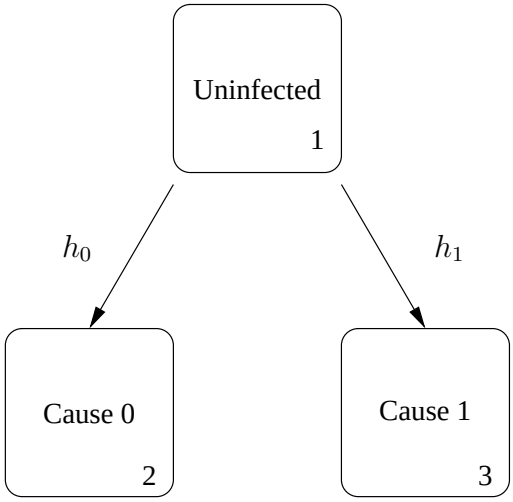


Figure 4.3: Possible graphical representation of competing risks model in multi-state framework

point. They offer a flexible framework for modelling many different kinds of longitudinal data, for example the conventional survival model could be viewed as a multi-state model with two states: failed or not failed. Figure 4.3 shows a possible graphical representation of the standard competing risks model (4.26) a multi-state framework.

Transitions between states are modelled through *transition hazards* (h_0 and h_1 in figure 4.3). Often the state structure is not unique and the formulation of different state structures can make the interpretation and fitting of the models much easier. The likelihood formulation is usually based around an assumption that the movement between states is governed by a Markov process, though this is not always the case. Hougaard (2000, 1999) and Commenges (1999) provide useful overviews of this methodology.

4.5.5 Change point models

Change point models are similar in concept to the multi-state models described in section 4.5.4 except that they assume that the distributional form of the entire survival process changes at one or more points in time (i.e. the entire process moves between states as opposed to individuals moving between states). Recently Ebrahimi et al. (1997) and Chung et al. (2005) have developed Bayesian models for k change points. Chung et al.

(2005) define the hazard in this case as

$$h(t) = h_1(t)I(0 \leqslant t \leqslant \tau_1) + \cdots + h_k(t)I(\tau_{k-1} < t < \tau_k) + h_{k+1}(t)I(t > \tau_k), \qquad (4.27)$$

where $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_k)$ is the vector of change point parameters with

$$I(A) = \begin{cases} 1 & \text{if } x \in A \\ \\ 0 & \text{otherwise.} \end{cases}$$

4.5.6 Random effects (frailties)

When modelling stochastic processes there is often unknown heterogeneity in the underlying risk. The idea of a random effect in the model is to try to account for any heterogeneity that cannot be explained by covariates either because they are unknown or cannot be obtained (Vaupel et al. 1979). Random effects may easily be handled using the Bayesian approach since this considers all parameters as random (Gilks et al. 1996). Traditionally in survival analysis random effects are known as frailties, and may be either spatially structured or unstructured, and for a hierarchical model can be defined at any level. In this section a model incorporating individual specific frailties will be discussed. Models including both fixed and random effects are referred to as *mixed models*.

Consider the following form (Shimakura 2003): let Z_i be a random variable for an individual i, where Z_i comes from a non-negative distribution with mean 1 and variance τ. Considering a model with a proportional hazards structure such as (4.12) with $\psi(\cdot) = \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)$; then for an individual i with observed failure or censoring time t_i and covariate vector $\boldsymbol{x}_i$, a frailty effect z_i can be included as:

$$h(t_i \mid \boldsymbol{x}_i) = z_i h_0(t_i) \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i). \qquad (4.28)$$

If there is no individual heterogeneity then (4.28) reduces to the standard proportional hazards model (4.12), else for $z_i > 1$ they experience an quicker failure rate and for $z_i < 1$

they experience a slower failure rate. Shimakura (2003) provides a list of references concerning the distributional forms that the Z_i can take, including: gamma (Clayton 1978), log-normal (McGilchrist and Aisbett 1991), power variance model (Aalen 1988), positive stable distribution (Hougaard 1986a) and the inverse normal distribution (Hougaard 1986b).

Spatial structure can also be incorporated into the frailty specification by considering a transformed frailty vector $\boldsymbol{W} = \log(\boldsymbol{Z})$, where $\boldsymbol{W}$ has a multivariate Normal distribution with a spatially structured correlation matrix. Here $\boldsymbol{W}$ is included additively in the linear regression parameters. Alternatively, another common way to specify $\boldsymbol{W}$ is to use a conditional autoregressive (CAR) normal distribution. In this case the mean response for an individual is conditional on the mean of its neighbours (Besag and Kooperberg 1995 and Besag et al. 1991).

4.5.7 Extensions to multivariate response variables

Methodology exists to extend the univariate survival framework to a multivariate one. The reader is referred to Hougaard (2000) for comprehensive details.

4.6 Conclusions

It can be seen that survival analysis offers a rich variety of options to deal with a host of different situations that may arise when studying failure time data. In particular the conventional survival approach is a straightforward way to deal not only with the censoring issue, but also to cope with non-normally distributed failure times. In addition issues such as time-dependent covariates, mixed survival groups and immunity to failure can all be handled, and in combination with Bayesian methodology, complex frailty models can be fitted to help account for unobserved heterogeneity in the data set.

In subsequent chapters the methodology discussed here will be used to develop some

modelling frameworks, based around survival techniques, that can be used to study the dynamics of infectious disease processes. As discussed in chapters 2 and 3 there are many facets affecting the spread of an infectious animal disease, ranging from biological aspects, such as species varying viral excretion and susceptibility, to the structural assumptions made in the model formulation itself, such as the specification of spatial and temporal dependence. In chapter 5 we will construct a simple preliminary model to explore some of these issues, and fit it to data from the FMD epidemic in Devon in 2001. The analysis will attempt to highlight ways in which inferences and predictions from the model are affected by various aspects relating to the assumptions made.

Chapter 6 will then explore in more detail ways in which resistance to infection can be dealt with in the survival framework, and we conduct a simple simulation study to draw attention to the potential problems encountered when resistance to infection is present in the data but unaccounted for. Chapter 7 will extend these ideas to a spatial setting with a further, more detailed, simulation study. In addition the potential predictive uses of the model are explored.

Chapter 5

Preliminary survival modelling of FMD

In the previous chapter we discussed various ways in which the survival framework can be adapted to deal with different situations. In this chapter we will explore some of the issues associated with modelling FMD using a simple traditional survival model. We investigate the effects of different modelling assumptions on the predictions of future failure times and risk of infection.

Potential ways in which this approach can be used to incorporate some of the key elements of the FMD spreading process; such as variable susceptibility and infection, and the inclusion of spatio-temporal correlation is also explored. Of particular interest is the derivation of a space-time varying *viral load* covariate to measure the infectiousness of IPs, based on models for the within-herd spread of the disease, allowing for different herd sizes and species type. This covariate is used as a means of determining exposure and targeting the survival modelling to 'at-risk' premises.

As discussed in chapter 2, our focus is concentrated on modelling the spread of FMD in Devon and not across the entire UK, so the terms 'global' and 'localised' here will refer to progression of the disease over the whole of the Devon region and at the between-premise

level respectively. As a convention, any holding environment that contains animals capable of contracting FMD will be referred to as a premise, since risk of infection from FMD is not limited to farm holdings only.

The structure of the chapter is as follows: section 5.1 contains more detailed discussion on the data set for the sub-epidemic of FMD in Devon in 2001. Examined in this section are some of the difficulties involved with the collection and interpretation of the epidemic data at this scale. In sections 5.2 and 5.3 we specify a preliminary survival model, including incorporation of covariates and censored information, likelihood formulation and the method of prediction. In sections 5.4 and 5.5 this preliminary model is fitted to the Devon data set with two initial naive covariates.

The initial covariates are refined in section 5.6, where the concept of viral load is introduced, and further developed in section 5.8, which discusses its use as a means of censoring the data set via exposure. Results from the viral load model fitted to the full data set and data set censored via exposure are given in sections 5.7 and 5.9 respectively.

Section 5.10 covers the potential uses of uninfected animal densities as surrogates for susceptibility in the models. Some conclusions are given in section 5.11.

5.1 The 2001 Devon data set

The 2001 UK FMD epidemic is the most completely documented major outbreak of animal disease to date (DEFRA 2004). Although the quality of the data can vary over time due to complications involved in data collection on a large scale (this is particularly evident in the earlier stages of the epidemic), it still provides a vital source of information for epidemiological researchers.

There were issues regarding both the logistical constraints involved in the collection of the data, and also various theoretical aspects concerning its interpretation and/or its collation into a coherent form for use by researchers. For example, quick decisions had

to be made about what factors were relevant, how they could be measured, who would measure them and how to ensure that the necessary steps would be taken to implement these procedures efficiently. In reality the British government was caught unawares by the speed in which the epidemic became established, and understandably data collection was relegated in favour of more urgent issues. Indeed much work had to be done post-hoc in order to collate the necessary information from multiple data bases and mixed sources into a usable form. One almost unanimous consensus arising from the epidemic is that more rigorous methodology needs to be instigated for data collection in the future, not so as to detract from important work in the field, but to help gauge a better understanding of the disease dynamics in case of further outbreaks.

A common example of the kinds of issues affecting large-scale data collection concerns accurate estimation of the total number of animals slaughtered 'at foot' i.e. slaughtered but not counted (DEFRA 2004). DEFRA estimate that 6 million animals were slaughtered nationwide as a result of the epidemic, but as many as 4 million additional young animals (though this estimate is thought to be high) may also have been slaughtered. Clearly large inconsistencies in estimations will impinge on the accuracy of any inferences derived from them.

Further issues arise when the georeferencing of premises is considered. The June 2000 census data included grid references for the point locations of holding premises. These were calculated from data collected by DEFRA using the Integrated Administration and Control System (IACS). When a farmer wishes to claim subsidies from the European Union (EU) they are required to provide a grid reference for the central point of each field that they manage. From this an overall grid reference for the premise is allocated as the location of the field closest to the centroid of all fields belonging to the holding premise (with adjustment if this lies outside the corresponding parish and ward boundaries). However this is only one of a range of ways in which farm premises can be geo-referenced into point locations. Durr and Froggatt (2002) investigated different methods by using the proportion of the farm area captured as the main factor in determining the optimal technique. Their conclusions, based on a case study in Cornwall, UK, was that the main farm building

was the best reference point and this contradicts the method used during the June 2000 census.

There is also confusion around the ownership and spatial dispersion of animals identified through census data. For example many landowners own sets of non-contiguous land parcels, each of which may be worked on either by the landowner themselves, or rented to other tenants. It is often very hard to keep track of exactly which areas of land are attributable to which particular tenants. In addition there is also the issue of quantifying the spatial dispersion of animals about each location; including which animals are located on which particular parcels of land. If the geographical and topographical complexity of the land is also considered then this further compounds the task of accurately quantifying the spatial spread and density of livestock across the study region. Kao (2001) in particular states the important role that landscape fragmentation plays in the spread of FMD and the particular problems associated with using point locations as a representation of the true location of the herd. He notes that in Devon for example, the average size of grazing areas for cattle and sheep is smaller than in Cumbria, yet the density of holding premises is similar. This indicates a more fragmented landscape in Devon than in Cumbria, which affects the pattern of disease spread across each region.

The original data set provided by VLA for Devon consisted of almost 10000 premises, of which only 171 became infected. The numbers of animals on infected premises are assumed to be accurate since they were recorded by visiting veterinarians. However for UIPs the numbers of animals were taken from census data collected in June 2000 and there is the possibility that these may vary significantly from their recorded values, since it is incredibly hard to keep an accurate track on the movements of a large number of animals. Some data does exist, in the form of the Cattle Traceability System (CTS) run by the British Cattle Movement Service (BCMS), however the database was not designed for use as an epidemiological tool and only holds information for movements directly into and out of premises.

The CTS was set-up in 1998 by MAFF to aid the lifting of the export ban on British

beef in the wake of the BSE crisis. Since 2001 it has been compulsory to report all cattle movements to the CTS, though certain exceptions exist. In 2005 DEFRA established the Rapid Analysis and Detection of Animal-related Risks project (RADAR) that has sought to make this data more widely available to researchers. There is currently a lot of work being done to extract more useful information from the data, such as contact tracing networks of cattle movements around the UK (Vernon et al. 2005). Although potentially very important in the longer term, this is an extremely large and complex process, from which detailed knowledge is yet to emerge, so it will be assumed here that the census counts are representative of the true size of each premise.

It was also necessary to make a series of additional assumptions about the spreading process. Since the density of animals has been identified as an important factor in the spread of FMD (Hugh-Jones and Wright 1970), uninfected premises that were located at the same grid reference were amalgamated. In addition information about land fragmentation and geographical complexity was unknown, and so we assume that animal density is isotropic i.e. that the spatial dispersion of animals around a point location is dependent on distance only.

Other information lacking from the data set concerned the environmental conditions over the course of the epidemic. It has been noted (Hugh-Jones and Wright 1970, Gloster et al. 1981, Donaldson 1983, Donaldson et al. 2001) that weather conditions, in particular temperature, wind speed and wind direction can play an important part in the spread of FMD, and yet no data was collected at the time of the outbreak. As a result we assume that the virus can not persist outside of the host for periods of longer than one day, and that the airborne distribution of virus spores follows an isotropic distribution over space.

Estimation of the lag between the actual date of infection and the date of report is also important. In practice veterinarians can estimate the age of the oldest lesion found in an infected herd, however the length of the incubation period associated with FMD (the time before clinical signs appear) can be anywhere between 1 and 14 days (Alexandersen et al. 2003b) - so even assuming that the animal with the oldest lesion was to first to become

infected it is still difficult to ascertain the exact date of entry of the virus into the herd. As such the date of infection was based on the age of the oldest identified lesion and the incubation period of the disease was assumed constant and identical for each animal.

Keeling et al. (2001b) cite a number of other potential sources of heterogeneity, including: farm level variability in biosecurity, dynamics of within-premise epidemics, and the relative infectivity and susceptibility of different species. Furthermore there are many difficulties associated with identifying clinical signs on infected animals (particularly in sheep). Some of these issues will be addressed in due course.

The final Devon data set consisted of 8729 premises, of which 171 became infected. For each premise the data consisted of a point location in terms of an (x, y)-coordinate derived from latitude and longitude, the number of pigs, cattle and sheep (any goats were treated as sheep since they share similar traits with regard to FMD), the date of infection, the start and end dates of culling (if applicable) and the type of cull (IP, DC, CP, Welfare, unknown). The spatial distribution of premises at the end of the epidemic is shown in figure 5.1.

The epidemic in Devon lasted 113 days from the date of the presumed initial infection date for the first case on 15^{th} February 2001 to the date of the final presumed infection on 8^{th} June 2001 . The temporal evolution of the epidemic is shown in figure 5.2. It can be seen that the epidemic peaked at around 34 days and then steadily declined with a few sparse infections around the tail end of the epidemic.

5.2 Specification of basic model

A desirable aim in this project is to attempt to develop a survival model that can be fitted sequentially as an epidemic progresses, helping to identify and explain patterns and trends in the data. A second, more important objective, is to use the results from this sequential model fit to predict the future path of the epidemic. One difficulty in any type of modelling situation is assessing how well the developed models fit the data. For the

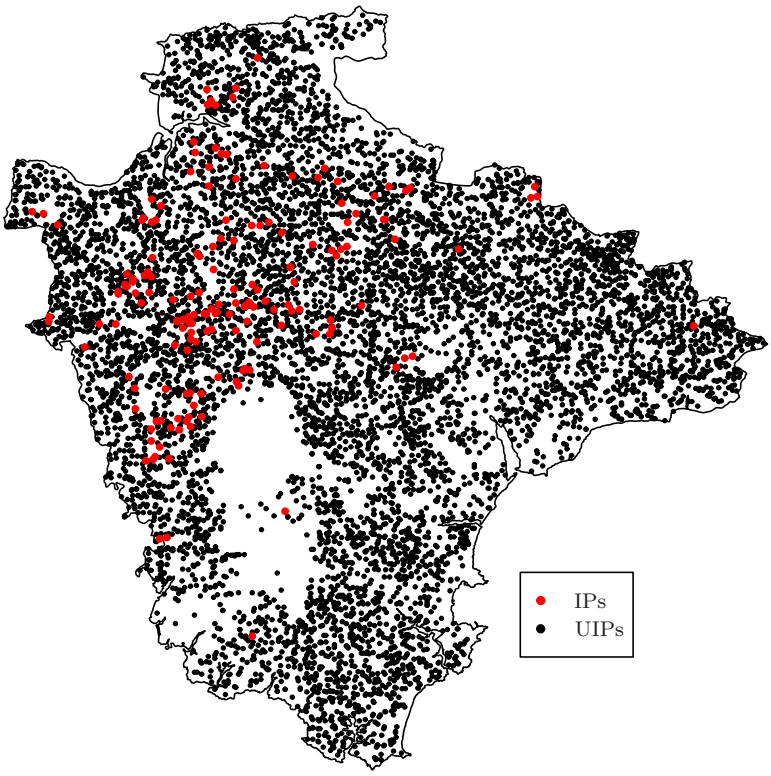


Figure 5.1: Spatial distribution of premises in Devon at the end of the 2001 epidemic

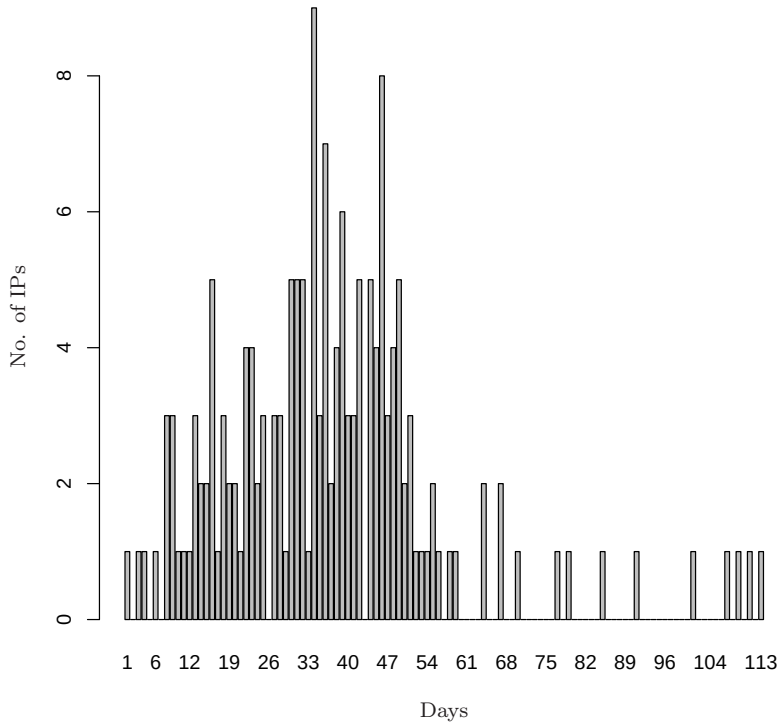


Figure 5.2: Temporal distribution of infections in Devon during the 2001 epidemic

preliminary investigation it was decided that a reasonable method of assessing fit would be to censor the data set at some arbitrary time point prior to the actual end of the epidemic, fit a model, and then use the resulting parameter estimates to predict future failure times for the remaining uninfected premises. In this way the predicted failure times can be compared to the actual failure times.

That said we consider a basic survival model formulation for the data described in section 5.1. It is important to think of the data as a discretisation of a continuous process, since although infection can occur at any time, observations are only recorded (at most) at one day intervals. Also, the future path of the epidemic at each stage depends on the history of the epidemic up to that point and it is vital that the model is developed with this in mind. It is therefore likely that time-dependent covariates will be of interest.

We specify an initial model through the hazard function, based on a Weibull distribution. We chose a parametric form so that the model is capable of predicting future infection times, and having both a shape and scale parameter means that the Weibull is reasonably flexible. The model is analogous to the discrete model (4.19) discussed in section 4.3.3, with discretisation over periods of one day, though the formulation used here is slightly different.

Consider initially a continuous random variable U representing survival time where $U > 0$ follows a Weibull distribution with hazard function

$$h_c(u) = \alpha\lambda u^{\alpha-1}, \tag{5.1}$$

and survivor function

$$S_c(u) = \exp(-\lambda u^{\alpha}). \tag{5.2}$$

Here α and λ are positive shape and scale parameters respectively. Now consider a discrete random variable T representing survival time where $T = 1, 2, \ldots$. We can view the discrete hazard at time t as being the probability that U lies in the interval $[t-1, t)$, given survival

to t i.e.

$$\begin{aligned} h(t) = P(T = t \mid T \geqslant t-1) &= P(t-1 \leqslant U < t \mid U \geqslant t-1) \\ &= \frac{P(t-1 \leqslant U < t)}{P(U \geqslant t-1)} \\ &= \frac{P(U < t) - P(U < t-1)}{P(U \geqslant t-1)} \\ &= \frac{P(U \geqslant t-1) - P(U \geqslant t)}{P(U \geqslant t-1)} \\ &= \frac{S_c(t-1) - S_c(t)}{S_c(t-1)} \\ &= 1 - \frac{S_c(t)}{S_c(t-1)} \\ &= 1 - \exp(-\lambda[t^\alpha - (t-1)^\alpha]), \end{aligned} \tag{5.3}$$

which is bounded in the region $(0, 1)$. It is sensible to view the hazard in this way since in an epidemic situation a premise that is confirmed to be infected at day t has a true infection time in the region $[t-1, t)$, (given that it wasn't confirmed to be infected at day $t-1$).

From the identities given in (4.8) and (4.9), the corresponding discrete survivor function is given by

$$\begin{aligned} S(t) = P(T \geqslant t) &= \prod_{j=0}^{t} (1 - h(j)) \\ &= \prod_{j=1}^{t} \exp(-\lambda[j^\alpha - (j-1)^\alpha]) \\ &= \exp\left(-\sum_{j=1}^{t} \lambda[j^\alpha - (j-1)^\alpha] \right) \qquad t = 1, 2, \ldots, \end{aligned} \tag{5.4}$$

and the probability function by $P(T = t) = h(t)S(t-1)$ i.e.

$$P(T=t) = \begin{cases} 1 - \exp(-\lambda) & t = 1 \\ [1 - \exp(-\lambda[t^\alpha - (t-1)^\alpha])] \times & \\ \quad \exp\left(-\sum_{j=1}^{t-1} \lambda[j^\alpha - (j-1)^\alpha]\right) & t = 2, 3, \ldots \end{cases} \tag{5.5}$$

So $S(t)$ is a decreasing function bounded above by 1 and below by zero where for fixed λ and α, $S(t) \to 0$ as $t \to \infty$. Likewise $P(T = t)$ is a proper probability function bounded in the interval $[0, 1]$.

If covariates are included through a log-link in the scale parameter λ and considered fixed (time-independent), then (5.5) is identical to the discrete model derived in section 4.3.3. In the case of time-varying covariates, specifying the model through the hazard function allows the dependence on the covariate history at previous time points to be conditioned out. So for a time-dependent covariate λ_t calculated at time t, model (5.3) becomes:

$$h(t) = 1 - \exp(-\lambda_{t-1}[t^\alpha - (t-1)^\alpha]), \tag{5.6}$$

and this leads to straightforward modifications to the survivor (5.4) and probability (5.5) functions. Hence the conditional probability of failure given survival is dependent on the covariate value at that point, however the overall survival and probability functions contain information on the entire covariate history.

In the remainder of the thesis model (5.6) will be referred to as the *conventional survival model*. As noted in section 4.1, knowledge of either the hazard, survivor or density (probability) functions will uniquely determine the survival distribution.

As previously highlighted, one advantage of the survival framework is that censored observations can be readily incorporated into the model. The Devon data set is technically interval-censored; that is that for any IP the actual date of infection is known only to within a set interval - calculated from the age of the oldest observed lesion and the maxi-

mum corresponding incubation period. In theory interval- or left-censoring could be used to compensate for these assumptions, and to provide estimates for the distributions of these times, however at this stage we consider the data to be right-censored, i.e. the effect of the incubation period is assumed to be constant and identical between animals. The actual infection date for any IP is taken to be the date of report minus the age of the oldest observed lesion.

An additional issue also arises with the handling of culled premises in the model. Culled IPs can be included as normal observations since they have experienced infection, though the culling will have the effect of reducing the levels of virus in the neighbouring region by removing infected animals, and this needs to be accounted for (section 5.6 discusses a way in which this can be done through the use of the viral load covariate). However most uninfected premises are culled because they are considered to be at 'high-risk' of infection. If these are treated as right-censored at the time of culling then the study population is no longer truly representative of the population at-risk, i.e. censoring is not random and may result in biased parameter estimates and predictions (Kalbfleisch and Prentice 2002).

For the time being culled UIPs are left out of the model, and the resulting censoring mechanism is assumed random. These observations still contain useful information about the spreading process, and in section 5.10 we introduce the concept of uninfected animal densities that can be used as a surrogate for susceptibility to infection. In this case some information from the culled UIPs is still included in the model through changes in the uninfected animal density over time.

Furthermore there is the possibility of latent infections in culled UIPs; where the premise is infected but culled before clinical signs appear. Here culled UIPs are assumed uninfected at the time of slaughter, however it may be possible to develop a way of imputing this information from data during the model fit if so required (see e.g. Deardon et al. 2006).

Given the above discussion, for a set of n observed failure (or censoring) times t_i, $(i =$

$1, \ldots, n$), the likelihood is given by:

$$L(\cdot) = \prod_{i=1}^{n} \left[P(T = t_i)\right]^{\delta_i} \left[P(T \geqslant t_i)\right]^{1-\delta_i}, \tag{5.7}$$

where δ_i is a binary variable with the value 1 if premise i is an IP or 0 if censored. Covariates $\boldsymbol{x}$ can be included through the scale parameter, using the specification $\lambda = \exp(\boldsymbol{\beta}^T \boldsymbol{x})$. As discussed in chapter 4 we will use Bayesian methodology and MCMC to fit the models, allowing full posterior distributions for the parameter estimates and predictions to be obtained.

5.3 Method of prediction

The Bayesian framework has the advantage that it can produce full predictive posterior distributions for the failure times for censored individuals. Ordinarily these predictive distributions can be routinely obtained from the MCMC run, however here it is possible that we will want to include complex time-dependent covariates, where the value of the covariate at each time point depends on the previous history of the epidemic. In this case a post-hoc approach must be adopted in which the posterior samples for the parameters are used to drive a series of simulated epidemics over a specified time period (e.g. E days) from which Monte Carlo estimates of different quantities can be derived.

To do this consider a matrix of K posterior samples for the m parameters obtained from a model fitted at time t. Firstly remove any premises from the data set that are neither susceptible to infection nor infective (e.g. any that have been culled or vaccinated for example). Then split the data into two groups - IPs and UIPs. The IP group (n_{inf} premises) then consists of all premises that are infected and contagious and the UIP group (n_{cens} premises) consists of all uninfected and susceptible premises.

For the UIPs set up an indicator matrix $\boldsymbol{\Phi}$ with the number of rows equal to the number of premises ($i = 1, \ldots, n_{\text{cens}}$) and the number of columns equal to the number of posterior samples. Initially set each row equal to zero ($\phi_{i\cdot} = 0$). Let $\mathbf{T}$ be the corresponding matrix

of predicted survival times with elements $\{t_{ik}\}$.

If $x_i(t)$ is a time-dependent covariate for an uninfected premise i at time t, the predictive algorithm for a model fitted at time t is given by:

1. Set $k = 1$, $\nu = t$, $\text{IP}^{(\nu)}$=IP and $\text{UIP}^{(\nu)}$=UIP.
2. Take k^{th} set of posterior samples and calculate $x_i(\nu)$ and $h_i(\nu \mid x_i(\nu)) = P(\nu \leqslant T < \nu + 1 \mid T \geqslant \nu, x_i(\nu))$ for all uninfected premises (using $\text{IP}^{(\nu)}$ if necessary).
3. Let u_i be a random sample from a $U(0, 1)$ distribution corresponding to premise i.
4. If $u_i < h_i(\nu \mid x_i(\nu))$ and $\phi_{ik} = 0$ then set $t_{ik} = \nu$ and $\phi_{ik} = 1$.
5. Set $\nu = \nu + 1$. Update $\text{IP}^{(\nu)}$ to include all new infected premises (i.e. $\{\text{UIP}^{(\nu-1)} \mid \phi_{ik} = 1\}$), and remove them from $\text{UIP}^{(\nu)}$ such that $\text{UIP}^{(\nu)} = \{\text{UIP}^{(\nu-1)} \mid \phi_{ik} = 0\}$.
6. If $\nu > E$ or there are no more uninfected premises remaining then go to step 7. Else go to step 2.
7. Set $t_{ik} = E$ for all remaining censored premises, set $k = k + 1$, $\nu = t$, $\text{IP}^{(\nu)}$=IP and $\text{UIP}^{(\nu)}$=UIP.
8. If $k \leqslant K$ then return to step 2; else END.

For a time-independent variable then $x_i(t) = x_i \; \forall \; t$. This predictive algorithm was coded in R.

One issue here is the determination of a reasonable value for the predictive time period, E. In an ideal situation the predicted epidemic should be allowed to run its full course; either until all premises become infected or the epidemic dies out. Determining when this has happened is difficult however, since the tail end of epidemics tend to be drawn out and often exhibit infrequent spark infections. In practice the models are intended to be fitted sequentially, and so as long as E is large enough and the gaps between the sequential fits are small enough this shouldn't be a major issue. To put this another way, premises whose

posterior predictive samples contain large numbers of censored values are effectively those that the model is predicting will not be 'at-risk' during the period before the time of the next model fit. A balance is required such that E is large enough to allow the predicted epidemic to develop, but not too large that it becomes computationally unfeasible.

Another problem is that premises that survive beyond E will have an unknown predicted failure time. This is similar to having a set of survival times where some observations are censored at E. Unfortunately this will bias any Monte Carlo estimates derived from the model. If a reasonable parametric form for the predictions can be assumed, then the contribution of each posterior predictive sample to any Monte Carlo estimate can be weighted by using a standard survival likelihood procedure, such as the one described in (4.20). Here the posterior samples for the predicted infection indicators, $\mathbf{\Phi}$, are used to define censored and observed failures. This requires an additional distributional assumption for the posterior predicted failure times. Another approach would be to consider the use of a non-parametric alternative such as Kaplan-Meier - see section 4.3.1 - to derive estimates of the survivor function.

For a reasonable model the censoring should therefore not have too large an adverse effect on the interpretation of the posterior predictions for 'at-risk' premises over the short term. Another option is to run the epidemic forwards until the time gap between successive infections exceeds a certain pre-determined value. The posterior distribution for the probability of infection in the next E days is also given by $\mathbf{\Phi}$, and this is perhaps a more robust measure for quantifying risk than the predicted survival time (in terms of interpretation).

The accuracy of the Monte Carlo estimates is therefore directly linked to the amount of 'censored' samples in the posterior. In the case where the number of these observations is large, then since short-term prediction is key this only affects the degree of confidence in using the predictive posterior samples to directly infer the time-to-infection, and not the risk of infection.

5.4 Preliminary covariates

Since localised spread of FMD is being modelled, two obvious initial covariates will be considered, both based on distance from IPs. The first, distance from the initial source of infection ($\mathrm{D}^{(\mathrm{S})}$), is a fixed time covariate and as such the complete likelihood for the conventional model simplifies to:

$$\begin{aligned} L(\alpha, \beta_0, \beta_1) &= \prod_{i=1}^{n} \Big([\exp(-\lambda_i (t_i - 1)^{\alpha}) - \exp(-\lambda_i t_i^{\alpha})]^{\delta_i} \\ &\qquad \times [\exp(-\lambda_i t_i^{\alpha})]^{1-\delta_i} \Big), \end{aligned} \tag{5.8}$$

due to the choice of discretisation. Here $\lambda_i = \exp(\beta_0 + \beta_1 \mathrm{D}_i^{(\mathrm{S})})$, where $\mathrm{D}_i^{(\mathrm{S})}$ is the distance between premise i and the source premise.

The second initial covariate, the nearest infected neighbour distance ($\mathrm{D}^{(\mathrm{I})}$), will be time-dependent, with the likelihood given by:

$$\begin{aligned} L(\alpha, \beta_0, \beta_1) &= \prod_{i=1}^{n} \Bigg(\Bigg\{ \left[1 - \exp(-\lambda_{i(t_i-1)} [t_i^{\alpha} - (t_i - 1)^{\alpha}]) \right] \\ &\qquad \times \exp\left(- \sum_{j=1}^{t_i - 1} \lambda_{i(j-1)} [j^{\alpha} - (j-1)^{\alpha}] \right) \Bigg\}^{\delta_i} \\ &\qquad \times \left[\exp\left(- \sum_{j=1}^{t_i} \lambda_{i(j-1)} [(j)^{\alpha} - (j-1)^{\alpha}] \right) \right]^{1-\delta_i} \Bigg), \end{aligned} \tag{5.9}$$

where $\lambda_{it} = \exp(\beta_0 + \beta_1 \mathrm{D}_{it}^{(\mathrm{I})})$ and $\mathrm{D}_{it}^{(\mathrm{I})}$ is the distance between premise i and the nearest infected premise at time t (alternative ways of using nearest infected neighbour distance can be seen in Lawson and Zhou 2005).

5.5 Initial model results

To complete the Bayesian specification of the models, prior distributions were assigned to the parameters. The intercept parameter β_0 was given an uninformative $N(0, 100)$ prior,

β_1 a $N(0, 10)$ distribution, and the shape parameter α a $G(0.1, 10)$ (i.e. mean of one and variance of 10) prior. The models were fitted in WinBUGS, but since neither (5.8) or (5.9) are included in the list of standard probability distributions, they had to be specified in a slightly different manner (see appendix A for details).

The models were quite sensitive to the choice of initial value, which is perhaps unsurprising since the mean and variance of the Weibull distribution are directly linked through the shape and scale parameters. To generate initial values, a value of α was sampled from an arbitrary gamma distribution. This value was then taken, and along with the range of the data was used to obtain reasonable upper and lower limits for a set of uniform distributions from which starting values for the regression parameters β_0 and β_1 could be randomly sampled. Details of this are given in appendix B.

The data set was censored at 50 days so that the predicted infection times could be compared to the actual infection times for subsequent IPs. With culled UIPs removed, the model was then fitted to 146 IPs and 8583 censored observations. Two chains were used with a burn-in of 5000 iterations and a further 30000 updates. The final results were thinned such that each posterior distribution consisted of 1000 samples, summary results of which are shown in table 5.1 for both the distance from source and nearest infected neighbour distance models. To aid convergence, in both cases the distances were divided by 10 before being used in the models.

After fitting, the returned convergence diagnostics were reasonable and plots of the posterior chains showed good mixing. This is further reinforced in each case by the $\hat{R}$ values, which quantify how well the chains have mixed, and the effective sample size given by n_{eff}. In the former case a value close to one indicates good mixing, with a general rule of thumb being to look for posteriors with $\hat{R} < 1.2$. The effective sample size gives the effective number of independent samples in the posterior, after accounting for autocorrelation due to the properties of the Markov chain. For a well-mixing set of chains this can usually be improved by running the model for longer and thinning more often.

The posterior distributions are significantly different from zero for both models, and in

		Mean	s.e.	2.5%	50%	97.5%	$\hat{R}$	n_{eff}
Distance from source	$\boldsymbol{\alpha}$	1.5477	0.1229	1.3150	1.5490	1.7961	1.0015	1000
	$\boldsymbol{\beta_0}$	-8.0071	0.4966	-8.9530	-7.9955	-7.0760	1.0055	1000
	$\boldsymbol{\beta_1}$	-0.0091	0.0007	-0.0107	-0.0091	-0.0077	1.0003	1000
Nearest infected neighbour	$\boldsymbol{\alpha}$	0.7744	0.0942	0.5954	0.7702	0.9607	1.0002	1000
	$\boldsymbol{\beta_0}$	-5.0139	0.4680	-5.9060	-5.0135	-4.1109	1.0005	1000
	$\boldsymbol{\beta_1}$	-0.0252	0.0026	-0.0303	-0.0252	-0.0205	1.0004	1000

Table 5.1: Posterior parameter estimates from models (5.8) and (5.9) fitted to the Devon data set at 50 days

particular the estimates for β_1, relating to the effect of the covariates on failure time, are negative in both cases. This corresponds to an increased probability of survival for premises further away from a source of infection. So the models are at least capturing some of the behaviour that we expect to see.

In order to study the predictive power of the model we choose to focus on the top ten 'most likely' future infections as predicted by the model (over a predictive period of 60 days). The mean predicted survival time and mean probability of failure over the predictive period for these premises are shown in table 5.2, and are relative to the current point, and not the beginning of the epidemic. Also, none of these values relate to observed future infected premises, for which the range of observed future infection times is between 1 and 63 days. So although all the parameter estimates are significant, neither covariate seems to capture the dynamics of the epidemic process and both models vastly overpredict the infection times.

As mentioned in section 5.3, the predictions are obtained by using the posterior samples to simulate over a finite future time period. This essentially leads to a situation where some posterior samples are 'censored' at the end of the predictive period (time E say). In order to return an interpretable *mean* predicted survival time, these 'censored' samples must be appropriately weighted since they represent predictions *greater* than E (rather than equal to E as would be the case using a traditional arithmetic mean). Here we assumed that

Distance from source		Nearest infected neighbour	
Mean survival time	**Mean probability of infection**	**Mean survival time**	**Mean probability of infection**
178	29.10%	397	14.00%
181	28.70%	407	13.70%
185	28.30%	428	13.00%
185	27.90%	462	12.20%
185	27.70%	466	12.00%
185	28.10%	470	12.10%
185	28.10%	477	11.80%
186	27.90%	479	11.80%
187	28.00%	479	11.80%
188	27.50%	480	11.80%

Table 5.2: Predictive output over a 60 day window for models (5.8) and (5.9) fitted to Devon data set at 50 days

for each premise the predictive posterior samples represented a random sample from an exponential distribution with a mean $\frac{1}{\lambda}$, and then estimated λ by maximising a likelihood function given by (4.20).

So the mean predicted survival times are influenced by the proportion of posterior samples 'observed to fail' (i.e. with values less than or equal to E). Posterior distributions with large numbers of 'unobserved' infections at the end of the predictive period will have heavily inflated mean predictions due to the weighting in (4.20). If the predictive time period is reasonably long, and a well-defined and well-fitting model is used then this should not be a problem. Indicative of the number of 'observed' infections in the posterior is the mean probability of becoming infected over the predictive period (also given in table 5.2). In distance from source model, even the most 'at-risk' premise has only approximately 30% of the predictive posterior consisting of 'observed' infections. For the nearest infected neighbour distance model this is down to approximately 14%, even though a long predictive period was used (60 days).

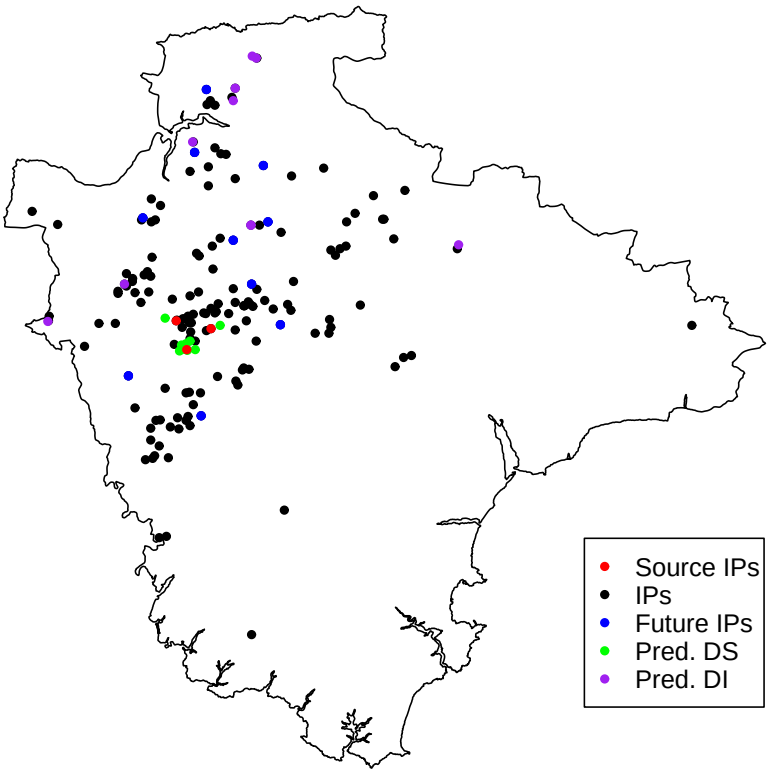


Figure 5.3: Comparison map of predictions from initial models

This posterior probability measure is a potentially useful alternative to using the predicted infection times as a means to assess risk since it can be derived directly from the posterior samples without some necessary weighting procedure. Another alternative way of summarising the results is to remove dependence on the parametric assumption used above (e.g. exponential) by instead using a non-parametric Kaplan-Meier technique to estimate the survivor curve over the predictive future period for each uninfected premise. However here this will simply result in a plateau of survival at a probability close to one for all premises.

To visualise the spatial pattern in the predictions, figure 5.3 shows the locations of the next ten actual future IPs (blue) compared to the top ten predicted from the models using distance from source (green) and nearest infected neighbour distance (purple). Also included are the source IPs (red) and the other IPs (black). It is clear from this that these covariates are not sufficient to explain the dynamics of the disease.

This is perhaps unsurprising using these very simple covariates. Neither includes contributions to the risk of infection from more than one IP at a time nor accounts for the fact that the infectiousness of IPs changes over time and in relation to the size of premise. Also there may be other factors (both spatial and non-spatial) such as susceptibility, that may vary between uninfected premises. This is apparent from figure 5.1 where there are large numbers of UIPs in-between many of the IPs, suggesting that the localised spreading process may not be solely based on proximity to nearby infected premises.

Another issue related to the latter point may simply be that there are too many premises included in the data that are not representative of the population 'at-risk' i.e. they are not exposed to the virus. It is not worth pursuing this at this stage but it will be looked at in more detail in section 5.8.

5.6 Viral load (VL)

The results in section 5.5 suggest that not enough information is being captured through the use of the distance from source or nearest infected neighbour distance covariates. This section focuses on the development of a more informative covariate, the *viral load* (VL), that uses information from infected premises to estimate the intensity of viral coverage at any spatial location at any point in time.

An animal that is infected with the FMD virus excretes different amounts of virus particles according to its species type and the length of time that it has been infected (Alexandersen et al. 2003a). In adult animals the disease is rarely fatal and animals can recover on their own without treatment - therefore as the length of time increases the amount of virus excreted by an infected animal will tend to zero. Modelling the distribution and magnitude of viral excretion of infected herds over time is an important aspect in developing a viable covariate to help model the spread of the disease.

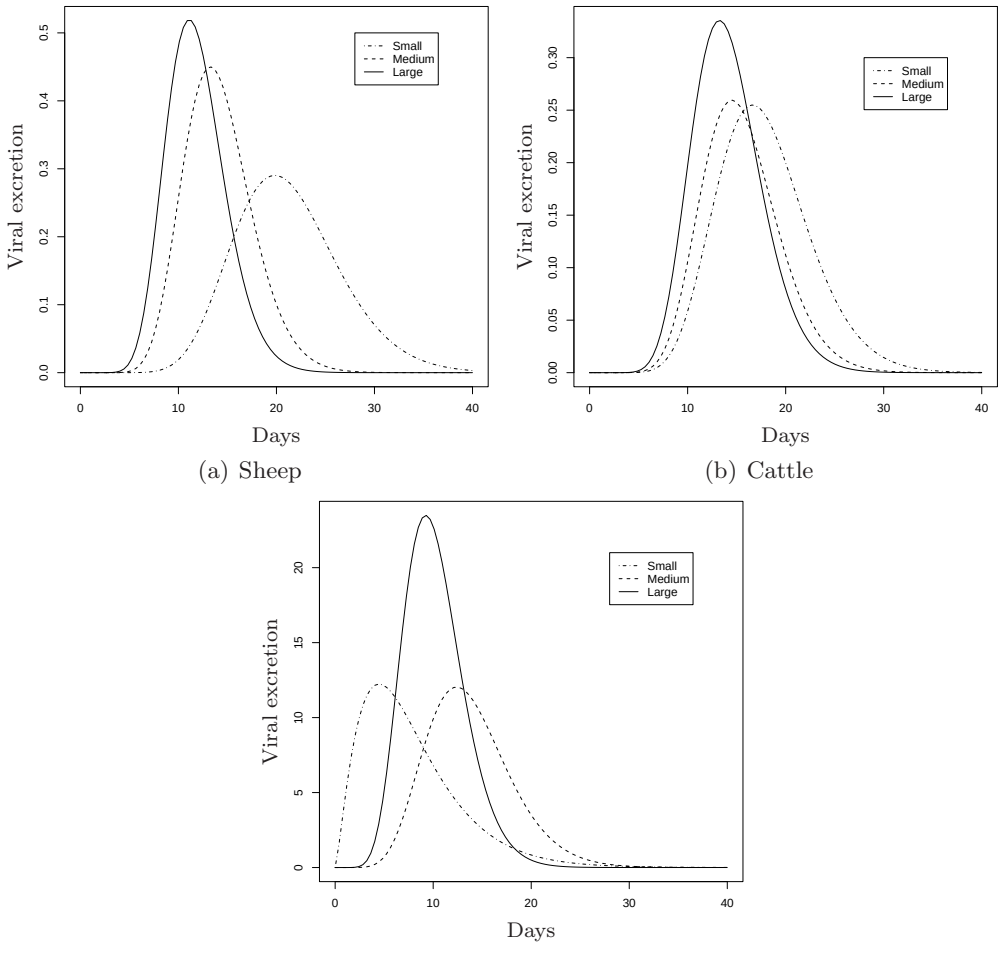
5.6.1 Infectivity functions

Some exploratory work done by VLA (Arnold 2005) produced empirical estimates of the total viral excretion at the end of the epidemic for herds of varying type and size. These figures were calculated by assuming an SEIR differential equation model for the within-herd transmission of the disease - assuming one initially infected animal - and then solving the resulting system of equations using a Runge-Kutta algorithm and summing up the infectivity of each animal at each time step. Further investigations revealed that various two-parameter gamma curves provide good fits to the empirical distributions of within-herd spread of the disease over time, with different values of the parameters dependent on the size and species of infected animal. Unfortunately these estimated curves were not available for all herd sizes, but the parameter estimates were provided for the median, lower and upper quartiles of herd sizes in Devon for each species (sheep, pigs and cattle).

Both the shape and scale of the herd infectivity changes with respect to different herd sizes. It was therefore decided to develop a measure of infectivity for an IP based upon these gamma curves by classifying each premise into small, medium or large with respect to the herd size of each species present, and then to use the parameter estimates provided for the lower, median and upper quartiles of each species (shown in table 5.3) to determine the shape of the infectivity curve over time. These *scaled infectivity functions* could then be multiplied by the actual number of animals present in the herd, and for each IP these values could be summed over all species to give an overall measure of viral excretion at any point in time.

Plots of the scaled infectivity functions are given in figure 5.4, with corresponding associated parameter estimates in table 5.3. It can be seen that pigs excrete far more virus than cattle or sheep. This echoes the experimental results of Alexandersen et al. (2003a). However the herd sizes for pig herds are generally much smaller than those for cattle or sheep (see table 5.4).

So the functional form for viral excretion at time t after infection, for a herd of species



(a) Sheep

(b) Cattle

(c) Pigs

Figure 5.4: Estimated scaled infectivity curves for different relative herd sizes in Devon by species type

	Herd size (H)	γ	ξ	η
Sheep	$H \leqslant 231$	15.38	1.38	3.83
	$231 < H \leqslant 601$	16.42	0.86	3.83
	$H > 601$	15.43	0.77	3.83
Cattle	$H \leqslant 81$	14.41	1.24	2.92
	$81 < H \leqslant 135$	15.66	0.99	2.92
	$H > 135$	15.72	0.90	2.92
Pigs	$H \leqslant 2$	2.36	3.29	125
	$2 < H \leqslant 26$	10.11	1.36	175
	$H > 26$	10.83	0.94	175

Table 5.3: Parameter estimates for scaled infectivity functions

	Lower	**Median**	**Upper**
Sheep	231	601	957
Cattle	80	135	206
Pigs	2	26	176

Table 5.4: Infected herd size information for Devon in 2001

k ($k = 1, \ldots, 3$ representing sheep, cattle and pigs) with relative herd size l ($l = 1, \ldots, 3$ representing small, medium and large) is given by the scaled *infectivity function* ($\mathcal{I}_{kl}$) where,

$$\mathcal{I}_{kl}(t) = \frac{\rho_{kl}}{\xi_{kl}^{\gamma_{kl}}\Gamma(\gamma_{kl})} t^{\gamma_{kl}-1} e^{\frac{-t}{\xi_{kl}}}. \tag{5.10}$$

The shape and scale are given by γ_{kl} and ξ_{kl} respectively and ρ_{kl} is

$$\rho_{kl} = \begin{cases} \eta_{kl} & \text{if relative herd size of species } k \text{ is } l, \\ 0 & \text{otherwise.} \end{cases}$$

Here the η_{kl} are constants of proportionality.

5.6.2 Viral load at a premise

For each species type the scaled infectivity functions essentially quantify the amount of virus excreted over time from an infected animal in a certain herd size. However caution is needed in this interpretation, since they should not be read as measures of viral excretion from *a single* infected animal, since buried in their definitions are additional factors relating to the relative temporal rate of infection between different herd sizes. These quantities are only really valid when scaled and applied to an entire herd.

To develop a measure for the total amount of virus acting at an arbitrary spatial location it is important to decide how the infection is dispersed over the entire spatial region. A key aspect of this particular form of smoothing is that when the smoothed values are integrated over space, the value of the integral should equal the total amount of virus excreted by all IPs.

This precludes the use of traditional spatial smoothing methods such as kernel regression, localised regression and spline smoothing, since they attempt to measure the *mean* value over space, and the smoothed mean would reflect the average infectivity *produced*, not the average infectivity *acting on* that location.

Instead we will use a modified form of kernel regression directed towards estimating the total amount of the virus *per unit area*. A bivariate kernel function will be used, centred on the point location of an IP and constrained so that the total volume under the curve equals one. The density of virus generated from each IP acting at any arbitrary point location can be calculated by taking the value of the kernel function at that point and multiplying the result by the total infectivity produced from the corresponding IP. The measure of viral load acting at any point location is equal to the sum of these weighted densities from all IPs.

The distributional form of the weighting function is then very important, as it not only controls the degree of smoothing but also the shape. We decided to use a bivariate normal distribution for this, since it has the properties required above. The variance, σ^2, controls

the amount of smoothing required and correct choice of σ^2 is important. If the weighting is too smooth it may not reflect accurately the importance of proximity to sources of infection, and if not smooth enough may preclude the possibility of the virus spreading.

Hence the VL at any spatial location, $\boldsymbol{s}$, is defined as the total amount of virus per unit area acting at $\boldsymbol{s}$ at time t, and is the sum across all IPs of the kernel-weighted infectivity functions (5.10), i.e.

$$\mathrm{VL}(\boldsymbol{s}, t) = \sum_{j|t_j<t} \left[\sum_k \left\{ n_{jk} \sum_l \mathcal{I}_{kl}(t - t_j) \right\} \right] \omega(\boldsymbol{s}, \boldsymbol{s}_j, \boldsymbol{\Sigma}), \tag{5.11}$$

where n_{jk} is the total number of animals of species k on the j^{th} IP located at $\mathbf{s}_j$, $(j = 1, \ldots, J)$ with infection time t_j. A suitable spatial smoothing function with bandwidth parameters $\boldsymbol{\Sigma}$ is given by $\omega(\cdot)$.

Here $\boldsymbol{s} = (s_1, s_2)$ where s_1 corresponds to easting and s_2 to the northing - though actually these are (x, y)-coordinates derived from latitude and longitude, since at this geographical scale the curvature of the earth is considered negligible. The smoothing function, $\omega(\cdot)$, is a bivariate normal distribution such that:

$$\omega(\boldsymbol{s}, \boldsymbol{s}_j, \boldsymbol{\Sigma}) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(\frac{-z}{2(1-\rho^2)}\right), \tag{5.12}$$

where

$$z = \frac{(s_1 - s_{j1})^2}{\sigma_1^2} - \frac{2\rho(s_1 - s_{j1})(s_2 - s_{j2})}{\sigma_1\sigma_2} + \frac{(s_2 - s_{j2})^2}{\sigma_2^2}, \tag{5.13}$$

and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

The correlation between easting and northing is given by ρ. In the absence of topographical or meteorological data that could help to identify directional behaviour in the local spreading process we decided that in the first instance the smoothing process should be assumed isotropic over space (i.e. covariance of 0), however anisotropy could potentially

be incorporated by changing the covariance matrix to reflect the degree of dependence required. In the isotropic case $\sigma_1 = \sigma_2 = \sigma$, and $\rho = 0$. Therefore,

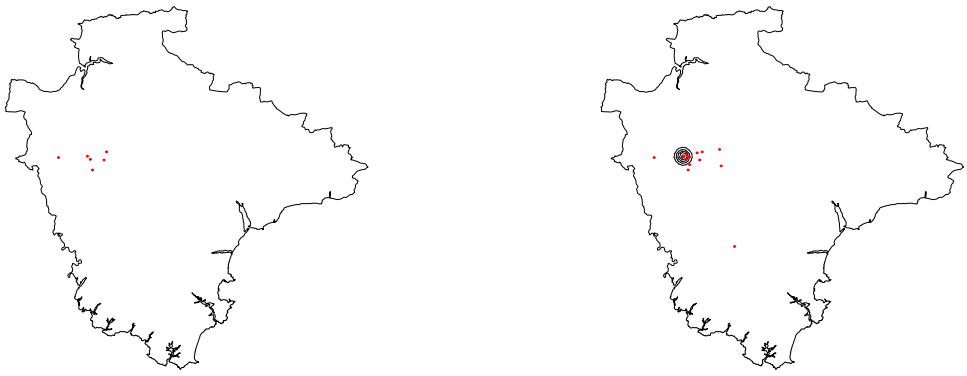
$$\omega(\boldsymbol{s}, \boldsymbol{s}_j, \boldsymbol{\Sigma}) = \omega(d_j, \sigma^2) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{d_j^2}{2\sigma^2}\right). \tag{5.14}$$

where d_j is the distance between premise $\boldsymbol{s}$ and premise $\boldsymbol{s}_j$. Some spatial maps of the viral load over time are shown in figure 5.5, with IPs indicated by red points.

Some discussion with collaborators at VLA suggested that in the Devon epidemic it was unlikely that localised spreading occurred over a range greater than 3km from an infected premise and so the bandwidth was fixed such that the kernel smoothing function had an effective radius of 3km. Also suggested was that the cumulative VL (denoted CV) over time might be a better covariate to use since there may be some temporal lag in the effect of the viral excretion (e.g. residual virus remaining from previous points in time). However a problem is that this results in a monotonically increasing function of viral load over time that has no capacity to reduce if sources of the virus in surrounding areas are removed. To combat this it was decided to use the average cumulative VL (denoted AV) as a covariate instead. This has the effect that it encompasses the temporal lag but also reduces over time during periods of no viral excretion in nearby IPs (shown in figure 5.6).

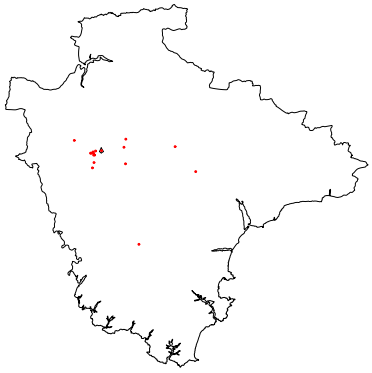
5.7 Results of model fitted with AV as covariate

The model likelihood takes the same form as (5.9) except this time $\lambda_{it} = \exp(\beta_0 + \beta_1 \mathrm{AV}_{it})$, where AV_{it} is the average cumulative viral load acting at location i at time t. The regression parameter β_0 was given a $N(0, 100)$ prior and β_1 a $N(0, 10)$ prior. The shape parameter α was given a $G(0.1, 10)$ prior (i.e. a mean of 1 and a variance of 10). Initial values were generated in the same way as before and two chains were used, again with a burn-in of 5000 iterations and a further 30000 updates, and with the final results thinned so that each posterior distribution consisted of 1000 samples.

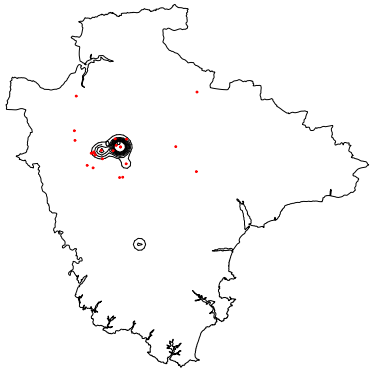


(a) Day 5

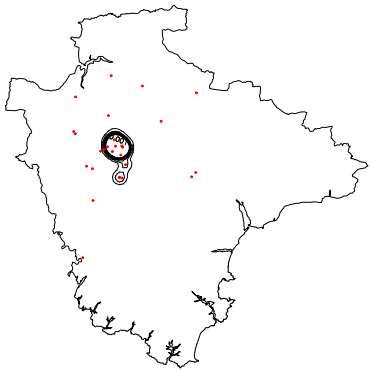
(b) Day 10



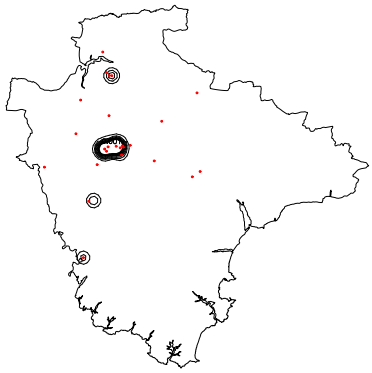
(c) Day 15



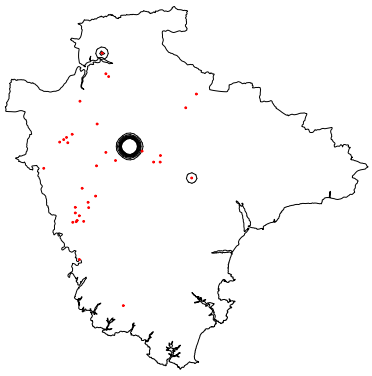
(d) Day 20



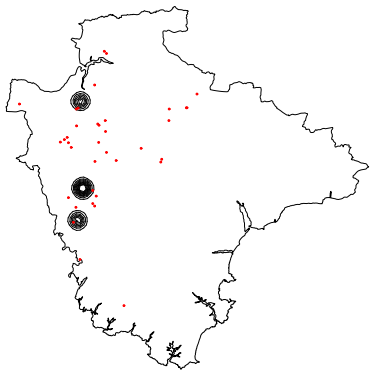
(e) Day 25



(f) Day 30

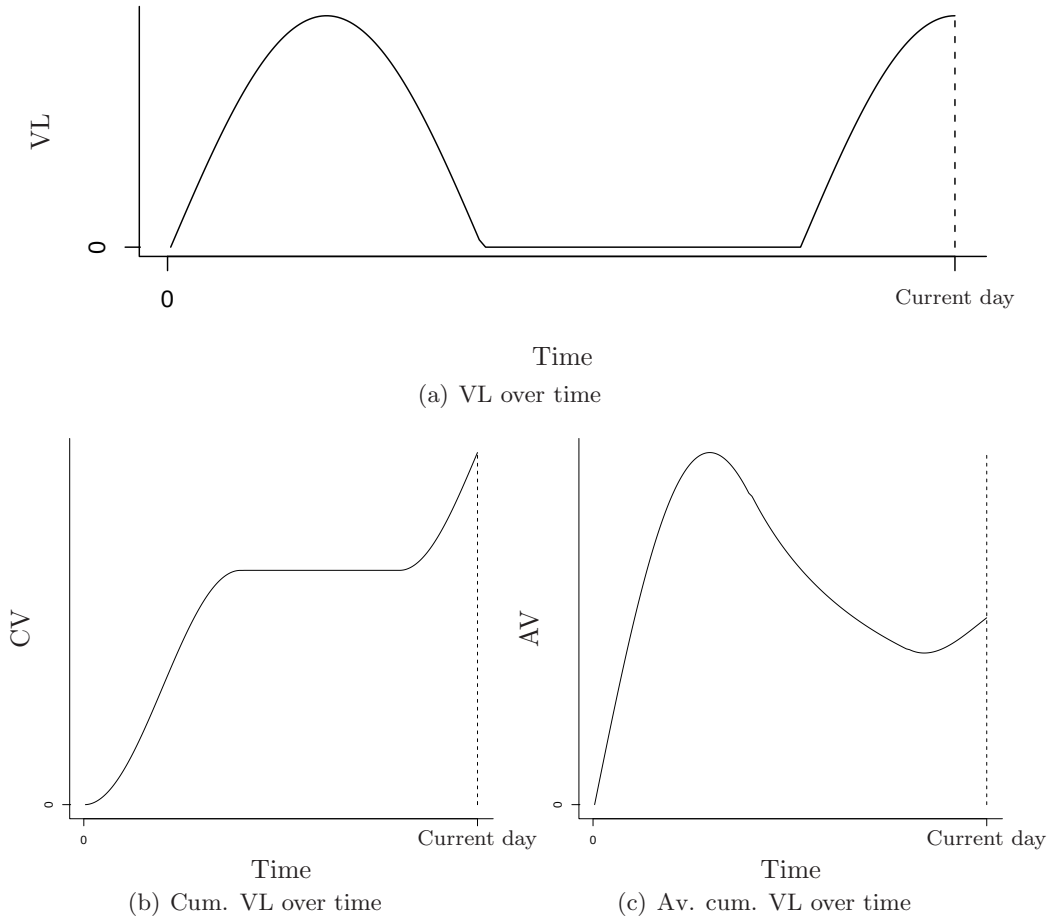


(g) Day 35



(h) Day 40

Figure 5.5: Spatial maps of viral load over time with 3km effective bandwidth



(a) VL over time

(b) Cum. VL over time

(c) Av. cum. VL over time

Figure 5.6: Theoretical VL, cumulative and average cumulative VL plots over time

	Mean	**s.e.**	**2.5%**	**50%**	**97.5%**	$\hat{\boldsymbol{R}}$	$\boldsymbol{n}_{\text{eff}}$
$\boldsymbol{\alpha}$	1.4919	0.1201	1.2619	1.4860	1.7381	1.0016	930
$\boldsymbol{\beta_0}$	-9.8749	0.4774	-10.8400	-9.8510	-8.9668	1.0019	810
$\boldsymbol{\beta_1}$	0.3198	3.1234	-5.8653	0.4691	6.5145	1.0017	870

Table 5.5: Posterior parameter estimates from model (5.9) with AV covariate fitted to Devon data set at 50 days

Mean survival time	Mean probability of infection
953	6.10%
953	6.10%
969	6.00%
974	6.00%
987	5.90%
988	5.90%
989	5.90%
1002	5.80%
1003	5.80%
1004	5.80%

Table 5.6: Predictive output over a 60 day window for model (5.9) with AV covariate fitted to Devon data set at 50 days

The parameter estimates from the model fitted to the full data set at 50 days are shown in table 5.5 and the predictive summary for the ten most likely future infections in table 5.6. The convergence diagnostics are good but the parameter estimate for the effect of the AV covariate is not significantly different to zero, suggesting that the covariate is not accurately capturing the variability in the disease spread. It is no surprise then that the model vastly overpredicts the survival times, even more so than for the simpler covariates used in section 5.4. A possible reason for this is that there is a large amount of confounding information in the data set, and this issue will be explored in the next section.

5.8 Exposure and censoring

The results shown in section 5.7 indicate that the predicted infection times from the model fit are much larger than the actual infection times. An intuitive reason for this overprediction could be that the population 'at-risk' is not representative - e.g. that there are many premises included in the model fit that do not contribute any useful information to the likelihood. There are over 8500 censored observations but only 146 infected observations

after 50 days in the Devon epidemic. In essence the censored information is 'swamping' the model and confounding the parameter estimates.

An interesting problem associated with modelling contagious epidemics is that exposure to the virus changes over space as well as time. Figure 5.1 shows the spatial distribution of the Devon premises, suggesting that there are large areas of the county that are not near to sources of infection. This ties in with the idea that those premises not exposed to the virus in any way will not be susceptible to infection from localised spread. If there are many censored observations included in the model that are not exposed to the disease then this will heavily bias the parameter estimates and lead to overprediction of the survival times.

Since the focus here is concerned with the localised spread of an infectious disease, these 'unexposed' premises may potentially be contributing little or no useful information to the likelihood. One way in which this problem could be tackled is to consider censoring the data via some measure of exposure to the virus; then for any point in time at which the model is fitted, only those premises 'exposed' to the virus are included. All that is needed is a method for determining exposure to the virus, and this can be achieved by considering VL.

VL is a measure of viral activity at a spatial location and can be calculated at any point in time up to current point in the epidemic. A logical step would be to classify premises as 'exposed' to the virus at time t if the VL value at that point exceeds a pre-determined threshold value. This effectively targets the model at those premises deemed most 'at-risk' from localised infection.

There are issues with this procedure however - for example in the case of premises that move into and out of exposure over time this results in multiple recorded censoring times for the same premise, though using average VL ensures that the covariate values for premises that fall out of exposure decrease over time, since if at any point the VL for an exposed premise falls below the threshold value then it is simply treated as zero in the cumulation. We decided that once a premise becomes exposed it remains exposed and the use of AV

results in the covariate value reducing over time when potential sources of infection are removed. Another important point is that this changes the focus of the survival model from modelling absolute time from the beginning of the epidemic to relative time from exposure. This is a desirable feature of this approach since it has a more reasonable biological interpretation as an infectious process, particularly since premises do not begin the epidemic in the same state as each other with regards to viral exposure. For example in the traditional setting, two premises with identical viral loads would be expected to have the same predicted mean survival time; however the point in the epidemic at which they attain that level of exposure will determine the absolute survival time. They are only comparable relative to exposure.

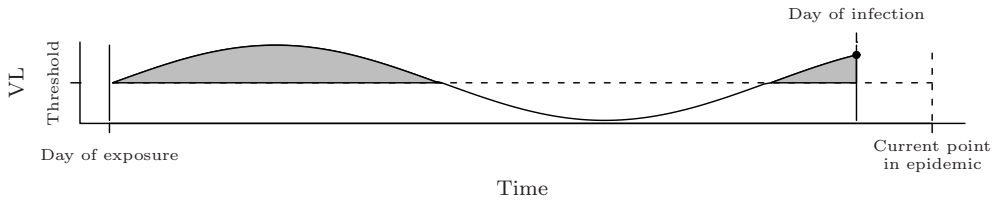
Figure 5.7 gives some examples of how the concept works for different premises. The shaded regions show the areas over which the VL is cumulated and this is averaged by dividing by the date of infection/censoring minus the date of initial exposure.

This leads to a situation where not all IPs are involved directly in the model fit since non-exposed IPs would have a relative survival time of zero. Information from these premises is still (indirectly) included through the AV covariate, but care must be taken with the determination of a reasonable threshold to reflect the degree of non-exposed infections believed to exist in the data.

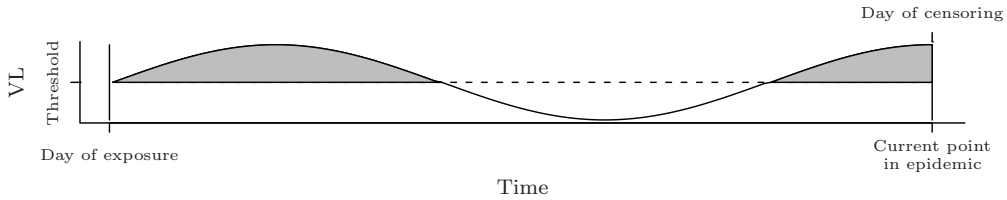
The predictive algorithm given in section 5.3 can be adapted to deal with exposure. To do this consider that exposure at time ν is determined by applying a threshold w^* to a time-dependent covariate $w(\nu)$. The covariate $x(\nu)$ that drives the epidemic is then given by a function $g(w(\nu))$.

Let δ^e_{ik} $(i = 1, \ldots, n)$ be an indicator variable for an uninfected individual such that

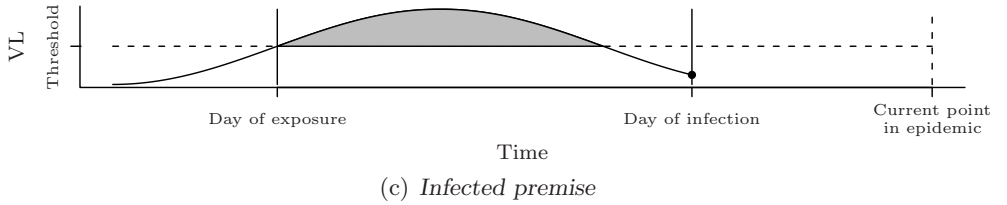
$$\delta^e_{ik} = \begin{cases} 1 & \text{if } \exists\ w_i(u), u = 0, \ldots, t, \text{ such that } w_i(u) - w^* > 0, \\ 0 & \text{otherwise,} \end{cases}$$



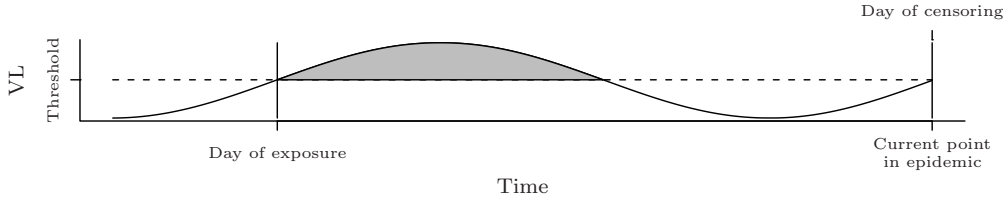
(a) *Infected premise*



(b) *Censored premise*



(c) *Infected premise*



(d) *Censored premise*

Figure 5.7: Theoretical threshold and exposure based on VL

and $\boldsymbol{T}^e$ be a matrix of exposure times with elements $\{t^e_{ik}\}$ such that initially

$$t^e_{ik} = \begin{cases} \inf\{u : w_i(u) - w^* > 0; u = 0, \ldots, t\} & \text{if } \delta^e_{i\cdot} = 1, \\ 0 & \text{otherwise.} \end{cases}$$

i.e. t^e_{ik} is the date of exposure for all currently exposed premises at time t.

Step 2 in the predictive algorithm then changes such that,

2. Take k^{th} set of posterior samples and calculate $w_i(\nu)$ (using $\text{IP}^{(\nu)}$ if necessary). If $w_i(\nu) - w^* > 0$ and $\delta^e_{ik} = 0$ set $\delta^e_{ik} = 1$ and $t^e_{ik} = \nu$. Calculate $x_i(\nu) = g(w_i(\nu))$ and

$$h_i(\nu \mid x_i(\nu)) = \begin{cases} P(\nu - t^e_{ik} \leqslant T < \nu - t^e_{ik} + 1 \mid T \geqslant \nu - t^e_{ik}, x_i(\nu)) & \text{if } \delta^e_{ik} = 1, \\ 0 & \text{otherwise,} \end{cases}$$

for all uninfected premises (using $\text{IP}^{(\nu)}$ if necessary).

5.9 Results for AV model fitted to data censored via exposure

The data set was censored using a threshold value of 5×10^{-08} at 50 days, leaving 110 IPs and 4079 UIPs. The model setup (e.g. priors, burn-in etc.) were specified as for the full model in section 5.7 and summaries of the posterior distributions are given in table 5.7. Predictive summaries are given in table 5.8.

Again the convergence diagnostics and mixing are good, but the β_1 parameter relating to the effect of the average cumulative viral load is not significantly different from zero. The point predictions are also large, though arguably better than for the full data set. Clearly exposure is not the sole source of the variation in the data, and so we need to identify other potential reasons for the overprediction.

	Mean	**s.e.**	**2.5%**	**50%**	**97.5%**	$\hat{R}$	n_{eff}
$\boldsymbol{\alpha}$	1.0407	0.0787	0.8974	1.0345	1.2021	1.0000	1000
$\boldsymbol{\beta_0}$	-6.8079	0.2707	-7.3550	-6.7960	-6.3220	1.0005	1000
$\boldsymbol{\beta_1}$	0.3802	3.1116	-5.8504	0.4317	6.6829	1.0045	330

Table 5.7: Posterior parameter estimates from model (5.9) with AV covariate fitted to Devon data set 'censored via exposure' at 50 days

Mean survival time	**Mean probability of infection**
520	10.90%
526	10.80%
526	10.80%
537	10.60%
545	10.40%
547	10.40%
549	10.30%
552	10.30%
553	10.30%
554	10.30%

Table 5.8: Predictive output over a 60 day window for model (5.9) with AV covariate fitted to Devon data set 'censored via exposure' at 50 days

One of the problems associated with censoring via exposure is that unexposed infected premises are not included in the model fit directly. The threshold is estimated by trying to reduce the total number of premises in the model whilst keeping as many IPs as possible. In the above case where the threshold was 5×10^{-08}, this resulted in 36 infected premises being left out of the model. In an epidemic situation, where events are relatively rare, this can constitute a large amount of information on the spreading process that is not being used in the estimation of the parameters. Of course if, as believed, these premises did not become infected through a localised process, then this is perhaps not unreasonable given that the model is assessing localised spread of the disease. However these so-called 'spark' or non-localised infections will still contribute to the epidemic process through viral excretion, and their effect in this case is felt through the VL covariate, since it measures the amount of viral pressure per unit area acting at a point location obtained from *all* infected premises over time.

Figure 5.8 shows the spatial distribution of premises in Devon at 50 days after being censored via exposure with a threshold of 5×10^{-08}. An example of the way that the VL captures some of the impact of 'spark' infections can be seen towards the east and south of the region, where proximity to non-localised IPs has resulted in stand-alone areas of UIPs being included in the model without a nearby source IP (since the corresponding IPs were *not exposed* at the point of infection). What is also clear from this figure is that although the modelling has been targeted at areas of higher-risk, the IPs are still fairly sparse in places, suggesting that there may be some other factors relating to the relative propensity of premises to become infected given the same level of exposure. Two main possibilities emerge: either premises exhibit varying resistance to localised spread, or the model is being confounded by a non-localised process. For the moment we concentrate on the former.

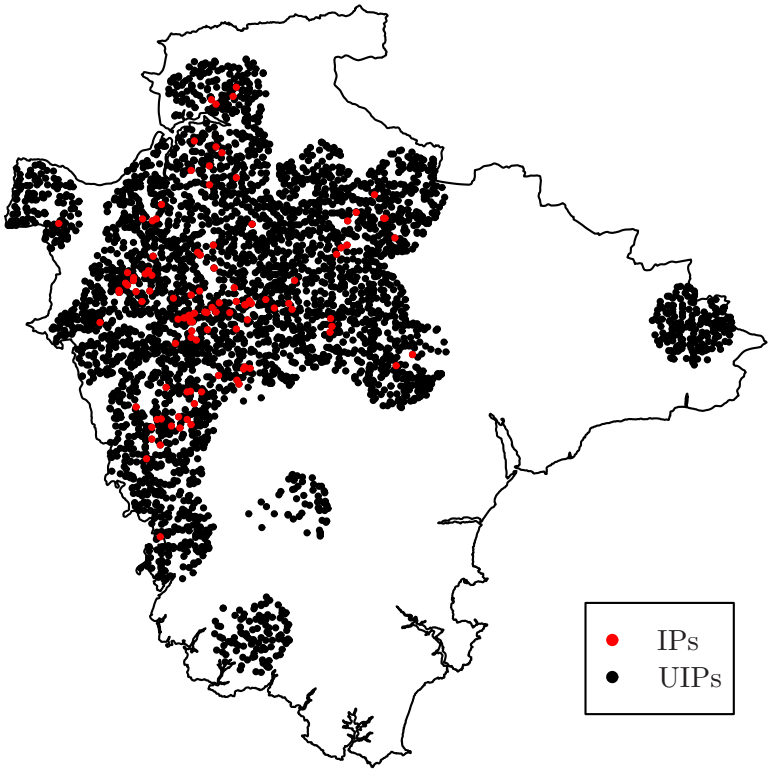


Figure 5.8: Spatial distribution of premises in Devon 'censored via exposure' at 50 days

5.10 Susceptibility to infection

Various papers (Keeling et al. 2001a, Deardon et al. 2006, Alexandersen et al. 2003a) have reported the importance of variable susceptibility with regard to herds of different sizes and species type. To this end two other models were fitted that include covariates that act as surrogates for susceptibility. The first used uninfected animal density, but did not differentiate between species, and the second modelled uninfected sheep, cattle and pig densities separately. The densities were calculated using the same bandwidth and kernel function as was used in the calculation of the viral load and the covariates were included through a log link in the scale parameter λ.

Hence λ_t in each case is given by:

$$\lambda_{it} = \exp(\beta_0 + \beta_1 \mathrm{AV}_{it} + \beta_2 \mathrm{AD}_{it}),$$

when using uninfected animal density (AD) as the susceptibility covariate, and

$$\lambda_{it} = \exp(\beta_0 + \beta_1 \mathrm{AV}_{it} + \beta_2 \mathrm{SD}_{it} + \beta_3 \mathrm{CD}_{it} + \beta_4 \mathrm{PD}_{it}),$$

when using species-specific densities for sheep (SD), cattle (CD) and pigs (PD).

The priors were given as for the previous models e.g. the shape parameter $\alpha \sim G(0.1, 10)$, the intercept β_0 $N(0, 100)$ and the regression parameter, β_1, relating to the effect of AV was assumed $N(0, 100)$. The parameters β_2, β_3 and β_4 were each given $N(0, 100)$ priors, and the MCMC chains run as before with a burn-in of 5000 iterations and a further 30000 updates. The posteriors were thinned to return 1000 samples.

Both models were fitted to the data set 'censored via exposure' at 50 days, and good convergence diagnostics were returned. The parameter estimates in each case are shown in table 5.9. Looking at the 2.5% and 97.5% percentiles indicate that the only covariate to have a non-zero effect on survival time is uninfected animal density (AD), suggesting that the effect of increasing AD is to increase the risk of infection. These results do not seem to tie in with those found in some other studies of foot-and-mouth disease, both with regards to experimental data (Alexandersen et al. 2003a) or actual epidemic data from the 2001 UK outbreak (e.g. Keeling et al. 2001a, Tildesley et al. 2006).

Focussing on the predicted survival times (table 5.10), we can again see that the predictions are poor and the infection times are overpredicted, with very few posterior samples returning an infection date less than 60 days into the future (14.3% in the highest case). This seems to suggest that the model is not correctly accounting for susceptibility, possibly because the model is assuming the effect of susceptibility on the hazard is related to scale only. Instead it may be that the actual shape of the survival distribution changes for different herds dependent on their susceptibility status, and this will be explored in the next chapter.

		Mean	s.e.	2.5%	50%	97.5%	$\hat{R}$	n_{eff}
Uninfected animal density	$\boldsymbol{\alpha}$	1.1344	0.0938	0.9611	1.1310	1.3240	1.0008	1000
	$\boldsymbol{\beta_0}$	-7.4321	0.3540	-8.1490	-7.4320	-6.7590	1.0001	1000
	$\boldsymbol{\beta_1}$	0.4545	3.2206	-5.7773	0.5276	6.7024	1.0022	690
	$\boldsymbol{\beta_2}$	6.9176	2.7698	1.4122	6.9125	12.1603	1.0003	1000
Uninfected species densities	$\boldsymbol{\alpha}$	1.1270	0.0974	0.9434	1.1240	1.3221	1.0006	1000
	$\boldsymbol{\beta_0}$	-7.2719	0.3592	-7.9884	-7.2660	-6.5958	1.0013	1000
	$\boldsymbol{\beta_1}$	0.5047	3.1682	-5.9030	0.5907	6.7280	1.0023	670
	$\boldsymbol{\beta_2}$	4.1790	2.8104	-1.3295	4.1165	9.5584	1.0001	1000
	$\boldsymbol{\beta_3}$	1.7190	3.1322	-4.4362	1.7105	7.8110	1.0016	960
	$\boldsymbol{\beta_4}$	2.3223	2.9623	-3.4215	2.2590	8.2260	1.0002	1000

Table 5.9: Posterior parameter estimates from models (5.9) with AV and additional susceptibility covariates fitted to the Devon data set censored via exposure at 50 days

Uninfected animal density		**Uninfected species densities**	
Mean survival time	**Mean probability of infection**	**Mean survival time**	**Mean probability of infection**
392	14.30%	454	12.40%
441	12.80%	464	12.20%
446	12.60%	470	12.00%
454	12.40%	481	11.80%
465	12.10%	483	11.70%
467	12.10%	487	11.60%
473	12.00%	488	11.60%
477	11.80%	488	11.60%
477	11.80%	491	11.50%
479	11.80%	501	11.30%

Table 5.10: Predictive output over a 60 day window for model (5.9) with AV and additional susceptibility covariates fitted to Devon data set at 50 days

5.11 Conclusions

In this chapter we have fitted a series of models to the Devon data set. Some initial exploratory models using distance from source infections and nearest infected neighbour distance over time failed to capture the dynamics of the localised spreading process. We then developed a viral load covariate that measured the amount of virus per unit area acting on a point location over time. These were based upon work done by collaborators at VLA fitting deterministic SEIR models to the within-herd spread of the disease for different herd sizes and species type. However this covariate also failed to capture the dynamics of the disease.

The viral load can be used to target the survival modelling to those areas with high viral coverage, in this way premises that may be confounding the parameter estimates due to the fact that they are not exposed and therefore contribute no useful information to the likelihood can be removed. This still left confounding factors that were not being accounted for, in particular 'spark' (non-localised) infections and premise varying susceptibility. Some of the extra variation due to these factors may be able to be captured through the inclusion of random effects (see section 4.5.6), however this will not allow us to determine what drives the competing processes or predict to premises not included in the model fit (e.g. due to censoring via exposure). Instead we need to examine different ways in which we can model these separate processes.

The possibility of 'spark' infections in the data set caused by means other than localised spread is backed up extensively in the literature, see e.g. Ferguson et al. (2001a,b), Keeling et al. (2001a) and Tildesley et al. (2006). In the case of FMD there are many potential sources of infection, and although the implementation of control measures such as housing of livestock and increased biosecurity help to alleviate some of the risk of disease spread, they do not confer complete immunity to premise infection. It is in this situation that knowledge from the Cattle Traceability System data (Vernon et al. 2005) would be useful if it could be effectively incorporated into the model structure. In order to model this various assumptions need to be made about the underlying process governing the spread of the

disease. When observations include knowledge of the cause of failure then a competing risks approach could be used, where the risk to each individual is treated as a combination of independent survival processes corresponding to each potential cause of failure (localised or 'spark').

The basic form of a competing risks model assumes all individuals are susceptible to infection from multiple causes of failure, and treats failures from each cause as censored with respect to the others. Two problems associated with this approach lie in the assumption of independence between the different failure processes and the requirement to observe the explicit means of failure. Adaptations to incorporate missing data exist but are not robust for large amounts of unobserved data. An alternative is to use a mixture model in which the overall survival process for an individual is modelled as a weighted combination of risks from each of the competing causes. This type of approach will be discussed in more detail in the next chapter, albeit in a slightly different context.

Although it is useful to note the potential use of both mixture models and competing risks models in the situations where different causes of failure occur, in the case of FMD the number of occurrences of spark infections is thought to be small in comparison to the number of localised infections. In addition, unexposed infections are removed from the data set through censoring via exposure, and although not included directly, information from these premises is incorporated into the model through the viral load covariate. Frailty effects could be introduced into the model to account for unobserved heterogeneity, but they do not contain information about factors affecting the spreading process and the effect on the predicted infection times would be to tighten up the error around the predictive mean. The inaccuracy of the predictive results from the models fitted in this chapter simply do not justify their use in this case.

It is therefore surmised that the overprediction of the survival times noted in the models fitted so far is most likely to be caused by premise-varying susceptibility. Differences in susceptibility between species and premises is well documented. Including uninfected animal density as a surrogate for susceptibility as an extra covariate in the hazard did

give a significant parameter estimate but did not vastly improve the predictions, although the species-specific densities did not. Subsequent chapters will look at possible alternative ways to model this and will use a series of simulation studies to explore the effects of resistance on an epidemic process, and to compare the relative advantages and disadvantages of various approaches used to model this.

Chapter 6

Modelling resistance to infection

In the previous chapter two sources of heterogeneity were identified as potential confounding factors that were not being accounted for in the modelling process, those of premise-varying susceptibility and the possible presence of non-localised infections. In this chapter we focus on alternative modelling approaches to capture premise-varying susceptibility.

Extensions to the conventional model that can be used in these situations include long-term survivor and mixture models (see chapter 4). In fact both these approaches are linked and the choice between them is greatly dependent on how the confounding factors are thought to affect the overall epidemic process. Section 6.1 will discuss some of the biological and epidemiological considerations regarding infectious animal diseases that can influence the choice of modelling strategy (with focus on FMD in particular).

Section 6.2 considers various candidate models, gives relevant mathematical detail, and discusses particular issues, advantages and deficiencies in their application. Sections 6.3 and 6.4 provide results and predictions from a simple simulation experiment to show comparisons between the different approaches in the context of modelling resistance to infection. Finally some conclusions are given in section 6.5.

6.1 Considerations in FMD and other animal diseases

There is strong evidence in the literature to support varying degrees of susceptibility between premises (e.g. Alexandersen et al. 2003a, Ferguson et al. 2001b, Keeling et al. 2001a, Gloster et al. 2005). We consider two main attributes affecting premise susceptibility. The first is concerned with differences between animal species. Alexandersen et al. (2003a) in particular investigate the excretion and transmission of FMD in pigs and cattle experimentally infected with the disease. They provide significant evidence to support the fact that although pigs excrete far more virus particles than sheep or cattle once infected, they are much less susceptible to contracting the disease through airborne infection. This is further backed up in Ferguson et al. (2001b) and Arnold (2005). Species differences in excretion of the disease is already accounted for in the VL covariate but differences in susceptibility have only been considered in a relatively simple manner so far.

The second factor corresponds to all other aspects of intra-premise heterogeneity, such as differences in farming practices, control policies and the physical attributes of each premise e.g. varying farm biosecurity and surrounding geographical topography (Keeling et al. 2001a). These effects could also be time-varying, particularly with reference to movement restrictions and other control policies that are implemented at different points in the epidemic.

In the context of a wider contagious animal disease epidemic we consider a situation where each individual farm premise exhibits some form of generic *resistance to infection* - for FMD this encompasses contributions from both differences in susceptibility caused by species type as well as varying premise level attributes. A sensible interpretation of resistance to infection would be that the hazard of localised infection for premises deemed 'resistant' would be smaller than that for 'susceptible' premises.

In the previous chapter resistance was directly incorporated into the hazard through the use of a specified covariate, affecting the relative probability of infection between premises at each time point but not the shape of the underlying baseline hazard function. An

alternative would be to consider that there are multiple survival processes governing the spread of the disease, and that the resistance measure controls the magnitude of susceptibility to infection from each competing process. A natural way to model this is to allow the overall hazard for an individual premise to be represented by a weighted combination of hazards, with the weights dependent on the probability of resistance. This allows susceptibility to affect not just the scale but also the shape of the corresponding survival distributions. In this chapter we will investigate two main extensions to the conventional survival model specification that provide tractable ways to do this - those of long-term survivor and mixture models.

6.2 Candidate models

We focus on two candidate models, the standard mixture model and a special case known as the long-term survivor model (see chapter 4). In general the long-term survivor framework is used when it is believed that there is a proportion of the population that is 'immune' to failure from the cause of interest. This type of model is commonly known as a 'cure rate' model, relating to situations in medical statistics where the proportion of the population is deemed 'cured' of failure from the cause of interest. In the case of modelling resistance to infection in individual farm premises, the long-term survivor approach treats all 'resistant' premises as *immune* with regard to transmission through localised means.

The mixture model approach extends this to allow the resistant group to experience failure. That is it considers that individuals from a population are subject to potential failure from multiple survival processes. It is worth noting in both these cases that although non-localised infections are not being modelled directly, information from unexposed infected premises is still included in the model through the VL covariate, since it has the ability to update on a day-by-day basis and is based on viral excretion from *all* IPs.

6.2.1 Mixture models

The use of mixture models in statistical analyses is well documented in the literature. One of the earliest recorded uses of this approach is found in Pearson (1894), in which he fitted a mixture of two normal distributions to measurements on the ratio of forehead to body length of 1000 crabs. The text by McLachlan and Peel (2000) provides a good introduction to modelling finite mixture models and includes a brief history of the field, as well as citing more comprehensive bibliographies and review articles (see McLachlan and Peel 2000, section 1.18).

The benchmark technique for fitting mixture models via maximum likelihood is to use the EM (expectation-maximisation) algorithm (Dempster et al. 1977). This was the first really practical alternative to the computationally intensive method of moments approach used by Pearson (1894), and opened the way for more complex mixtures to be considered. However recently Bayesian methodology and MCMC in particular are becoming more popular (see Titterington et al. 1985, McLachlan and Basford 1988 and McLachlan and Peel 2000). An intensive review of mixture models is not the intention here, instead an overview of Bayesian methodology for the application of the mixture approach to survival modelling will be given and some newer developments discussed.

Other recent papers on using mixture models in a Bayesian framework can be found in e.g. Diebolt and Robert (1994), Escobar and West (1995), Richardson and Green (1997), Roeder and Wasserman (1997), Stephens (2000a) and Stephens (2000b).

From now on only mixture models developed in a survival context will be considered. For N individuals whose survival times can be grouped into J ($J > 1$) categories, the standard survival (mixture) density for an individual i can be written

$$f(t_i \mid \boldsymbol{x}_i, \boldsymbol{\Psi}) = \sum_{j=1}^{J} p_j f_j(t_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_j), \tag{6.1}$$

where $f_j(t_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_j)$ is the (proper) component survival density for the j^{th} group at time

t_i, with m-vectors of covariates $\boldsymbol{x}_i$ and parameters $\boldsymbol{\theta}_j$. The p_j are the mixing probabilities such that $0 < p_j < 1$ and $\sum_{j=1}^{J} p_j = 1$, and $\boldsymbol{\Psi}$ denotes the full vector of unknown parameters such that $\boldsymbol{\Psi} = (p_1, \ldots, p_J, \boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_J})$. The component densities, $f_j(\cdot)$, do not have to be identical since as long as they are proper then the constraints placed on the p_j's ensures that the overall survival density $f(\cdot)$ is also proper.

A common way in which (6.1) can be specified is by considering the use of a categorical (or multinomial) latent variable Z_i (for individual i) as a grouping indicator. The premises could be assigned *a priori* to a group by some measure pre-determined by the analyst - such as in the case of competing risks where the cause of failure can be observed. Furthermore, if the causes are assumed to be independent then (6.1) can instead be modelled by using a competing risks framework.

In practice explicit information on the cause of failure is often unknown and in the case of modelling FMD it is assumed that the process driving the mixing is that of resistance to infection rather than differences in the specific cause of failure. So instead the Z_i's can be treated as unobserved multinomial distributed random variables, with the probability of membership of group j being p_j. The conditional density function for $T_i \mid Z_i = z$ is given by $f(t_i; \mathbf{x_i}, \boldsymbol{\theta_z})$ where

$$Z_i \sim \text{Mult}_J(1; p_1, \ldots, p_J).$$

After marginalising out the Z's, the survival distribution for individual i is given by (6.1). Although this approach requires some assumptions about the nature of the mixing parameters p_j, it does allow for random variation in their definition. The simplest scenario is to leave the p_j constant, however perhaps a more useful (and intuitively more reasonable) approach would be to let them vary either over space or time, and/or with relation to a covariate of some kind.

Covariates can be included in the model in various ways, depending on how they are believed to influence the survival times. If believed to affect the probability of group membership then the covariates can be incorporated through the mixing parameters p_j. The simplest way to do this would be to use a logistic link function i.e. for individual i

with m-vector of covariates $\boldsymbol{x}_i$,

$$p_{ij} = \frac{\exp(\boldsymbol{\gamma}_j^T \boldsymbol{x}_i)}{1 + \exp(\boldsymbol{\gamma}_j^T \boldsymbol{x}_i)},$$

where $\boldsymbol{\gamma}$ is an m-vector of parameters.

If believed to directly influence the underlying hazard then a natural way to include covariates would be as a function of one or more parameters in some (or all) of the component hazard functions. There is nothing stopping their inclusion in both the hazard and mixing parameters if required, though care must be taken if the same covariates are used in each, due to the correlation structure that would arise as a result.

These arguments also apply to random effects, which can be included in either the hazard functions for each group, or in the mixing parameter, and can be either spatial or non-spatial (see section 4.5.6 for detail on possible frailty specifications). For example an intuitive method would be to use individual level spatial and non-spatial effects as additive terms in the component hazard functions, with $N(0, \tau)$ distributions for the non-spatial frailties and conditional autoregressive normal distributions for the spatial frailties (Besag et al. 1991). Alternatively the frailties could be sampled from a spatially-correlated gamma distribution and applied multiplicatively to the baseline hazard (Henderson et al. 2002, Shimakura 2003).

With regard to censored individuals, if the grouping mechanism is assumed to be independent of the failure and censoring mechanisms then each individual can be grouped regardless of their failure status. Right-censored observations can then be included by truncating the lower bound of the corresponding component survival distribution to the correct censoring time.

Given a mixture distribution of the form (6.1) with random censoring, then in the Bayesian framework the posterior distribution is:

$$P(\boldsymbol{\Psi} \mid \boldsymbol{t}) \propto \left(\prod_{i=1}^{N} \left[\sum_{j=1}^{J} p_j \{ f_j(t_i \mid \boldsymbol{\theta}_j) \}^{\delta_i} \{ S_j(t_i \mid \boldsymbol{\theta}_j) \}^{\delta_i - 1} \right] \right) P(\boldsymbol{\Psi}), \tag{6.2}$$

where δ_i is a binary variable taking the value 1 if individual i is infected and 0 if censored. It is usually desirable to have noninformative priors on the unknown parameters, $\boldsymbol{\Psi}$, so that the model isn't influenced by badly defined prior information. In the case of mixture models the use of improper priors leads to improper posteriors; which leads to further complications when deriving inferences from the resulting distributions.

A poor choice of prior distribution can also lead to the problem of label-switching (discussed subsequently). A more detailed discussion about the choice of prior distribution can be found in McLachlan and Peel (2000), chapter 4.

An additional issue is that of identifiability. The identifiability problem for a statistical model was discussed in section 4.4.3. Furthermore, if the distribution of interest $f(\cdot)$ is comprised of a mixture of distributions that contain components from the same parametric family, then the problem of label-switching can occur (also known as the *identifiability of mixtures*).

This happens if the prior distribution contains no information to help the model distinguish between the groups, since the posterior distribution will be invariant under different permutations of the component labels. For prediction purposes this is not usually an issue, but it is crucial for a Bayesian model if the posterior is to be used to make inferences about the mixture.

To illustrate this consider two parametric mixture distributions $f(\boldsymbol{t} \mid \boldsymbol{x}, \boldsymbol{\Psi})$ and $f(\boldsymbol{t} \mid \boldsymbol{x}, \boldsymbol{\Psi}^*)$ where

$$f(\boldsymbol{t} \mid \boldsymbol{x}, \boldsymbol{\Psi}) = \sum_{j=1}^{J} p_j f_j(\boldsymbol{t} \mid \boldsymbol{x}, \boldsymbol{\theta}_j),$$

and

$$f(\boldsymbol{t} \mid \boldsymbol{x}, \boldsymbol{\Psi}^*) = \sum_{j=1}^{J^*} p_j^* f_j(\boldsymbol{t} \mid \boldsymbol{x}, \boldsymbol{\theta}_j^*),$$

and all the component densities belong to the same parametric family. The mixtures are said to be identifiable when

$$f(\boldsymbol{t} \mid \boldsymbol{x}, \boldsymbol{\Psi}) \equiv f(\boldsymbol{t} \mid \boldsymbol{x}, \boldsymbol{\Psi}^*)$$

if and only if the number of components $J = J^*$ and the component labels can be permuted such that

$$p_j = p_j^* \qquad \text{and} \qquad f_j(\boldsymbol{t} \mid \boldsymbol{x}, \boldsymbol{\theta}) = f_j(\boldsymbol{t} \mid \boldsymbol{x}, \boldsymbol{\theta}^*).$$

This problem is usually handled by imposing an appropriate constraint on some or all of the parameters $\boldsymbol{\Psi}$ or their prior distributions. Caution must be used however, since recent work (e.g. Richardson and Green 1997, 1998 and Celeux et al. 2000) has shown that in certain circumstances (e.g. for two parameter mixture model with mixing probabilities ≈ 0.5) forced ordering can bias the parameter estimates and does not always prevent label-switching.

Celeux et al. (1996) suggest some methods for detecting label-switching in simulation studies, but they rely on knowing the 'true' values of the parameters. Richardson and Green (1997) suggest post-processing of the posterior simulations according to different label choices. Stephens (1997a) and Stephens (1997b) suggest relabelling of the MCMC output, and this approach is extended in Stephens (2000a). Another general relabelling algorithm is suggested in Celeux (1997) and Celeux (1998).

One possible way of circumventing this problem is to include covariate information in the mixing parameters p_j so that the mixture is driven by information in the data; however this still requires that the inclusion of the chosen covariates will lead to the mixing being well defined. An alternative way to view a mixture model is to force an ordering on the mixing probabilities, possibly through the use of an *ordinal* model.

An ordinal approach is useful when including covariates as part of the grouping probability as it forces a relative ordering to the data rather than an absolute probability. Consider as above that you have a regression term $\mu_{ij} = \boldsymbol{\gamma}_j^T \boldsymbol{x}_i$ for individual i in group j that is dependent on a set of covariates $\boldsymbol{x}_i$. Instead of using a logistic link to the mixing parameters p_j as before, a $(J-1)$-vector of random variables is introduced ($\boldsymbol{\nu}$), that represents the cut points between groups. For an individual i, $\nu_{ij} = p_{i1} + p_{i2} + \cdots + p_{ij}, (j = 1, \ldots, J)$, where p_{ij} is the probability of belonging to group j. Hence ν_{ij} is the cumulative probability of being in any ordered group up to j, and can be estimated by the model.

If the number of groups in the model is unknown, then one way of estimating how many components are in the model is to fit a series of models each with a different number of mixture components, and then use some goodness-of-fit measure that penalises for extra components. Richardson and Green (1997) suggest employing reversible jump MCMC techniques to estimate the number of components while Stephens (2000b) offers an alternative technique that views the parameters of the model as a (marked) point process and constructs a birth-death Markov process with an appropriate stationary distribution.

6.2.2 Long-term survivor models

It seems that Boag (1949) was the first to publish a paper discussing the use of a survival approach with a so-called 'cure' proportion. He wished to estimate the proportion of women cured of breast cancer in a population, and to do this he used a parametric model that was a mixture of two distributions, the first a log-normal distribution representing the survival times of those who develop breast cancer (*susceptibles*), and the other a degenerate distribution allowing for the essentially infinite survival times of those that had been cured (*immunes*). He also allowed the model to include a proportion of patients that were still alive at the end of follow-up but who suffered a recurrence of the disease rather than death. He used maximum likelihood to fit the model and treated deaths from causes other than breast cancer as a censoring mechanism.

A key extension to Boag (1949) is that of Berkson and Gage (1952), who used a long-term survivor framework to model the proportion of patients cured in a population suffering with stomach cancer. They noted that if cured individuals existed then the death rates of long-term survivors should drop to the baseline death rate of the population. To allow for a cured proportion they used a mixture of an exponential distribution and a degenerate distribution.

The standard form for a long-term survivor model as proposed by Berkson and Gage (1952) is given by

$$S(t) = p + (1 - p)S^*(t), \tag{6.3}$$

where $S^*(t)$ is the survivor function for the susceptible proportion, and p is the proportion of immune individuals in the population. In the context of FMD p is the proportion of premises considered resistant to infection. This framework has formed the basis for much of the subsequent literature in this field. It will be the convention in this thesis to refer to models containing an 'immune' or 'cured' proportion as *long-term survivor* rather than *cure rate* models, since in the context of FMD the principal concern is with modelling *resistance* to infection rather than *immunity* from infection.

Another important paper regarding the development of long-term survivor models was that of Farewell (1977), who extended the approach of Berkson and Gage (1952) to predict the proportion of women immune to breast cancer from a population using a series of recorded covariates. He wanted to investigate how these risk factors might influence not only the time to the development of the cancer but also the proportion who eventually developed the disease. He used a mixture of a Weibull and a degenerate distribution to model the susceptible and immune proportions respectively. The covariates were included in the model through a log-link function in the scale parameter of the Weibull distribution, and through a logistic-link in the probability of immunity. This is probably the most common model for long-term survivor data.

An extensive history of the development of long-term survivor models can be found in Maller and Zhou (1996). This provides a good introductory text on the subject and includes many examples of the use of these models in a wide range of different fields of study. This thesis will give a brief review of some of the more recent developments, notably the extension of these models to a Bayesian framework and the incorporation of spatial information. In particular the approaches of Chen et al. (1999) and Banerjee and Carlin (2004) will be discussed.

Banerjee and Carlin (2004) develop a Bayesian long-term survivor model that incorporates interval-censoring and a spatial dependence structure. They use this to model smoking cessation data (aggregated into areas), in which the time-to-relapse is recorded for patients who resume smoking after an initial attempt at quitting. They consider using a single

latent binary variable Z to represent the 'propensity to relapse', but allow the probability of relapse to vary across both individuals and regions (i.e. Z_{ji} is the latent variable for the i^{th} individual in the j^{th} region).

The authors assume an implicit proportional hazards structure in the susceptible group through the use of both gamma and Weibull distributions for the time-to-relapse T_{ji}. Conditional upon the Z_{ji}'s the T_{ji}'s are independent with survivor and density functions $S^*(t_{ji}; \boldsymbol{\Psi}_{ji})$ and $f^*(t_{ji}; \boldsymbol{\Psi}_{ji})$ respectively. If the Z_{ji}'s are assumed Bernoulli distributed with parameter $1 - p_{ji}$, then after marginalising over the Z_{ji}'s the survival distribution for the i^{th} individual in the j^{th} region is given by

$$S(t_{ji}; \boldsymbol{\Psi}_{ji}) = p_{ji} + (1 - p_{ji})S^*(t_{ji}; \boldsymbol{\theta}_{ji}), \tag{6.4}$$

where $\boldsymbol{\Psi}_{ji}$ is the complete vector of parameters, $1 - p_{ji}$ is the probability of relapse and $\boldsymbol{\theta}_{ji}$ is the vector of parameters relating to the survival distribution for the individuals at risk. This allows the cure proportion and time-to-relapse for the susceptible group to vary over different spatial regions.

Spatial frailties are also included through a link function in the scale parameter of the survival distribution for the susceptible group. These are jointly modelled with spatial effects in both the baseline hazard and cure proportions through the use of a multivariate conditional autoregressive (MCAR) prior (see Gelfand and Vounatsou 2003, Carlin and Banerjee 2003).

Chen et al. (1999) devise a different modelling strategy using a series of latent variables to represent the underlying biological process rather than the single latent variable assumed by Banerjee and Carlin (2004). They suggest a number of advantages of their approach over that of the standard long-term survivor model given by (6.3). They note that in the presence of covariates the standard long-term survivor model does not have a proportional hazards structure, and if covariates are included in the cure parameter through a standard binomial regression model then for many types of improper prior distributions improper posteriors are obtained. Also that the assumption of a single latent variable representing

the underlying survival process does not have a sound biological interpretation for certain types of processes such as modelling time to relapse for a cancer patient.

They develop a Bayesian model for modelling long-term survivors that has a proportional hazards structure, and in contrast to Banerjee and Carlin (2004) allows the covariates to influence the probability of an individual being immune (or in their case cured) rather than the underlying process governing the time to development of the disease. They include discussions on prior elicitation and how their model relates to the standard cure rate model (6.3). They fit it to a data from a melanoma clinical trial using maximum likelihood.

Although this approach is not directly relevant to the problem at hand, the underlying biological arguments surrounding its conception are interesting, and could perhaps be used in other epidemic situations. It can also be written as a standard cure rate model and a short discussion of the mathematical detail follows.

The biological principal governing their model formulation is that after treatment an unknown number of carcinogenic cells remain, of which it takes only one of those cells to develop cancer in order for the patient to experience relapse. So if N represents the number of carcinogenic cells for an individual left active after initial treatment, N is assumed to follow a Poisson distribution with mean θ. Define Z_k as a series of independent and identically distributed latent random variables ($k = 1, \ldots, N$) denoting the time taken for the k^{th} carcinogenic cell to produce a detectable cancer mass. The time to relapse of the cancer T can be defined as,

$$T = \min(Z_k, 0 \leqslant k \leqslant N).$$

Here $P(Z_0 = \infty) = 1$ and N is independent of the sequence $Z_1, Z_2, \ldots$. The overall survivor function for an individual is given by

$$\begin{aligned} S(t) &= P(\text{no cancer by time } t) \\ &= P(N = 0) + P(Z_1 > t, \ldots, Z_N > t, N \geqslant 1) \\ &= \exp(-\theta) + \sum_{k=1}^{\infty} S_Z(t)^k \frac{\theta^k}{k!} \exp(-\theta) \\ &= \exp(-\theta + \theta S_Z(t)) = \exp(-\theta F_Z(t)), \end{aligned} \tag{6.5}$$

where $S_Z(t)$ and $F_Z(T)$ are the survivor and distribution functions for the i.i.d. latent variables Z_k. The authors note that the (6.5) would still be valid for any data set with long-term survivors that can be thought of as being generated by an unknown number of latent competing risks.

The survivor function for the susceptible cells is given by

$$S^*(t) = P(T > t \mid N \geqslant 1) = \frac{\exp(-\theta F(t)) - \exp(-\theta)}{1 - \exp(-\theta)}, \tag{6.6}$$

and from this the authors show that the model (6.5) can be written as a standard long-term survivor model (6.3) in the form

$$S(t) = \exp(-\theta) + (1 - \exp(-\theta))S^*(t). \tag{6.7}$$

In this case the cure proportion $p = \exp(-\theta)$ (see Chen et al. 1999, for details).

It can be seen that neither (6.5) or its corresponding density function, given by $f(t) = \theta f_Z(t) \exp(-\theta F_Z(T))$, are proper since $p = S(\infty) = \exp(-\theta)$, however the overall hazard function

$$h(t) = \theta f_Z(t), \tag{6.8}$$

has a proportional hazards structure with covariates modelled through the cure parameter θ. Proper posteriors also arise for regression coefficients $\boldsymbol{\gamma}$ even under improper priors (though this does not hold if N is considered Bernoulli). So the two different approaches have various advantages and disadvantages dependent on the underlying beliefs about the process driving survival.

6.3 Simple simulation study

In this section a simple simulation study is used to investigate the effect of resistance to infection on the parameter estimates obtained from the conventional, long-term survivor

and mixture models. In the actual FMD epidemic under study there was evidence in the literature (e.g. Keeling et al. 2001a) that different premises exhibited different levels of susceptibility, and this was also alluded to by the results in the previous chapter and the spatial distribution of premises in Devon i.e. various areas experiencing high concentrations of the virus but containing large numbers of uninfected premises (figure 5.1). For a highly contagious disease such as FMD it would seem counterintuitive for this to be the case if all premises were actually equally susceptible.

In order to replicate this behaviour in the simulated epidemic, individuals were assigned to either a resistant or a susceptible group dependent on a simulated covariate y (representing e.g. biosecurity - sampled from a mixture of $N(20, 2^2)$ and $N(40, 2^2)$ distributions). Group membership was considered fixed - that is that resistance to infection is time-independent and does not change during the epidemic. A positive x covariate (representing e.g. size of herd) was sampled for each individual from a $G(2/5, 25/2)$ distribution. The survival process for the susceptible group was formulated in such a way that increasing values of x result in shorter survival times. The survival process for the resistant group was deemed independent of x and was simply governed by an underlying baseline hazard function common to both groups.

A small data set containing 500 individuals was generated, with a 4:1 ratio of susceptible to resistant individuals. A plot of the infectious covariate x against failure times is given in figure 6.1, with susceptible individuals shown in black and resistant individuals in red. The simulation was censored at 72 days, giving a total of 200 IPs and 300 UIPs.

A series of models were then fitted to the data with varying degrees of censoring and resistance. In the first instance the resistant premises were removed from the data set and the conventional, long-term survivor and mixture models fitted to IPs plus increasing numbers of censored observations. Subsequently resistant premises were included, firstly in the IPs only and finally in both the IPs and censored individuals. The model frameworks were based on the discrete-time Weibull models discussed in sections 4.3.3 and 5.2, i.e. for an individual with failure/censoring time t_i and covariates x_i and y_i, the survivor function

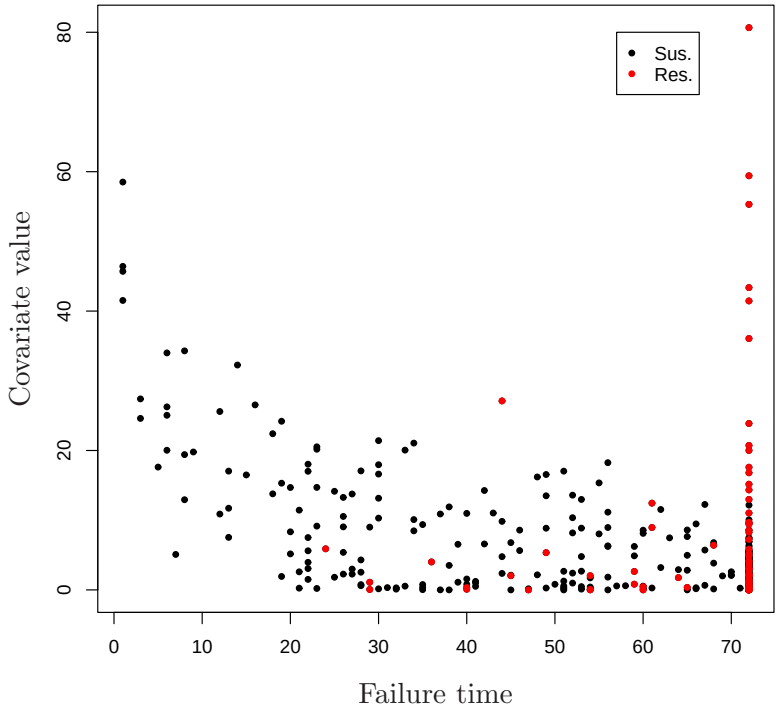


Figure 6.1: Plot of failure times against infectious covariate for non-spatial simulation

for the conventional model is:

$$S(t_i) = \exp(-\lambda_i t_i^{\alpha}), \tag{6.9}$$

where $\lambda_i = \exp(\beta_0 + \beta_1 x_i)$. Similarly, the survivor functions for the long-term survivor and mixture models are given by:

$$S(t_i) = p_i + (1 - p_i)S_1(t_i) \tag{6.10}$$

and

$$S(t_i) = p_i S_0(t_i) + (1 - p_i)S_1(t_i) \tag{6.11}$$

respectively. Here $S_1(t)$ has the form (6.9) with $\lambda_i = \exp(\beta_0 + \beta_1 x_i)$ as before, and $S_0(t)$ is given by (6.9) with $\lambda_i = \lambda = \exp(-\eta_0)$. (Note that in the simulation $\beta_0 = \eta_0$.) The probability of being *susceptible*, p_i, was given a logistic link to the covariate y, i.e.

$$\log\left(\frac{p_i}{1 - p_i}\right) = \gamma_0 + \gamma_1 y_i.$$

The models were fitted in WinBUGS using uninformative priors on the parameters. The intercept parameters β_0, η_0 and γ_0 were given $N(0, 100)$ distributions, a $N(0, 10)$ distribution was given to β_1 and the shape parameter α was given a $G(0.1, 10)$ distribution. In order to prevent label-switching, γ_1 was constrained to be positive by assigning it a $G(0.1, 10)$ prior distribution. As discussed in the previous chapter the models were sensitive to the choice of initial value used, and so initial values were generated in a similar way by using the range of the data to fix sensible limits over which to randomly sample the starting points of the chain. Details of how to derive the WinBUGS code for the models is given in appendix A and details about generating initial values in appendix B.

Two chains were used with a burn-in of 10000 iterations and a further 40000 updates. Again the posterior samples were thinned so that 1000 values were returned. The convergence diagnostics were reasonable and posterior summaries of the parameter estimates are shown in table 6.1. For reasons of space we report the results for the cases of low censoring (IPs and 50 UIPs) and high censoring (IPs and all UIPs), though the results from intermediate levels of censoring and repeated simulations reinforce the general patterns discussed here.

Essentially the effect of resistance in the data set is to produce influential outliers that cause heavy bias in the posteriors for the parameters of the conventional model. Looking initially at the regression parameters for the survival processes rather than the mixing, it can be seen that when resistance is absent from the data set the parameter estimates are reasonably well estimated even under increased levels of censoring. In addition the long-term survivor and mixture models seem to satisfactorily reproduce the results from the conventional fit. There is some discrepancy between the posterior interval and true value for the β_1 parameter at low levels of censoring for each of the three fitted models, highlighting the potential biases caused by excluding censored information from the model. At higher levels of censoring the posterior estimates are much better.

As soon as resistance is introduced the conventional model estimates begin to break down. Even for low levels of censoring (where the only resistant premises are those that are also

infected) the posterior for the β_1 parameter gets markedly worse. The long-term survivor model also exhibits this behaviour, though to a lesser degree, whereas the mixture model stands up better in the case of both high- and low-censoring.

In order to fully understand this we need to examine how well the latter models are capturing the resistance. Table 6.2 gives a summary of the numbers of individuals mis-specified with regard to estimated resistance status in the model. The overall resistance status for an individual was calculated from the mean posterior probability of susceptibility such that if it was greater than 50% the individual was classified as susceptible and less than 50% as resistant. It can be seen that when resistance is absent both the long-term survivor and mixture models correctly identify all premises as susceptible, whereas when resistance is introduced the long-term survivor model begins to overpredict the numbers of susceptibles. This is unsurprising since the model formulation in the latter case treats *all* infected individuals as susceptible, leading to some confounding in the posterior estimates for the mixing parameters γ_0 and γ_1. The mixture model on the other hand allows for variability in resistance between IPs, and as such performs better when estimating the susceptibility status.

6.4 Predicted survival times

The simulation study in the previous section allows us to compare actual and predicted survival times. Predictive posterior distributions of failure times for censored individuals were obtained from the models fitted in 6.3, and figure 6.2 shows plots of the actual failure times post censoring against the mean predicted failure time for the case where resistance is present in the data set. The solid line corresponds to the correct specification with the upper and lower dashed lines representing twice the actual value and half the actual value respectively. Parkes (1972) defined predictions lying outside this range as being in 'serious error'. It can be seen that all the models have a tendency to overpredict the actual survival times (e.g. be too optimistic). Also there is a lot of variation in the accuracy of these individual level point predictions, and this particular issue has been noted in the literature

		No resistance							Resistance							
		Par.	**Mean**	**s.e.**	**2.5%**	**50%**	**97.5%**	$\hat{R}$	n_{eff}	**Par.**	**Mean**	**s.e.**	**2.5%**	**50%**	**97.5%**	$\hat{R}$	n_{eff}
Low censoring - estimates (actual values)	**Conv**	α (2)	1.9721	0.1219	1.7340	1.9715	2.2173	1.0367	46	α (2)	1.8878	0.1147	1.6639	1.8845	2.1220	1.0017	1000
	LTS		1.9753	0.1230	1.7330	1.9780	2.2181	1.0037	400		1.9871	0.1126	1.7830	1.9850	2.2061	1.0042	450
	Mix		1.9924	0.1396	1.7280	1.9865	2.2901	1.0180	86		2.1688	0.1331	1.9199	2.1630	2.4292	1.0021	720
	Conv	β_0 (−9.8982)	-8.8188	0.5410	-9.9233	-8.8115	-7.7869	1.0406	42	β_0 (−9.8982)	-8.2039	0.4925	-9.1981	-8.2005	-7.2525	1.0015	1000
	LTS		-8.8294	0.5431	-9.8751	-8.8450	-7.7746	1.0036	420		-8.6847	0.4823	-9.6481	-8.6695	-7.8238	1.0034	480
	Mix		-8.9351	0.6513	-10.3703	-8.8955	-7.7689	1.0176	87		-9.6315	0.5980	-10.7600	-9.6325	-8.4909	1.0015	1000
	Conv	β_1 (0.1833)	0.1327	0.0116	0.1108	0.1326	0.1556	1.0254	75	β_1 (0.1833)	0.0946	0.0100	0.0752	0.0945	0.1132	1.0022	940
	LTS		0.1330	0.0116	0.1104	0.1328	0.1551	1.0019	800		0.1114	0.0097	0.0929	0.1116	0.1299	1.0014	1000
	Mix		0.1369	0.0156	0.1110	0.1352	0.1730	1.0082	190		0.1490	0.0153	0.1213	0.1487	0.1801	1.0004	1000
	LTS	γ_0 (NA)	-15.4995	5.5621	-28.3085	-14.6200	-7.0389	1.0047	320	γ_0 (−6.91)	-12.9570	4.9591	-25.3068	-11.9100	-6.1467	1.0000	1000
	Mix		-14.1417	5.9871	-27.0823	-13.3350	-4.2838	1.0003	1000		-13.3667	5.5506	-25.8320	-12.8300	-4.4387	1.0022	680
	LTS	γ_1 (NA)	0.2043	0.1762	0.0067	0.1533	0.6637	1.0051	470	γ_1 (0.23)	0.2176	0.1190	0.0352	0.1987	0.5040	1.0002	1000
	Mix		0.2585	0.2378	0.0069	0.1844	0.9018	1.0031	820		0.5248	0.2346	0.1631	0.4998	1.0701	1.0021	730
	Mix	η_0(NA)	-2.1609	9.6979	-19.0523	-4.1445	19.1730	1.0023	650	η_0(-9.8982)	-8.9917	0.5728	-10.1305	-8.9475	-7.9190	1.0034	440

		No resistance							Resistance							
		Par.	**Mean**	**s.e.**	**2.5%**	**50%**	**97.5%**	$\hat{R}$	n_{eff}	**Par.**	**Mean**	**s.e.**	**2.5%**	**50%**	**97.5%**	$\hat{R}$	n_{eff}
High censoring - estimates (actual values)	**Conv**	α (2)	1.9653	0.1324	1.7160	1.9635	2.2390	1.0127	120	α (2)	1.3981	0.0968	1.2190	1.3955	1.5960	1.0040	380
	LTS		1.9999	0.1325	1.7520	2.0010	2.2620	1.0696	27		1.8905	0.1196	1.6680	1.8910	2.1320	1.0066	230
	Mix		2.0497	0.1476	1.7778	2.0400	2.3650	1.0088	250		2.0090	0.1245	1.7699	2.0035	2.2480	1.0004	1000
	Conv	β_0 (−9.8982)	-9.7457	0.5928	-10.9303	-9.7400	-8.6207	1.0147	100	β_0 (−9.8982)	-6.8331	0.4083	-7.6693	-6.8105	-6.1020	1.0033	460
	LTS		-9.8996	0.5930	-11.0900	-9.8915	-8.8059	1.0690	28		-9.3552	0.5247	-10.3910	-9.3555	-8.3808	1.0049	310
	Mix		-10.1819	0.7004	-11.7805	-10.1500	-8.8851	1.0078	200		-9.9394	0.5596	-10.9903	-9.9280	-8.9109	1.0002	1000
	Conv	β_1 (0.1833)	0.1843	0.0111	0.1626	0.1840	0.2050	1.0149	110	β_1 (0.1833)	0.0322	0.0041	0.0240	0.0322	0.0398	1.0038	950
	LTS		0.1864	0.0110	0.1645	0.1867	0.2076	1.0343	55		0.1644	0.0095	0.1462	0.1646	0.1822	1.0001	1000
	Mix		0.1931	0.0143	0.1668	0.1919	0.2252	1.0156	99		0.1875	0.0111	0.1664	0.1870	0.2100	1.0001	1000
	LTS	γ_0 (NA)	-15.3780	5.5191	-28.3335	-14.5300	-7.2272	1.0022	1000	γ_0 (−6.91)	-9.5555	1.4992	-13.0303	-9.3925	-7.0219	1.0017	880
	Mix		-16.2349	5.8441	-30.1138	-15.4900	-6.8224	1.0023	1000		-19.3944	5.4810	-31.1513	-18.9650	-10.3688	1.0013	1000
	LTS	γ_1 (NA)	0.2016	0.1754	0.0052	0.1557	0.6808	1.0098	1000	γ_1 (0.23)	0.2144	0.0382	0.1463	0.2111	0.2980	1.0025	600
	Mix		0.3873	0.3163	0.0080	0.3176	1.0985	1.0492	63		0.6582	0.1993	0.3364	0.6410	1.1151	1.0038	1000
	Mix	η_0(NA)	-3.3734	8.8417	-17.2440	-7.2695	16.6873	1.0021	710	η_0(-9.8982)	-9.9708	0.5573	-11.0908	-9.9515	-8.9342	1.0029	510

Table 6.1: Parameter estimates from models fitted to non-spatial simulation

			No resistance				Resistance		
			Misspec. as Res	Misspec. as Sus.	Total		Misspec. as Res	Misspec. as Sus.	Total
Censoring	Low	LTS	0	0	0	LTS	0	31	31
		Mix	0	0	0	Mix	2	0	2
	High	LTS	0	0	0	LTS	0	100	100
		Mix	0	0	0	Mix	0	0	0

Table 6.2: Estimated numbers of individuals misspecified with regards resistance in non-spatial simulation

in the past, most recently in Henderson and Keiding (2005). The authors conclude that even in the case where the statistical model is known to be true and with no uncertainty in the parameter estimates, individual level predictions of survival times are of limited practical use as a prognostic tool.

As an illustrative example consider a plot of the posterior distribution for an individual with an arbitrary covariate value ($x = 0.836$) obtained from the conventional model fitted to the data with no resistance and high censoring (figure 6.3). Comparing this empirical predictive distribution against the theoretical distribution obtained in the case where the true parameter values are known shows a good association. However the actual failure time is one realisation from this distribution, and so even if the parameters are accurately estimated the range of the predictive distribution can still be large. In the case of many real-life survival processes large amounts of uncertainty remain even when using predictive intervals rather than a mean or median point prediction (see Henderson and Keiding 2005). In practical terms there are often further unknown heterogeneities that exacerbate problems of prediction. This is especially true in the case of epidemic data where each realisation of the underlying survival process directly affects the path of the epidemic. In the non-spatial simulation the covariate affecting failure time is not linked to the history of the epidemic process, however in real infectious disease situations just one extraneous infection could have a large effect on the dynamics of the disease. This further invalidates the use of long-term predictive windows and reinforces the potential importance of trying

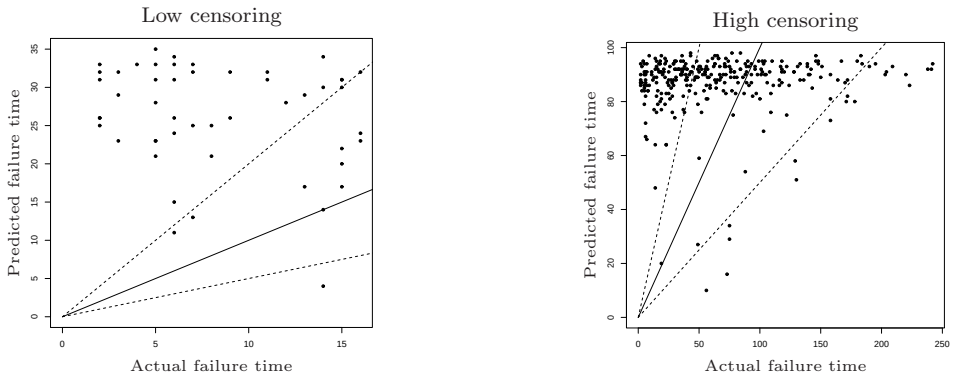
to use sequential model fitting techniques to minimise the effect of these rogue infections.

The important point to note from these comparisons is that the long-term survivor and mixture models perform better than the conventional approach, both in terms of getting more point predicted survival times within the error bands and also in correctly identifying resistant and susceptible premises.

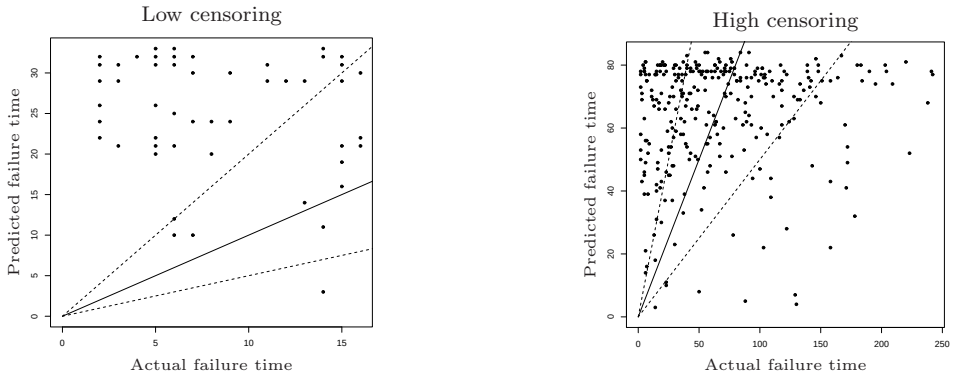
6.5 Conclusions

The primary focus of this chapter was to investigate modelling of resistance to infection in the survival framework. Two extensions to the conventional survival model - those of the long-term survivor and mixture models - were explored as a means to alleviate the bias caused by the resistant process in the simulation study. It was seen that both methods had advantages over the conventional approach, though the extent to which the parameter estimates improved was greatly dependent on how well the models captured the mixing process. The mixture model seemed to do better since it allows for variation in susceptibility in the infected premises, which the long-term survivor model does not. Another potential advantage of the mixture approach over other methods (e.g. Keeling et al. 2001a) of including resistance to infection is that it allows the *shape* of the underlying probability distribution to change between the susceptible and resistant groups instead of just the magnitude.

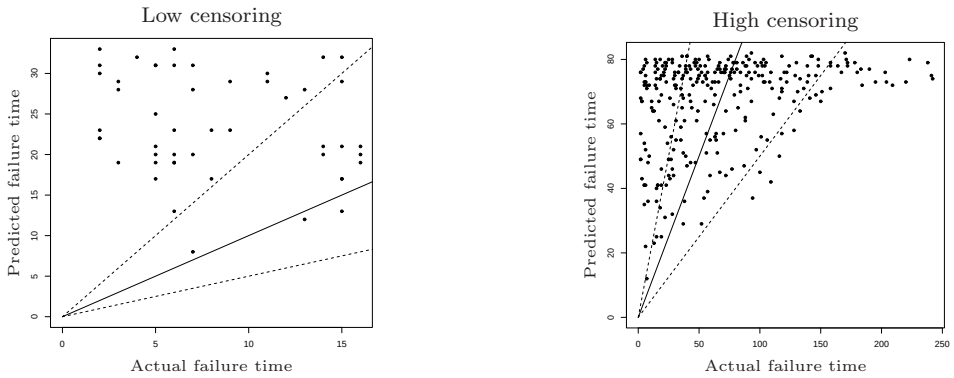
A secondary issue that emerged from this chapter concerns the accuracy of individual level survival predictions. In practice there are many unknown or unobserved facets relating to survival that are not accounted for but may have a large influence on the predicted failure times. However there is evidence in the literature that survival predictions are much more useful when viewed at the population level, and some potentially useful population measures and their uses in predicting the path of a spatial epidemic will be shown in section 7.2. The earlier part of chapter 7 will investigate the effects of resistance to infection



(a) Conventional model



(b) Long-term survivor model



(c) Mixture model

Figure 6.2: Plots of actual vs. predicted failure times for non-spatial simulation

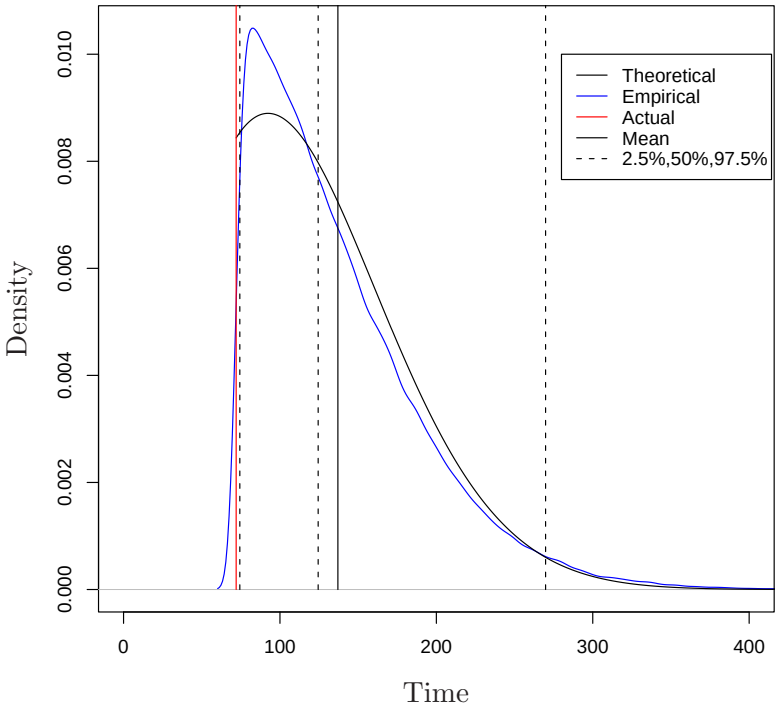


Figure 6.3: Predictive posterior distribution comparison for arbitrary individual obtained from conventional model fitted to data set with no resistance and high censoring for non-spatial simulation

and mixture and long-term survivor models in a more sophisticated spatial simulation setting.

Chapter 7

Applications of mixture modelling to infectious animal disease epidemics

The previous chapter discussed potential ways in which issues regarding non-localised infections and resistance to infection could be dealt with in a survival framework. The results from a simple non-spatial simulation study suggested that including large numbers of censored observations in the models results in far less bias to the predicted survival times than occurs when including even small numbers of resistant premises.

In this chapter a more complex spatial simulation study will be carried out in order to compare the long-term survivor and mixture approaches to the conventional model in a situation where the simulated epidemic encompasses spatial as well as temporal structure. Here the simulations more accurately reflect the dynamics of a real animal disease outbreak and will be based on a spatially random subset of locations of farm holding premises in Devon, with artificial covariates used to control various aspects of the simulated spread.

There are three main sections to the chapter. The first, section 7.1, gives details about the simulation study, including the algorithm used to generate the data, some discussion about

the simulated data sets, the mathematical form of the models used and some adjustments to the predictive algorithm (discussed in section 5.3) to incorporate resistance to infection. Comparative results are also given.

Section 7.2 focusses on ways in which these techniques can be usefully applied to spatial epidemic data - in order to produce spatial and temporal predictive hazard maps and the use of this information as a means of targeting control policies to high-risk areas.

Finally, in section 7.3 these ideas are applied to the data set from the 2001 FMD epidemic in Devon (see chapter 5). Some overall conclusions are given in section 7.4.

7.1 Spatial simulation study

7.1.1 Details of the simulation

Two simulated covariates were used to represent different characteristics of each premise. The first, x, relates to the potential infectiousness of the premise were it to become infected, and the second, y, to susceptibility. The x covariate was sampled from a normal distribution and the y covariate from a mixture of two normal distributions with different means but the same variance. The mixing parameters were set by the corresponding level of resistance required in the data set. In the case of FMD these two covariates most likely represent size and biosecurity respectively, though for the simulation they are assumed to be generic measures that may relate to corresponding factors important in the pathogenesis of the disease under study. The dynamics of the epidemics were controlled by the varying the values of the parameters in the simulations.

One source premise was assumed and at any time point a simplified version of the VL (pseudo-viral load, PVL, based on x) was calculated, with exposure determined by a threshold value decided beforehand. Two survival processes were used - one (localised) based on relative time from exposure and driven by PVL, and the other ('spark') a baseline hazard, independent of covariates and based on absolute time from the beginning of the

epidemic. Only premises that were classified as *susceptible* were assumed at risk from the localised infective process whereas *all* premises were assumed at risk from the non-localised process. This adds some measure of additional heterogeneity into the susceptible group - reflecting our prior belief in the way that the exposure mechanism in a real-life animal disease epidemic may affect the recorded data set.

The PVL measure is given by:

$$\mathrm{PVL}(\boldsymbol{s}, t) = \sum_{j|t_j<t} \mathcal{I}_{\mathrm{P}}(t - t_j, x_j)\omega(\boldsymbol{s}, \boldsymbol{s}_j, \boldsymbol{\Sigma}), \tag{7.1}$$

where $\mathcal{I}_{\mathrm{P}}$ is the pseudo-infectivity function given by a gamma curve

$$\mathcal{I}_{\mathrm{P}}(t, x) = \frac{\eta x}{\xi^{\gamma}\Gamma(\gamma)} t^{\gamma-1} e^{\frac{-t}{\xi}}, \tag{7.2}$$

with shape and scale parameters γ and ξ respectively. The constant η helps to scale $\mathcal{I}_{\mathrm{P}}$ to ensure reasonable values. The form of (7.2) was chosen because it mirrors the type of behaviour that would be expected for the within-herd spread of FMD for a generic animal species over time. A bivariate normal kernel function was used for the distance decay term $\omega(\cdot)$ with covariance matrix $\boldsymbol{\Sigma}$ controlling the bandwidth and hence the degree of spatial smoothing in the simulation.

The PVL allows the analyst to control the shape and magnitude of viral coverage over time to create simulated epidemics with varying dynamics. (Note that in the simulated epidemics as for the real epidemic, the average cumulative PVL - PAV - will be used to drive the hazard functions with the PVL used to determine exposure.)

The probability of resistance to infection was controlled by a logistic-link function to the covariate y as,

$$p_i = \frac{\exp(\gamma_0 + \gamma_1 y_i)}{1 + \exp(\gamma_0 + \gamma_1 y_i)}, \tag{7.3}$$

again allowing the analyst to control the degree of resistance present in the data set. Here resistance to infection was considered time-independent i.e. premises were classified as

either resistant or susceptible at the outset and their status did not change over time.

In order to conduct the simulation, let $\boldsymbol{D}$ be the complete data set for N premises, where each premise has covariates x and y, a variable t representing infection/censoring time and a binary indicator δ, where $\delta = 1$ for an infected premise and $\delta = 0$ for an uninfected (censored) premise. In addition let t^e be a variable representing exposure time and δ^e be another binary indicator, with $\delta^e = 1$ if a premise is *exposed* and $\delta^e = 0$ if *not exposed*.

At the beginning of the simulation set δ, δ^e and PVL to 0, and t and t^e to NA for all premises. Then determine a threshold value for exposure to the virus. In addition set a maximum time E for the epidemic to run. The set of epidemic parameters is given by $\boldsymbol{\Psi}$ and these need to be set before the simulation. The number of 'source' premises (S_0) can then be randomly sampled from the data. Label these premises $1, \ldots, S_0$ and set $t_1, \ldots, t_{S_0} = 0$ and $\delta_1, \ldots, \delta_{S_0} = 1$.

The simulation algorithm is given by:

1. Set $t = 1$.
2. Let $\text{IP}^{(t)}$ be the set of infected premises such that $\delta_i = 1$ and $\text{UIP}^{(t)}$ be the set of uninfected premises such that $\delta_i = 0$ at time t.
3. Calculate $\text{PVL}_j(t)$ for each of the C members of $\text{UIP}^{(t)}$ $(j = 1, \ldots, C)$.
4. If PVL for premise j exceeds the threshold value for exposure and $\delta_j^e = 0$, then set $\delta_j^e = 1$ and $t_j^e = t$.
5. Calculate
$$d_j = \begin{cases} t - t_j^e - 1 & \text{if } t > t_j^e \text{ and } \delta_j^e = 1, \\ 0 & \text{otherwise.} \end{cases}$$

6. Then calculate

$$h_j(t \mid \mathbf{\Psi}) = \begin{cases} \min\{P_L(d_j \leqslant T < d_j + 1 \mid T \geqslant d_j, \mathrm{PAV}_j(t-1)), \\ \qquad P_S(t-1 \leqslant T < t \mid T \geqslant t-1)\} & \text{if } \delta_j^e = 1, \\ P_S(t-1 \leqslant T < t \mid T \geqslant t-1) & \text{otherwise.} \end{cases}$$

   where $P_L(\cdot)$ represents the *localised* survival process and $P_S(\cdot)$ the *spark* process.

7. Let u_j be a random sample from a $U(0,1)$ distribution corresponding to premise j and set $t_j = t$ and $\delta_j = 1$ if $u_j < h_j(t \mid \mathbf{\Psi})$.

8. Set $t = t+1$. If $t > E$ or there are no remaining uninfected premises then go to step 9. Else go to step 2.

9. END.

This returns a data set that records the absolute survival time from the beginning of the epidemic regardless of whether premises became infected through spark or localised infections. In this way the analyst can set a threshold value for the simulation, but also investigate the effects of fitting models to data that has been sorted via exposure using alternative threshold values.

7.1.2 The simulated epidemics

The parameters $\mathbf{\Psi}$ relating to $\mathcal{I}_{\mathrm{P}}$ were defined by taking approximations of the parameter estimates for the best-fitting gamma curve for a median size cattle herd in Devon (given in table 5.3). The parameters relating to the localised infection process were set by taking a infected premise with a mean x covariate and fixing a probability of infection after a pre-determined number of days for an uninfected premise located a set distance away. The distances and times could be changed to reflect the degree of infectiousness required. A similar technique was used to fix the baseline hazard for the spark process - though in this

	N	IPs	UIPs	No. spark infections	No. localised infections	No. resistant
No resistance	1000	379	621	47	332	0
Low resistance	1000	189	811	38	151	329
Medium resistance	1000	191	809	42	149	481
High resistance	1000	179	821	51	128	661

Table 7.1: Summary values for simulated spatial epidemics at 50 days

case no covariate dependence was assumed. Resistance to infection was controlled by (7.3) though the $\boldsymbol{\gamma}$ parameters remained the same and the y covariates were changed to reflect the degree of resistance required. In addition no culling was used in these simulations in the first instance.

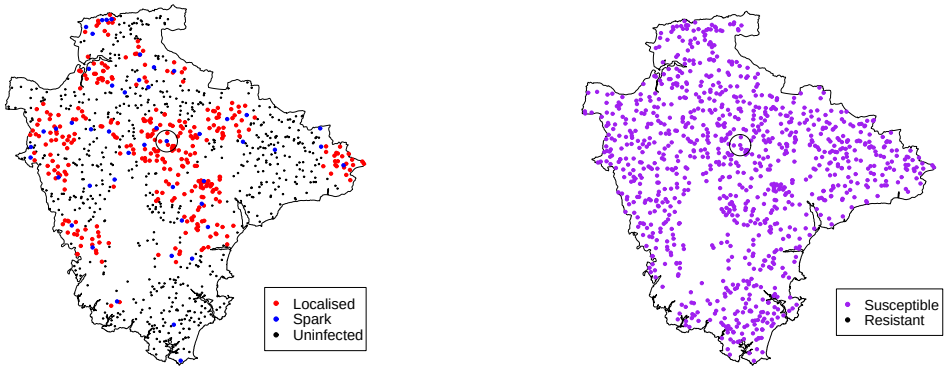
In order to compare the three survival approaches four initial epidemics were simulated with varying degrees of resistance to infection in the data set - none, low ($\approx 25\%$), medium ($\approx 50\%$) and high ($\approx 75\%$). A discrete Weibull distribution was used to drive survival for both the resistant and susceptible groups.

Figure 7.1 and table 7.1 give comparative summaries for a set of four simulated epidemics over 50 days, all based around the same subset of 1000 premises from Devon with the same initial source premise (circled). Figure 7.1 helps to show the differences in spatial patterns that can arise from increased resistance levels, and it can be seen that as the level of resistance in the data set increases then number of localised infections decreases.

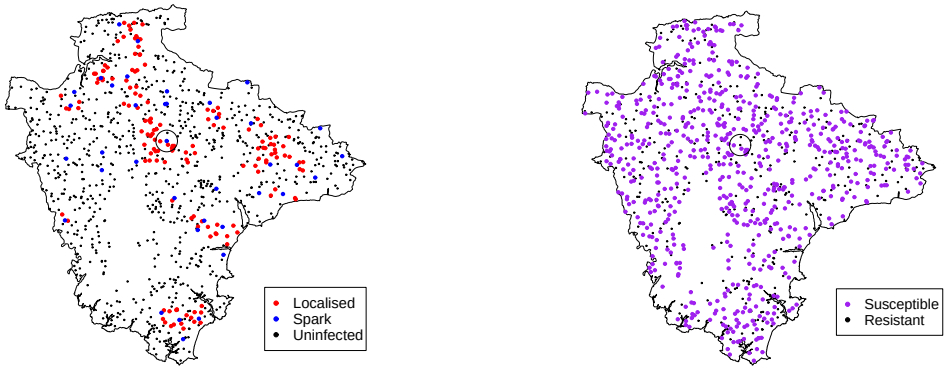
It can be seen from figure 7.2 that each of the four simulated epidemics still seem to be growing in size after 50 days - possibly as a result of the lack of control policies used in the simulation. This will be investigated in more detail in section 7.2.

7.1.3 Model formulations and prediction

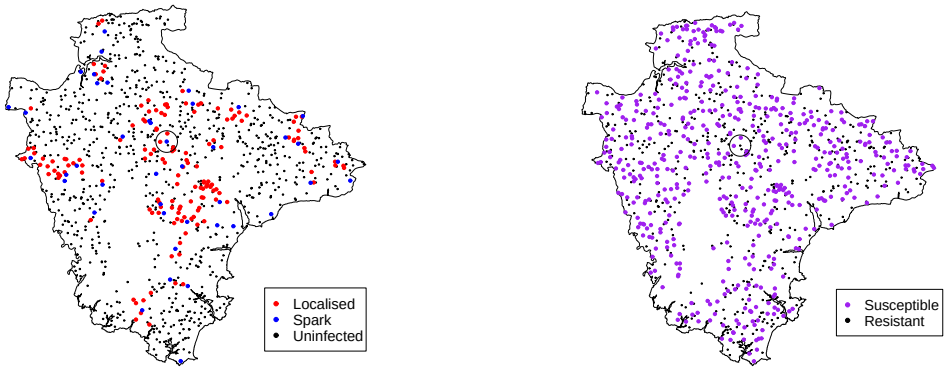
In this section the predictive power of the conventional approach will be tested against that of the long-term survivor and mixture model approaches. The form of the hazard



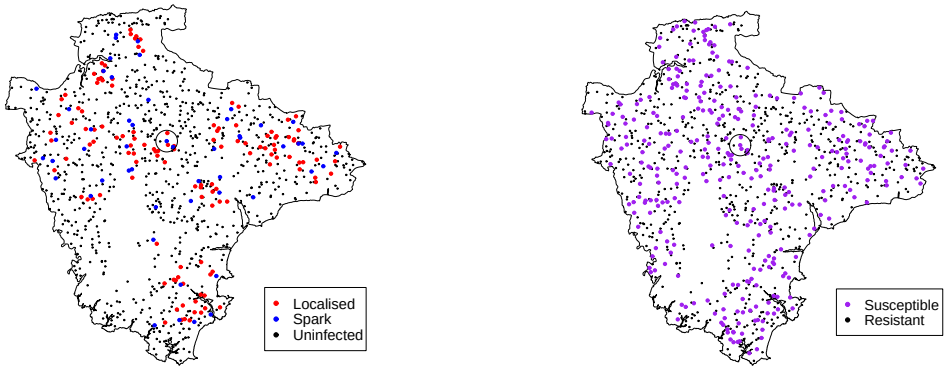
(a) No resistance



(b) Low resistance



(c) Medium resistance



(d) High resistance

Figure 7.1: Spatial maps of simulated epidemics at 50 days

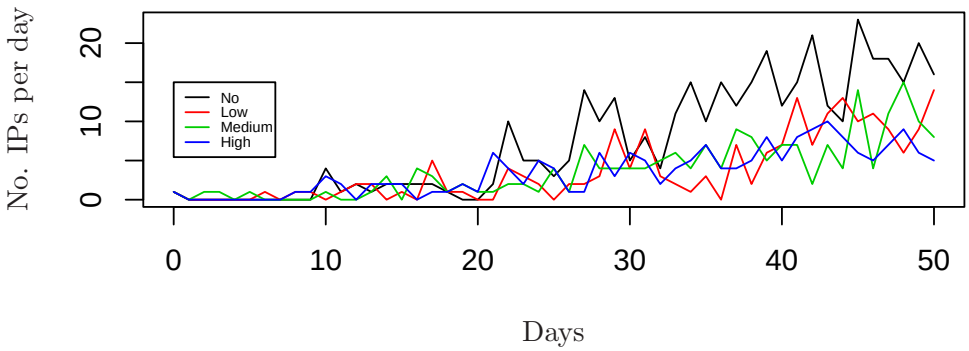


Figure 7.2: Epidemic plots for spatial simulations over time

function for the conventional model at time t for premise i is given by

$$h_i(t) = 1 - \exp(-\lambda_{i(t-1)}[t^\alpha - (t-1)^\alpha]), \tag{7.4}$$

where covariates are included in the scale parameter as $\lambda_{i(t-1)} = \exp(\beta_0 + \beta_1 \text{PAV}_{i(t-1)})$. This has corresponding survivor and probability functions,

$$S_i(t) = \exp\left(-\sum_{j=1}^{t} \lambda_{i(j-1)}[j^\alpha - (j-1)^\alpha]\right), \tag{7.5}$$

and

$$f_i(t) = [1 - \exp(-\lambda_{i(t-1)}[t^\alpha - (t-1)^\alpha])] \exp\left(-\sum_{j=1}^{t-1} \lambda_{i(j-1)}[j^\alpha - (j-1)^\alpha]\right), \tag{7.6}$$

respectively.

Discrete analogies of the continuous model framework for the long-term survivor and mixture models are also required. For the former only the survival process for the susceptible group is explicitly modelled - the resistant premises are assumed to be immune to infection. This is represented by a latent Bernoulli variable Z_i with parameter $(1 - p_i)$ for premise i, where $Z_i = 1$ if susceptible and 0 if resistant. In this case the survival process for the susceptible group is equivalent to the conventional survivor function (7.5). This

leads to the overall survivor function for a premise i at time t to be

$$S_i(t) = p_i + (1 - p_i) \left\{ \exp\left(-\sum_{j=1}^{t} \lambda_{i(j-1)} [j^\alpha - (j-1)^\alpha] \right) \right\}, \tag{7.7}$$

where p_i is the probability of resistance with $\text{logit}(p_i) = \gamma_0 + \gamma_1 y_i$. This version of the long-term survivor model includes covariates in both the probability of resistance and the component hazard for the susceptible group. As discussed in Banerjee and Carlin (2004) care must be taken since improper priors lead to improper posteriors. In this case non-informative proper priors will be used instead.

The marginal survivor function for resistant premises is equal to one always (i.e. it will never become infected). Conversely the marginal probability function is $f_i(t) = 0$ for resistant premises and given by (7.6) for susceptible premises. The overall probability function for a premise i at time t is therefore:

$$f_i(t) = (1-p_i)\{1-\exp(-\lambda_{i(t-1)}[t^\alpha-(t-1)^\alpha])\} \exp\left(-\sum_{j=1}^{t-1} \lambda_{i(j-1)} [j^\alpha - (j-1)^\alpha] \right). \tag{7.8}$$

The mixture model is similar to the long-term survivor model except that both the susceptible and resistant groups have non-degenerate survival processes. As before the survivor function for susceptible premises (SS_i) is given by (7.5) and for the resistant premises (SR_i) by

$$SR_i(t) = SR(t) = \exp\left(-\lambda \sum_{j=1}^{t} [j^\alpha - (j-1)^\alpha] \right). \tag{7.9}$$

The hazard for the resistant group is assumed independent of any covariate influence ($\lambda = \exp(\eta_0)$), and the shape parameter α to be the same as that for the susceptible process.

The survivor function for a premise i at time t for the mixture approach is then given by:

$$S_i(t) = p_i SR(t) + (1 - p_i) SS_i(t). \tag{7.10}$$

The overall hazard functions for both the mixture and long-term survivor models can be

determined from the formula

$$h_i(t) = \frac{f_i(t)}{S_i(t)}, \tag{7.11}$$

derived in section 4.1.

A further issue is the possible inclusion of frailty effects into the models. In the case of the mixture and long-term survivor models, the possible presence of spatial structure in both the component hazards and the mixing may result in confounding of the parameter estimates, leading to difficulties in inference. As a result of this we do not use frailties in the following model fits.

7.1.4 Prediction

To test and compare the predictive power of the different approaches the simulated epidemics were censored via exposure at 43 days and the parameter estimates used to predict over the remaining week. The correct threshold for exposure was assumed known, no frailty effects were used and the predictive algorithm described in section 5.3 was adjusted to allow for susceptibility and censoring via exposure. An extension for the latter issue was covered in section 5.8, but to avoid confusion we will amalgamate all previous versions of the predictive algorithm, including an adjustment for susceptibility, to one form described below.

Using the definitions in section 5.3, consider a matrix of K posterior samples for the m parameters obtained from a model fitted at time t. Remove any premises from the data set that are neither susceptible to infection nor infective (e.g. any that have been culled or vaccinated for example), and then split the data into two groups - IPs and UIPs. The IP group (n_{inf} premises) consists of all premises that are infected and contagious and the UIP group (n_{cens} premises) consists of all uninfected premises.

For the UIPs set up an indicator matrix $\mathbf{\Phi}$ with the number of rows equal to the number of premises ($i = 1, \ldots, n_{\text{cens}}$) and the number of columns equal to the number of posterior samples. Initially set each row equal to zero ($\phi_{i\cdot} = 0$). Let $\mathbf{T}$ be the corresponding matrix

of predicted survival times with elements $\{t_{ik}\}$.

Define $w_i(\nu)$ to be a time-dependent covariate for an individual i, where $x_i(\nu) = g(w_i(\nu))$ drives the epidemic process. A pre-determined threshold value, w^*, applied to $w_i(\nu)$ determines exposure at time ν such that an indicator variable δ^e_{ik} $(i = 1, \ldots, n)$, is defined as:

$$\delta^e_{ik} = \begin{cases} 1 & \text{if } \exists\ w_i(u), u = 0, \ldots, t, \text{ such that } w_i(u) - w^* > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, let $\boldsymbol{T}^e$ be a matrix of exposure times with elements $\{t^e_{ik}\}$ where initially

$$t^e_{ik} = \begin{cases} \inf\{u : w_i(u) - w^* > 0; u = 0, \ldots, t\} & \text{if } \delta^e_{i\cdot} = 1, \\ 0 & \text{otherwise.} \end{cases}$$

i.e. t^e_{ik} is the date of exposure for all currently exposed premises at time t.

Finally, if y_i is an observed covariate relating to resistance then define δ^s_{ik} $(i = 1, \ldots, n)$ to be an indicator variable for an uninfected individual such that

$$\delta^s_{ik} = \begin{cases} 1 & \text{if individual } i \text{ susceptible,} \\ 0 & \text{otherwise.} \end{cases}$$

The full algorithm for predicting the course of the epidemic over a time period E is:

1. Set $k = 1$, $\nu = t$, $\text{IP}^{(\nu)}$=IP and $\text{UIP}^{(\nu)}$=UIP.

2. Take k^{th} set of posterior samples and calculate $w_i(\nu)$ (using $\text{IP}^{(\nu)}$ if necessary) and $p_i = \frac{\exp(\gamma_0+\gamma_1 y_i)}{1+\exp(\gamma_0+\gamma_1 y_i)}$, where γ_0 and γ_1 are parameters estimated by the model.

3. Let u_i be a random sample from a $U(0, 1)$ distribution corresponding to premise i and set $\delta^s_{ik} = 1$ if $u_i < p_i$ and $\delta^s_{ik} = 0$ otherwise.

4. If $w_i(\nu) - w^* > 0$ and $\delta^e_{ik} = 0$, then set $\delta^e_{ik} = 1$ and $t^e_{ik} = \nu$. Calculate $x_i(\nu) =$

$g(w_i(\nu))$ and

$$h_i(\nu \mid x_i(\nu)) = \begin{cases} P_j(\nu - t^e_{ik} \leqslant T < \nu - t^e_{ik} + 1 \mid T \geqslant \nu - t^e_{ik}, x_i(\nu)) & \text{if } \delta^e_{ik} = 1 \text{ and } \delta^s_{ik} = j, \\ 0 & \text{otherwise,} \end{cases}$$

for all uninfected premises (using $\text{IP}^{(\nu)}$ if necessary). Here $j = 1, 2$, and $P_0(\cdot)$ relates to the hazard for resistant premises and $P_1(\cdot)$ to the hazard for susceptible premises.

5. Sample a new u_i from a $U(0, 1)$ distribution for premise i.

6. If $u_i < h_i(\nu \mid x_i(\nu))$ and $\phi_{ik} = 0$ then set $t_{ik} = \nu$ and $\phi_{ik} = 1$.

7. Set $\nu = \nu + 1$. Update $\text{IP}^{(\nu)}$ to include all new infected premises (i.e. $\{\text{UIP}^{(\nu-1)} \mid \phi_{ik} = 1\}$). Update $\text{UIP}^{(\nu)}$ such that $\text{UIP}^{(\nu)} = \{\text{UIP}^{(\nu-1)} \mid \phi_{ik} = 0\}$.

8. If $\nu > E$ or there are no more uninfected premises remaining then go to step 9. Else go to step 4.

9. Set $t_{ik} = E$ for all remaining censored premises, set $k = k + 1$, $\nu = t$, $\text{IP}^{(\nu)}$=IP and $\text{UIP}^{(\nu)}$=UIP.

10. If $k \leqslant K$ then return to step 4; else END.

7.1.5 Comparative results

The formulation of the WinBUGS code for models with non-standard likelihoods is discussed in appendix A. The shape parameter α and mixing parameter γ_1 were both given $G(0.1, 10)$ priors to ensure positivity. In the latter case this was to prevent label switching (though technically we constrain this parameter to be negative in the model so that an increase in y leads to an increase of susceptibility). The linear intercept, β_0, in the scale parameter of the hazard for the susceptible group was assumed $N(0, 10)$ distributed, as was the mixing intercept γ_0. The linear intercept in the scale parameter of the hazard for the resistant group, η_0, was also given a $N(0, 10)$ prior.

The mixture model was run for longer, with a burn-in of 10000 and a further 70000 updates. The long-term survivor model had a 20000 burn-in followed by 40000 updates and the conventional model 5000 and 40000 respectively. Two chains were used in each case and initial values generated as before.

Table 7.2 gives a comparative account of some summary statistics for the estimated parameters returned from the different model fits. In each case the convergence and mixing was reasonable, and the mean, median, standard error and 2.5^{th} and 97.5^{th} percentiles were returned along with the potential scale reduction factor $\hat{R}$ and the effective number of parameters n_{eff}.

Table 7.3 gives some summary results for the corresponding predicted infection times relating to different aspects of the prediction process. The first part of the table gives the total number of premises currently uninfected (at 43 days) and the number of those that are susceptible. The second part of the table compares the actual number of premises that become infected in the subsequent week with the number predicted by the models, followed by the proportion of these actual subsequent infections that were correctly predicted. The third part of the table is analogous to the second, but corresponding to the susceptible population only. The final part gives the proportion of premises correctly predicted to be resistant or susceptible by the models.

We can see from table 7.2(a) that when there is no resistance to infection in the data set the parameter estimates governing the localised spreading process (α, β_0 and β_1) obtained from the conventional model fit are reasonably accurate. Furthermore we can see that the long-term survivor and mixture model results reinforce the results from the conventional model; indeed it can be seen from table 7.3 that all premises included in the model fit were predicted to be susceptible by both the mixture and the long-term survivor models. It is worth noting at this point that the estimates for the η_0 parameter governing the spark infection process are not comparable to the true value used in the simulation. This is because the model fit relates to a baseline hazard relative to time from exposure rather than absolute time from the beginning of the epidemic as was used in the simulation.

(a) No resistance

			Mean	**s.e.**	**2.5%**	**50%**	**97.5%**	$\hat{R}$	n_{eff}
Parameters (actual values)	α (2)	Conv	1.9159	0.1598	1.6218	1.9105	2.2311	1.0043	1000
		LTS	1.8915	0.1510	1.5980	1.8910	2.1900	1.0022	960
		Mix	1.9140	0.1580	1.6090	1.9135	2.2260	1.0044	340
	β_0 (-5.91)	Conv	-5.7564	0.3732	-6.5371	-5.7520	-5.0559	1.0071	1000
		LTS	-5.6936	0.3440	-6.3615	-5.7000	-5.0039	1.0040	410
		Mix	-5.7468	0.3728	-6.5001	-5.7285	-5.0269	1.0051	300
	β_1 (0.043)	Conv	0.0439	0.0027	0.0384	0.0439	0.0496	1.0041	370
		LTS	0.0438	0.0027	0.0383	0.0438	0.0493	1.0017	880
		Mix	0.0438	0.0028	0.0389	0.0436	0.0494	1.0000	1000
	γ_0 (NA)	LTS	-1.3516	3.5610	-8.4250	-1.1745	5.2545	1.0139	110
		Mix	-0.8740	2.6827	-5.6912	-0.9245	4.2907	1.0010	1000
	γ_1 (NA)	LTS	3.5537	5.2158	0.0000	1.6955	17.9607	1.1479	64
		Mix	1.0337	0.9822	0.0504	0.7273	3.5161	1.0042	360
	η_0	Mix	0.4035	3.0894	-5.4662	0.4953	6.4595	1.0003	1000

(b) Low resistance

			Mean	**s.e.**	**2.5%**	**50%**	**97.5%**	$\hat{R}$	n_{eff}
Parameters (actual values)	α (2)	Conv	1.4498	0.1137	1.2350	1.4445	1.6721	1.0156	99
		LTS	1.7294	0.2048	1.3421	1.7270	2.1596	1.0134	120
		Mix	1.6529	0.1910	1.2900	1.6540	2.0173	1.0004	1000
	β_0 (-5.91)	Conv	-3.9057	0.2800	-4.4642	-3.8900	-3.3650	1.0203	77
		LTS	-5.2869	0.4754	-6.2446	-5.2740	-4.3853	1.0122	130
		Mix	-5.0914	0.4431	-5.9485	-5.0795	-4.2549	1.0004	1000
	β_1 (0.043)	Conv	-0.0024	0.0011	-0.0049	-0.0023	-0.0006	1.0083	320
		LTS	0.0427	0.0042	0.0346	0.0426	0.0516	1.0010	980
		Mix	0.0430	0.0041	0.0350	0.0429	0.0512	1.0050	1000
	γ_0 (6.91)	LTS	6.4815	1.0820	4.4749	6.5000	8.5161	1.0131	120
		Mix	6.8882	1.2190	4.7070	6.8035	9.4391	1.0011	1000
	γ_1 (0.23)	LTS	0.2428	0.0338	0.1777	0.2434	0.3101	1.0174	100
		Mix	0.2322	0.0369	0.1660	0.2302	0.3081	1.0006	1000
	η_0	Mix	-9.0891	1.3120	-12.1408	-8.9865	-6.8989	1.0009	1000

Table 7.2: Posterior parameter estimates from models fitted to simulated epidemics with varying resistance to infection at 43 days

(c) Medium resistance

			Mean	s.e.	2.5%	50%	97.5%	$\hat{R}$	n_{eff}
Parameters (actual values)	α (2)	Conv	1.2385	0.1067	1.0339	1.2370	1.4511	1.0019	1000
		LTS	1.8097	0.1911	1.4379	1.8015	2.2081	1.0023	640
		Mix	1.6865	0.2079	1.2998	1.6800	2.1091	1.0020	730
	β_0 (-5.91)	Conv	-3.8245	0.2761	-4.3861	-3.8205	-3.2967	1.0007	1000
		LTS	-4.5728	0.4098	-5.3782	-4.5695	-3.7809	1.0009	1000
		Mix	-5.1958	0.5025	-6.1822	-5.2025	-4.2460	1.0019	800
	β_1 (0.043)	Conv	-0.0027	0.0011	-0.0050	-0.0027	-0.0008	1.0006	1000
		LTS	0.0157	0.0019	0.0120	0.0158	0.0193	1.0023	660
		Mix	0.0423	0.0041	0.0342	0.0423	0.0506	1.0000	1000
	γ_0 (6.91)	LTS	5.5601	0.6677	4.2789	5.5615	6.9101	1.0116	160
		Mix	6.9506	0.9581	5.1807	6.9300	8.8927	1.0001	1000
	γ_1 (0.23)	LTS	0.2116	0.0239	0.1679	0.2105	0.2608	1.0070	250
		Mix	0.2322	0.0308	0.1748	0.2308	0.2930	1.0002	1000
	η_0	Mix	-8.9272	0.9640	-10.8703	-8.9355	-7.2066	1.0003	1000

(d) High resistance

			Mean	s.e.	2.5%	50%	97.5%	$\hat{R}$	n_{eff}
Parameters (actual values)	α (2)	Conv	1.3671	0.1163	1.1449	1.3680	1.5951	1.0029	510
		LTS	1.8789	0.1775	1.5480	1.8730	2.2290	1.0021	710
		Mix	1.7792	0.2053	1.3908	1.7820	2.1770	1.0000	1000
	β_0 (-5.91)	Conv	-4.4338	0.2965	-5.0311	-4.4260	-3.8590	1.0035	430
		LTS	-4.6896	0.4136	-5.4682	-4.6880	-3.9567	1.0041	370
		Mix	-5.4700	0.5061	-6.4783	-5.4645	-4.4870	1.0016	930
	β_1 (0.043)	Conv	-0.0035	0.0012	-0.0060	-0.0035	-0.0013	1.0033	450
		LTS	0.0074	0.0013	0.0044	0.0075	0.0096	1.0045	500
		Mix	0.0410	0.0040	0.0331	0.0410	0.0490	1.0045	330
	γ_0 (6.91)	LTS	4.6158	0.5892	3.5359	4.6055	5.8160	1.0168	110
		Mix	5.3000	0.6482	4.0637	5.2540	6.5872	1.0017	1000
	γ_1 (0.23)	LTS	0.1910	0.0241	0.1487	0.1900	0.2406	1.0145	130
		Mix	0.1777	0.0241	0.1336	0.1757	0.2294	1.0025	1000
	η_0	Mix	-8.4632	0.8467	-10.0803	-8.4890	-6.8179	1.0002	1000

Table 7.2: Posterior parameter estimates from models fitted to simulated epidemics with varying resistance to infection at 43 days (cont.)

		Numbers of premises		Infections in next week			Infections to susceptible premises in next week			Susceptibility
		UIPs	Susceptible UIPs	Actual	Predicted	Correct predictions	Actual	Predicted	Correct predictions	Correct predictions
No resistance	Conv	741	741	120	110	76.67%	120	110	76.67%	NA
	LTS	741	741	120	111	76.67%	120	111	76.67%	100.00%
	Mix	741	741	120	111	77.50%	120	111	77.50%	100.00%
Low resistance	Conv	883	563	72	0	0.00%	68	0	0.00%	NA
	LTS	883	563	72	45	43.06%	68	45	45.59%	85.28%
	Mix	883	563	72	36	37.50%	68	36	39.71%	85.28%
Medium resistance	Conv	875	415	66	0	0.00%	61	0	0.00%	NA
	LTS	875	415	66	40	40.91%	61	40	44.26%	84.80%
	Mix	875	415	66	34	31.82%	61	34	34.43%	86.63%
High resistance	Conv	867	240	46	0	0.00%	39	0	0.00%	NA
	LTS	867	240	46	25	32.61%	39	25	38.46%	82.24%
	Mix	867	240	46	18	30.43%	39	18	35.90%	88.70%

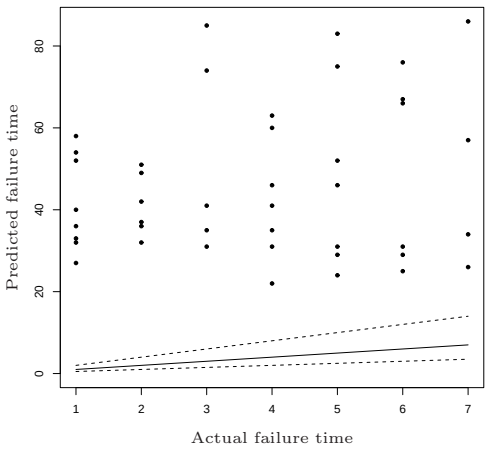
Table 7.3: Posterior predicted summary values for models fitted to simulated epidemics at 43 days

We can also see that as the level of resistance in the data set increases the conventional model begins to break down. This is most evident in table 7.3 and results in vastly over-predicted survival times. This further backs up the results from the non-spatial simulation conducted in the previous chapter, where even small levels of resistance can cause large inaccuracies in the predicted survival times. In contrast the long-term survivor and mixture models perform markedly better, and give good results in predicting resistance in the data (even though the mean predicted survival times are still too large). This is illustrated also by the plots in figure 7.3 that show the actual vs. mean predicted survival times (post censoring) for future IPs from each of the conventional, long-term survivor and mixture models, in the case where there is a high level of resistance in the data set. Predicted survival times of greater than 100 days post censoring were left out for clarity (3, 4 and 9 premises for each of the models respectively). The dashed lines correspond to Parkes' estimates of 'serious error' as described in section 6.3.

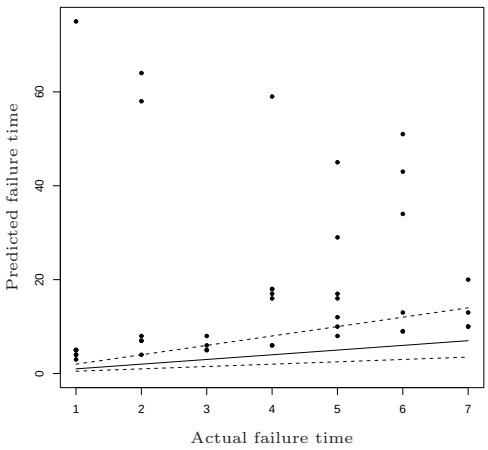
This is also seen in tables 7.2(b), 7.2(c) and 7.2(d), where the parameter estimates for γ_0 and γ_1 are relatively good even in the case where there is low resistance being exhibited. The estimates for α, β_0 and β_1 are likewise markedly better than those for the conventional model in each case, although the individual level survival predictions are unreliable (see section 6.3). We will explore some techniques for eliciting useful information on a global scale in the next section.

7.2 Spatial hazard maps and simulated work

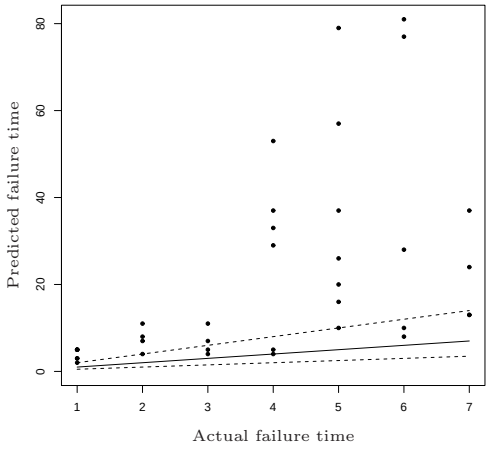
In the previous chapter some potential advantages of using the long-term survivor and mixture models over the conventional survival model were explored in a situation where resistance to infection was evident in the data set. The results show that under certain circumstances (i.e. that the mixing is well defined) the mixture and long-term survivor models help to alleviate some of this bias that arises from the misclassification of resistant premises. There remains a lot of variation in the premise-specific predictions of survival times for censored premises when compared to their actual values however.



(a) Conventional model



(b) Long-term survivor model



(c) Mixture model

Figure 7.3: Plots of actual vs. predicted failure times from spatial simulation with high levels of resistance

So the question arises: how can the results from these models be realistically used in the understanding, prediction and implementation of control policies during an actual animal disease epidemic?

The majority of modelling work done on infectious animal disease epidemics focuses on estimation of the basic reproductive number R_0, giving the mean number of secondary infections from each primary infection. Spatial maps of R_0 help to identify areas of high- and low-risk of infection, where $R_0 > 1$ relates to areas where an epidemic situation is highly likely and $R_0 < 1$ to areas where the infection will eventually die out. Though not identical, the discrete hazard function is at least comparable to this in some sense, since it gives the conditional probability of infection in the next day given survival to that point. (This allows spatial maps to be produced in which 'high-risk' areas can be identified - though there is not a threshold attached as in the case of R_0.)

A problem occurs when making inferences about the estimated values of the hazard. The results in section 7.1.5 show that even in the case where all factors affecting disease spread are known, there is still large statistical variation in the infective process that makes accurate prediction of survival times and hazards very difficult. Care must therefore be taken when using these estimates to determine the scale of the epidemic in relation to absolute probabilities and numbers of infections at each time point. It is still possible to use the hazard maps to identify areas of high- and low-risk based on the relationships between the estimations; e.g. the models should capture some kind of 'relative' spatial and temporal variability in risk even though the actual probabilities of infection for individuals are often underestimated.

In addition to this, if a series of spatial maps are produced over time then the spatial aspect can be integrated out to give the hazard over time for the entire region. (Note that since a Bayesian framework is being used, full posterior distributions for the predictions can be obtained - though in this case the spatial maps are based on a `loess` smoothed map of the means.)

To investigate this an epidemic was simulated over a period of 50 days using a spatially

random sample of 2000 Devon premises as its base. This was conducted using the same method as in section 7.1.2, though this time the grouping covariate y was given some spatial structure (in order to keep the spatial structure the same the mixing parameters were allowed to change to reflect differing degrees of resistance). In addition we also added an incubation period of 3-7 days between the date of infection and the date of report. The VL covariate for each model was calculated using only those IPs that had been reported at that point in time (though it was assumed that the date of the oldest lesion found on a premise could be accurately quantified and reflected the actual date of infection).

Three source premises were used and to prevent the epidemic from growing too quickly a contiguous 'cull' policy was also introduced two weeks after the initial infection. Here contiguity was determined by a probabilistic process, in which two premises were classified as 'contiguous' with a high probability if they were within 1km of each other, and with a low probability if they were within 2km of each other. This seemed a reasonable method to estimate contiguity and/or dangerous contacts in the population of susceptible UIPs; acting essentially as a cross between the actual CP/DC used across the entire UK in 2001 and the 3km ring cull that was applied in certain regions.

The 3km ring cull policy was not used in the simulation for two reasons: the first because it was never implemented in Devon in 2001 and the second because it was far too aggressive, especially considering that the simulation was based on a spatially randomly-thinned subset of the population. The 2km contiguous policy described above, whilst not exactly replicating the response used in Devon, did at least capture the notion of contiguity whilst balancing the number of premises culled to something more reasonable. It also reflected the increased aggression of the simulated infective process compared to the actual one. Note also that we make no distinction between culling and vaccination in this sense. The assumption is that the infected and uninfected animals are removed from the data on the date of removal, and that no residual viral excretion remains. We refer to this procedure as 'culling' for consistency.

Here IPs were culled within 24hrs of infection, and CP/DC UIPs within 48hrs of infection.

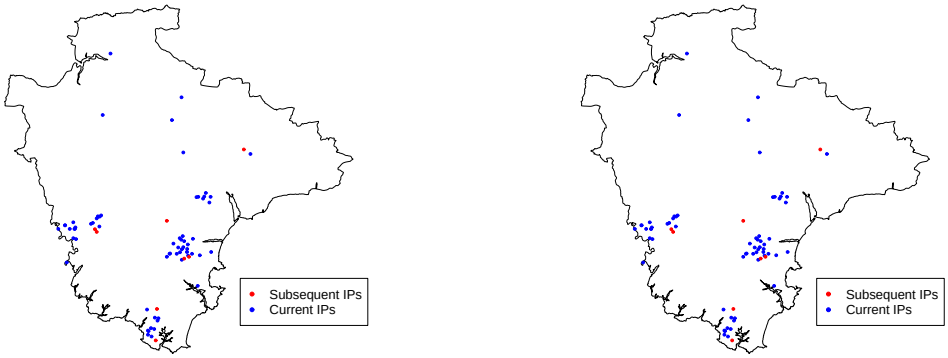
A prompt cull strategy was therefore assumed from the outset. In the real epidemic the 24/48hr prompt cull policy was introduced at the end of March 2001, and before this it sometimes took up to a week before IPs were culled. In addition the simulation also assumes that this 24/48hr period includes time to disposal of the slaughtered carcasses and that the disposal mechanism does not affect the infective process.

7.2.1 Hazard maps for simulated epidemic

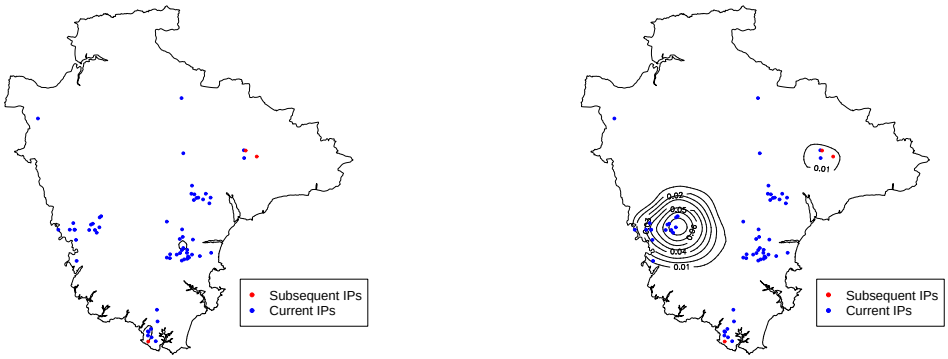
Figure 7.4 gives comparative contour maps of hazards of infection from the conventional and mixture models fitted between day 14 and 36 of the simulated epidemic at two day intervals. These help to show how the epidemic progressed over time and space. It can be seen that the mixture model shows much more variation in the predicted hazard surface than the conventional model, due to the fact that it is capturing some of the variation relating to susceptibility. (Note that all these maps are on the same scale to aid interpretation.)

These maps are useful in identifying the effects of localised spread and 'spark' infections on the path of the epidemic. If fitted on a day-by-day basis and the average taken over space then the hazard over time can be obtained, which if scaled by the number of uninfected premises at each time point provides an estimate of the number of premises expected to become infected. Figure 7.5 shows this estimate plotted against the actual number of IPs at each day.

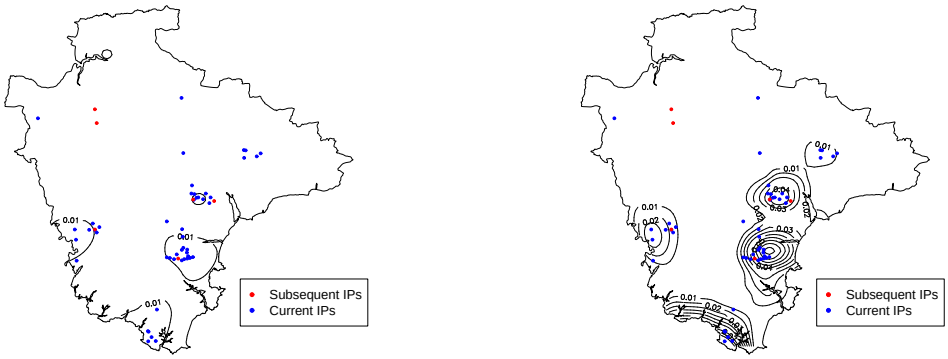
It is clear that the mixture model is capturing far more of the variability than the conventional model, and from figure 7.5 it can be seen that the mixture seems to capture the shape of the epidemic curve well. It does seem to overpredict the risk of infection when compared to the actual numbers of infections, but it is a marked improvement on the smooth estimate obtained from the conventional curve.



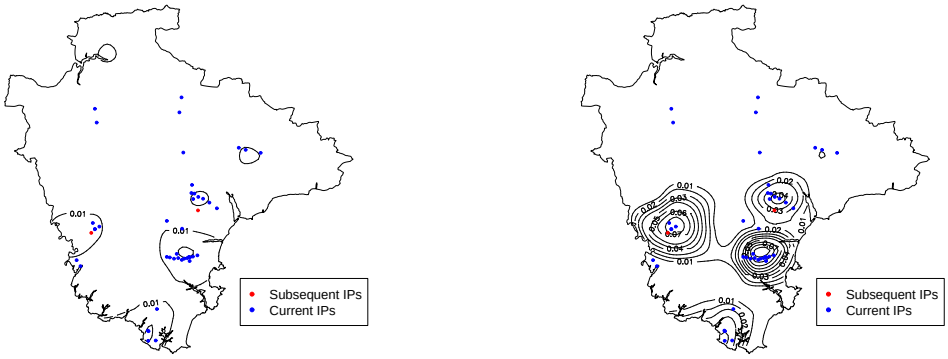
(a) Day 14



(b) Day 16

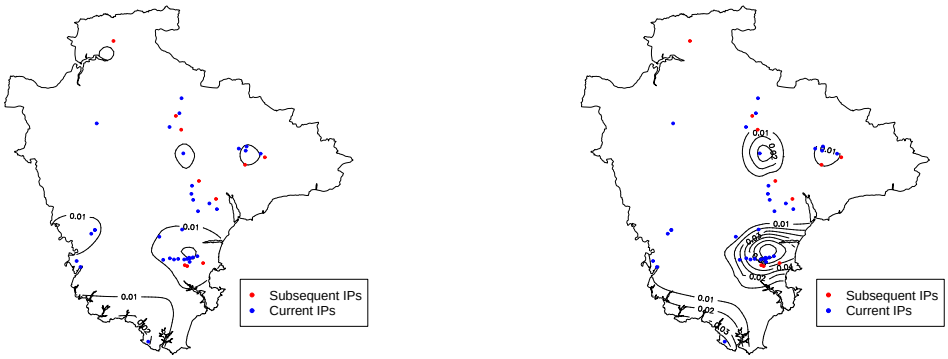


(c) Day 18

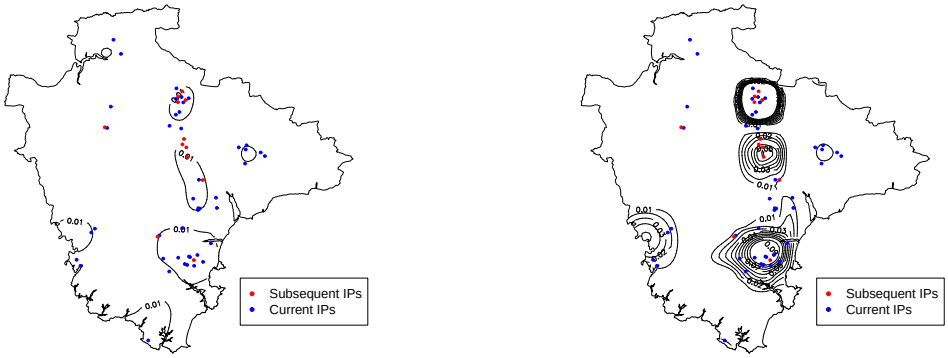


(d) Day 20

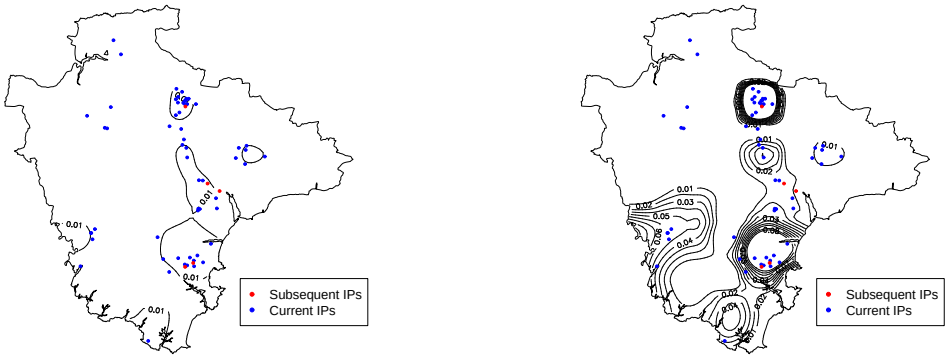
Figure 7.4: Comparative contour maps of hazards of infection in the next day, from conventional (left) and mixture (right) models fitted from day 14 of simulated epidemic



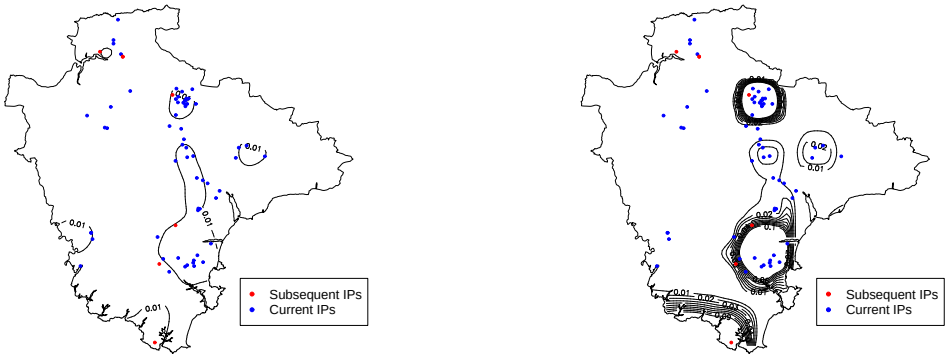
(e) Day 22



(f) Day 24

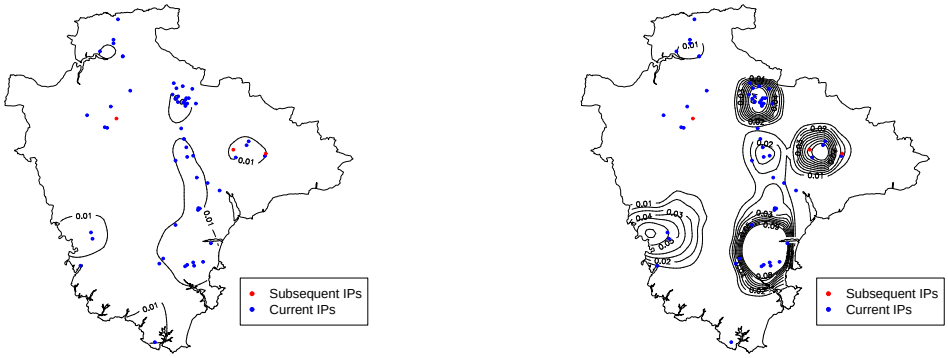


(g) Day 26

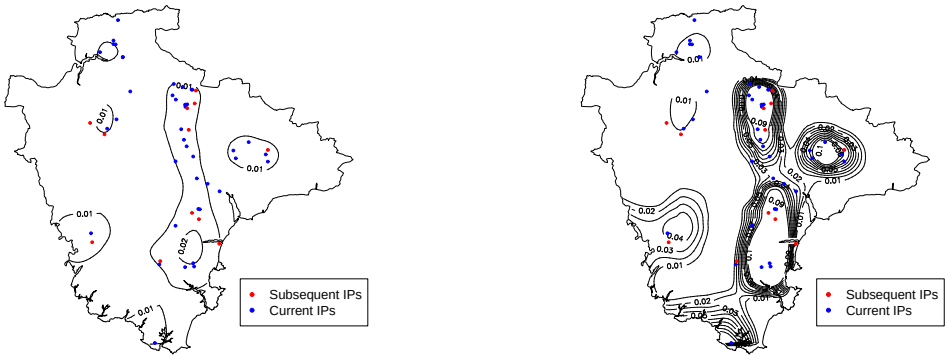


(h) Day 28

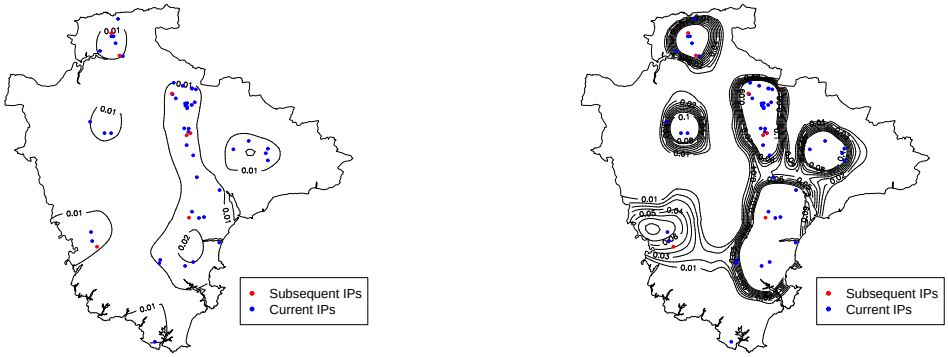
Figure 7.4: Comparative contour maps of hazards of infection in the next day, from conventional (left) and mixture (right) models fitted from day 14 of simulated epidemic (cont.)



(i) Day 30



(j) Day 32



(k) Day 34

Figure 7.4: Comparative contour maps of hazards of infection in the next day, from conventional (left) and mixture (right) models fitted from day 14 of simulated epidemic (cont.)

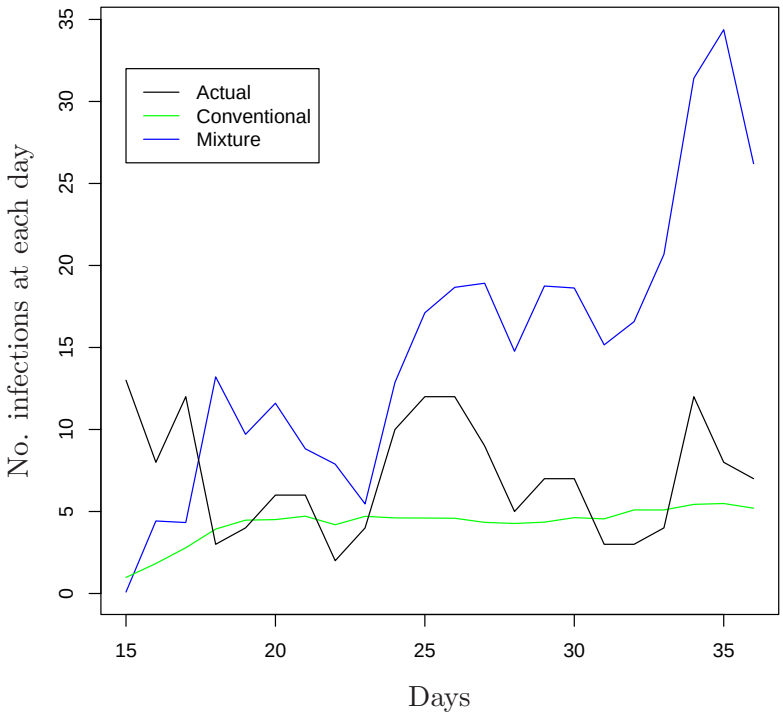


Figure 7.5: Estimated hazard over time for simulated epidemic (weeks 3-5)

7.2.2 Predictive uses

The hazard maps produced in the previous section give the probability of infection in the next day given survival to that point. These are useful in identifying areas of high- and low-risk and can perhaps be used to help guide reactive control policies; however the one day time frame required between successive model fits perhaps limits the effectiveness of these policies due to the difficulties associated with their implementation over such a short space of time. It may be better to consider the conditional probability of infection over a longer period of time as a means of predicting risk.

The nature of the epidemic is that the future course at each day is dependent on the history up to that point. Care must be taken to ensure that the period of time over which predictive inference is required is not too long, since even small numbers of influential and unforecasted 'spark' infections can significantly change the entire dynamics of the epidemic process. It was decided to limit the period of time over which to predict to be one week. This allows a longer forecast than the ordinary hazard maps, providing a more informative

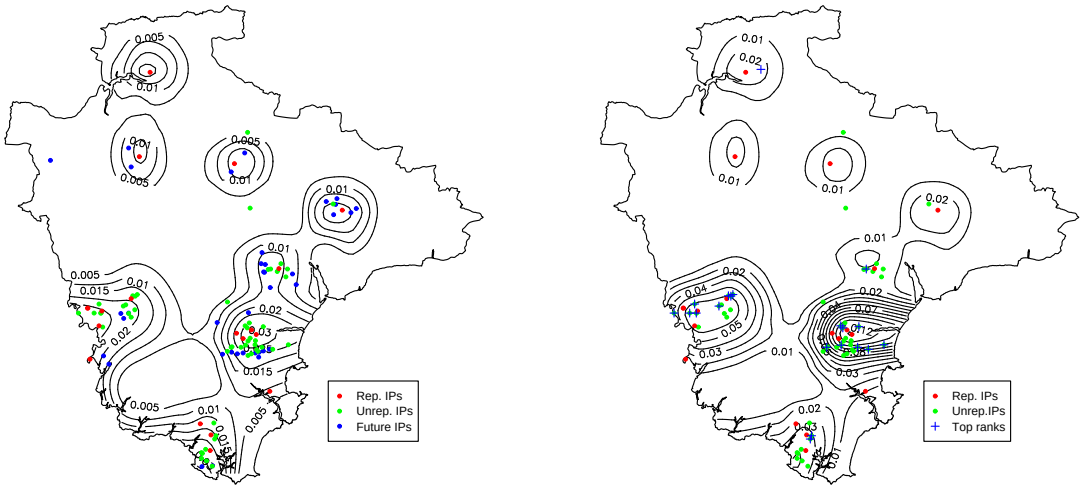
time-scale over which to implement control policies, whilst limiting the deleterious effects of rogue spark infections on the long-range predicted course of the epidemic.

Spatial maps of the hazard of infection in the next week were produced, along with a second set of maps based on ranking premises according to their predicted future infection time. Examples of these are shown in figure 7.6 for models fitted at two, three and four weeks respectively. In the case of the rank map each posterior predictive sample was ordered by time to infection, and predictive posterior distributions for the probability of being in the next r infections were recorded.

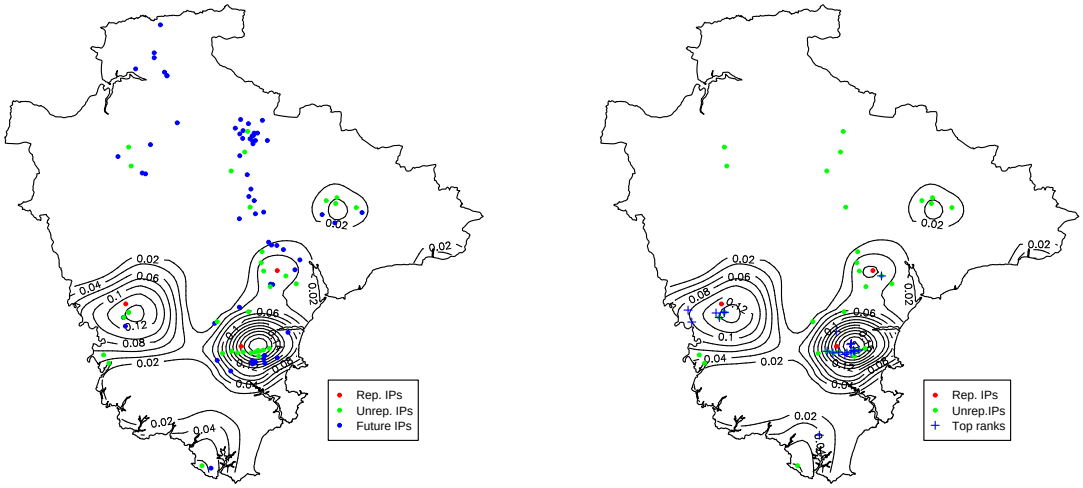
On each of the hazard and rank maps the green points correspond to those premises that are currently infected (and infective) but have yet to be reported, and the red points correspond to reported and infective premises (i.e. the ones contributing to the VL). So the simulated infective process is driven by both the green and red premises, but the predictive model is driven by the red premises only.

On the hazard map the blue points give the locations of premises that subsequently become *infected* (but not necessarily reported) in the next week, and on the rank map (generated based on the top 20 ranks) the blue crosses relate to the actual top 20 premises in order of mean predicted rank. It is worth noting that the blue crosses can relate to premises that have yet to be reported, and a similar argument regarding the locations of the most 'at-risk' premises could also apply to the map of the hazard, though they have not been shown here.

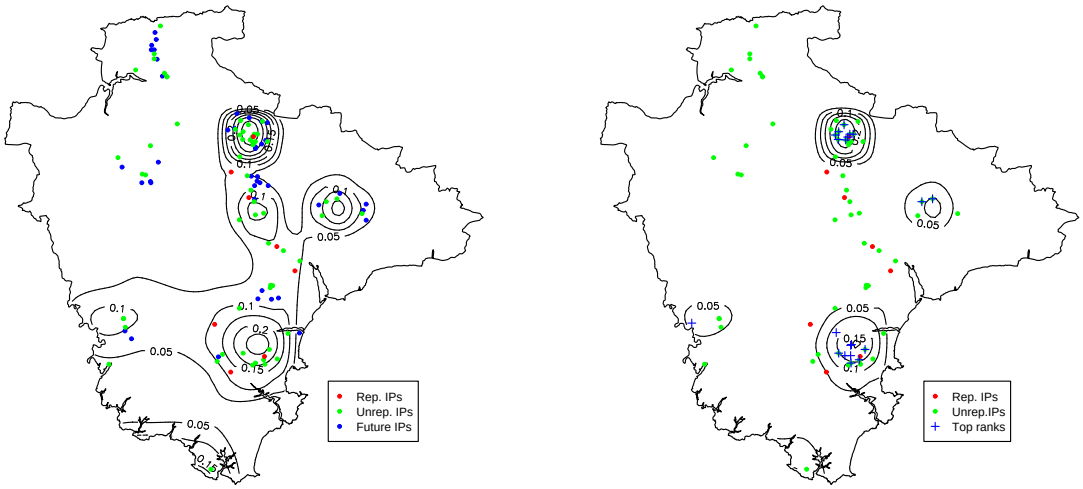
The important point is that the predicted maps seem to capture the dynamics of the epidemic reasonably well with regard to identifying areas of high- and low-risk in relation to the information gleaned from the reported IPs. However as shown, particularly in the hazard map in figure 7.6(b), if the disease is aggressive then latent infections can cause difficulties with accurate prediction. In practice of course the predictive maps can be updated as the epidemic progresses to encompass new information such as spark infections, new reports or changes in control policy that may have a significant impact on the predictions, and this is a desirable feature of our VL approach.



(a) Day 14



(b) Day 21



(c) Day 28

Figure 7.6: Predictive risk maps of probability of infection in next week (left) and probability of belonging to top twenty future IPs (right)

Also, in contrast to the previous section, each of these maps is on a different scale across the time periods. In addition it is also worth noting that the maps based on ranking premises seem to be a bit more localised, though of course this will depend greatly on the extent of ranking used to generate the probabilities (since the same span has been used in the `loess` smooth in both all cases). These maps act as potentially useful guides to the evolution of the outbreak, and the next section will discuss possible ways in which they may be used to help influence control policies.

7.2.3 Targeting control policies

A lot has been written in both the veterinary literature and the media regarding the choice and effectiveness of control policies for animal disease epidemics (the recent furore surrounding the role of badgers in the spread of bovine tuberculosis is a prime example). In the case of the 2001 UK FMD epidemic there has been a lot of research, conducted both at the time of the outbreak and in hindsight, investigating the effects of various control procedures. Recently work has focussed on whether alternative strategies may have been more effective (DEFRA 2002a), not only in reducing the size of the epidemic but also the numbers of animals culled.

Ferguson et al. (2001a) investigated the relative effects of contiguous culls, ring culls and ring vaccination strategies on the basic reproductive number R_0. Published just under 3 months after the outbreak began in May 2001, it concluded that a ring cull or vaccination strategy would be essential in bringing the epidemic under control, although the ring cull was predicted to be more efficient in reducing the extent of the epidemic than vaccination. This was reinforced further by Ferguson et al. (2001b), published in October 2001, which concluded that the ring cull strategy was essential in bringing the epidemic under control in the parts of the UK in which it was used. They note however that it would have been far more effective if the policy had been instigated earlier. Keeling et al. (2001a), also published in October 2001, supports the point that an intensive neighbourhood culling policy is key to controlling disease spread; though the authors question whether an extended 3km

ring cull was necessary, arguing that a prompt (24/48hr) CP/DC cull implemented far sooner would have been more effective. They question to what extent the decline in the epidemic was due to culling alone, or whether other factors such as reduced numbers of susceptible premises had a large effect. In addition they note that the definition of a neighbourhood surrounding an IP will be situation- and disease-dependent, as can be seen by the contrasting dynamics of FMD spread in Devon and Cumbria.

A consensus common to all of these papers regards '... the importance of rapid implementation of properly focussed disease control policies' (Keeling et al. 2001a), whether that be culling, movement restrictions, increased biosecurity, vaccination or (as more likely) a combination of these. In addition to this are the logistical and financial constraints involved in instigating these policies, and many recent papers have sought viable solutions to these problems. Tildesley et al. (2006) investigate a reactive vaccination strategy for cattle that combines prompt IP and DC culling with a ring vaccination program. They focus on finding the optimal radius for a ring vaccination policy that minimises the effect of the epidemic but is constrained by the total number of cattle that can feasibly be vaccinated each day. They conclude that an optimum 35,000 livestock vaccinations a day would be more effective than using a prompt IP and DC cull combined with CP culling.

Morley and Chang (2004) investigate the effect of the standard policies used in the UK in 2001 on a potential outbreak in the USA. They conclude that the use of such policies would constitute a large risk to the USA if the disease were to enter the country. Instead they suggest a pre-emptive method of control that can reduce the time-lag between infection and report by using mobile PCR (polymerase chain reaction) units. These can detect the presence of the virus in RNA and DNA before clinical signs appear. The authors suggest stationing these units at points of entry into the country as well as coupling this with a mathematical model (in their case a cellular automata approach) to predict potential areas of high-risk to which mobile PCR units could be deployed should a source of the virus be identified.

These examples highlight the need for effective control policies and a means of determin-

	No. inf.	No. culled
No culling	749	
IP cull	587	494
Contig. cull	363	1433
Target cull	352	725

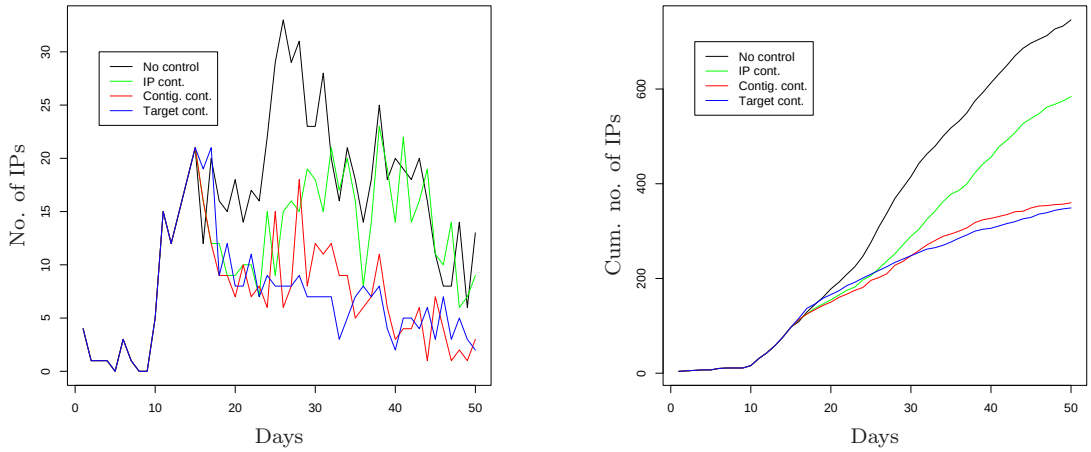
Table 7.4: Comparative numbers of infected and culled premises from simulated epidemics with varying control policies

ing the optimal focal points for the implementation of these procedures. In addition they should incorporate methodology that can deal with new sources of infection. A combination of different techniques, dependent on the situation, coupled with prompt instigation is probably the key to developing a balanced policy.

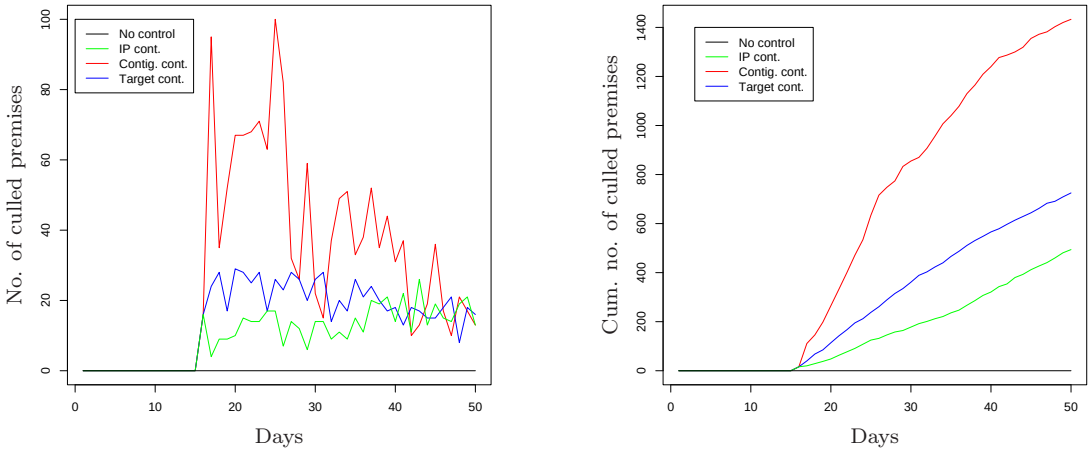
Here we do not attempt to investigate which is the better option, vaccination or culling, but simply offer a means of potentially targeting a control policies through the use of predictive spatial hazard maps of infection generated over one week periods. To do this four epidemics were simulated using identical data sets. The first had no cull policy, the second an IP-only cull, the third a 2km CP/DC (ring) cull policy and the fourth a targeted policy based on culling IPs and the top twenty most 'at-risk' premises according to the ranking predictions described in section 7.2.2.

The simulations were allowed two weeks to get established before culling was introduced. All IPs were culled within 24hrs of report and all CP/DCs within 48hrs of report. The ring cull worked on a day-by-day basis whereas the targeted cull worked by obtaining the top twenty 'at-risk' premises from a mixture model fitted at intervals of one week. A comparative plot of the epidemics is given in figure 7.7 with a corresponding table of results in table 7.4.

From these plots and results it seems that the target policy is certainly as effective, if not more effective than the contiguous cull policy in bringing the epidemic under control, resulting in less infected premises and substantially less culling. The IP-only cull results



(a) Plots of number of IPs



(b) Plots of numbers of culled premises

Figure 7.7: Plot of simulated epidemics with varying control policies

in less culled premises but more infections, and the no cull policy results in much larger numbers of IPs, though the results suggest that it is on its way to burning itself out on the basis of running out of susceptible premises to infect by the end of the 50 day period.

7.3 Application to real data set

The results from the previous section highlight the potential advantages of using a mixture or long-term survival approach over the conventional survival model when resistance to infection is present in the data set. However the accuracy of the posterior parameter estimates depends largely on how well the mixing is defined. For the long-term survivor model all infected premises are treated as susceptible, and this can confound the parameter estimates by not accounting for bias from non-localised resistant infections. The mixture model on the other hand can allow for this, but can't distinguish between susceptible infections caused by localised or non-localised sources. Also, these methods assume a resistance status for each premise that remains fixed for the entire duration of the epidemic. In our model this is based on some covariate measure relating to susceptibility. Two issues arise here, the first that a reasonable covariate measure be identified, and the second that the susceptibility status does not change over time.

In the latter case this could be encompassed by using a time-dependent covariate in the mixing probability, however for the time being we will treat the susceptibility status as fixed. With regard to defining a reasonable covariate to assess resistance, it seems sensible in the first instance to use total and species-specific uninfected animal densities in the same way as in chapter 5, however since it is considered fixed it will be taken to be the density *just prior* to infection or censoring.

We will fit just the mixture model here, since it seems to be more robust to the infection of resistant premises than the long-term survivor model. As a result of some preliminary model fits, we decided to parameterise our model slightly different to the standard form (7.10) used in the simulations. In order to try to gauge some indication about the sig-

nificance of the susceptibility parameters in explaining the resistance in the data, it was important that there weren't constraints placed on the mixing parameters (such as was used to control label-switching). Instead we decided to formulate the component hazards such that in the presence of no viral load they had the same baseline hazard function, and then force an ordering upon them through the regression parameters β_1 and β_2. So the model for the Devon data is given by:

$$S_i(t) = p_i SR_i(t) + (1 - p_i)SS_i(t), \tag{7.12}$$

where the two groups have survivor functions of the form (7.5), but $\lambda_{i(t-1)} = \exp(\beta_0 + \beta_1 \text{PAV}_{i(t-1)})$ for one group and $\lambda_{i(t-1)} = \exp(\beta_0 + \beta_2 \text{PAV}_{i(t-1)})$ for the other group constrained such that $\beta_1 < \beta_2$. The mixing parameter p_i was given by:

$$\log\left(\frac{p_i}{1-p_i}\right) = \gamma_0 + \gamma_1 AD_i, \tag{7.13}$$

or

$$\log\left(\frac{p_i}{1-p_i}\right) = \gamma_0 + \gamma_1 SD_i + \gamma_2 CD_i + \gamma_3 PD_i, \tag{7.14}$$

for total uninfected animal density, and species-specific densities respectively.

The model was fitted to the data set censored via exposure at 50 days with threshold 5×10^{-08}. Each of β_0, β_1 and the $\boldsymbol{\gamma}$ parameters were given $N(0, 10)$ priors, with α given a $G(0.1, 10)$ prior as before. The parameter β_2 was constrained to be greater than β_1 by letting $\beta_2 = \eta + \beta_1$ with η given a $G(1, 1)$ prior. Two chains were used with a burn-in of 10000 and a further 50000 updates. The posteriors were thinned to return 1000 samples.

A summary table of the posterior parameter estimates is given in table 7.5 with associated prediction summaries for the top ten highest risk premises given in table 7.6. The extent of overprediction exhibited in the results from the previous models fitted in chapter 5 appears to be less here, and the mean posterior estimates are derived from more complete posterior distributions (33% of samples returned a predicted infection time within the next 60 days for the most high-risk premise), though none of these top ten predicted premises

		Mean	s.e.	2.5%	50%	97.5%	$\hat{R}$	n_{eff}
Uninfected animal density	α	1.1090	0.0960	0.9294	1.1035	1.3100	1.0001	1000
	β_0	-7.0694	0.3283	-7.7662	-7.0560	-6.4700	1.0001	1000
	β_1	0.5244	3.1707	-5.6352	0.5470	6.6205	1.0008	1000
	β_2	1.4983	3.3196	-4.8698	1.6145	8.2122	1.0012	1000
	γ_0	-0.1114	3.0635	-6.4397	-0.0456	5.7362	1.0010	1000
	γ_1	-0.0913	3.1962	-6.2895	-0.1064	5.7877	1.0020	1000
Species-specific animal density	α	1.1004	0.0955	0.9283	1.0970	1.3060	1.0040	500
	β_0	-7.0431	0.3310	-7.7343	-7.0245	-6.4467	1.0037	410
	β_1	0.5309	3.2081	-5.8432	0.6158	6.5861	1.0060	250
	β_2	1.5066	3.3209	-5.0475	1.5585	7.8257	1.0089	170
	γ_0	-0.0778	3.1153	-6.1621	-0.1832	5.8501	1.0000	1000
	γ_1	-0.0409	3.2872	-6.5429	0.0187	6.2918	1.0017	870
	γ_2	0.0585	3.2378	-6.3455	0.0288	6.7122	1.0035	430
	γ_3	-0.0138	3.1020	-6.0193	-0.1042	5.9381	1.0009	1000

Table 7.5: Posterior parameter estimates from the mixture model (7.12) with regression covariate AV, and uninfected animal and species-specific densities in the mixing parameter - fitted to Devon data set 'censored via exposure' at 50 days

Uninfected animal density		Species-specific densities	
Mean survival time	**Proportion observed**	**Mean survival time**	**Proportion observed**
130	33.00%	132	32.60%
136	31.70%	137	31.60%
138	31.30%	138	31.60%
142	30.80%	143	30.40%
186	25.30%	214	23.10%
187	25.20%	216	22.70%
192	24.80%	228	21.80%
194	24.60%	231	21.60%
197	24.40%	234	21.20%
198	24.30%	241	20.70%

Table 7.6: Predictive output over a 60 day window for mixture model (7.12) with regression covariate AV, and uninfected animal and species-specific densities in the mixing parameter - fitted to Devon data set 'censored via exposure' at 50 days

were actual infections.

The regression parameters β_1 and β_2 were still not significantly different to zero and there was no label-switching. The covariates associated with susceptibility are clearly failing to capture this process adequately. A series of other models with slightly different specifications regarding the component hazards and mixing (analogous to the mixtures fitted to the simulated data) also backed up the results shown here.

7.4 Conclusions

In this chapter we have explored the uses of mixture and long-term survivor models in a spatial epidemic setting. The results from the spatial simulation study have shown that if resistance is present in the data set then these approaches offer various advantages over the conventional survival model - in that they can deal with confounding aspects due to

heterogeneous susceptibility and help to remove bias from the predicted infection times. In general the mixture model is potentially the more useful of the two alternatives, since it allows for both the resistant and susceptible groups to have non-degenerate survival distributions. These frameworks can also be adapted to include covariates in both the mixing parameters and component hazards depending on how they are thought to affect the dynamics of the disease.

We have also shown how measures derived from the basic properties of survival analysis can be applied to the problem of quantifying the risk of infection. The hazard function in particular is a useful tool that can be used as a means of assessing the current risk in a continuously developing epidemic situation. In addition the ability to be able to extend predictions over periods of greater than one day is a particularly useful feature.

The simulations also show that it is important that the mixing is well-defined if the mixture model is to outperform the conventional approach. It can be seen that in situations where the mixing between susceptible and resistant premises can be reasonably quantified, the mixture model gives a greater degree of accuracy when targeting 'at-risk' premises. This is due to the fact that it incorporates individual-level variability, not only in the external epidemic mechanism but also in the inherent propensity of a premise to succumb to infection.

Spatial and temporal hazard maps also provide good ways of visualising areas at risk. In this case, where localised regression was used to smooth the predicted hazards over space, a smaller span will result in different absolute values of the hazard at the predicted points, and so the maps must be interpreted with care. This is particularly necessary when susceptibility is dependent upon premise-specific measures. (Note that VL and AD can be calculated at any point in space and time, so an alternative mechanism here would be to superimpose a fine grid over the study region and predict directly to each grid point using the method described in section 7.1.4. This may not hold for all epidemic situations.)

An important result is that the ability to be able to correctly distinguish between those premises more likely to be susceptible to infection could help to greatly reduce the amount

of culling needed in order to bring an epidemic under control. The potential advantages of a targeted control strategy over a standard CP/DC/ring policy are shown in a series of simulated epidemics, in which the target cull results in far fewer premises needing to be culled in order to bring the epidemic under control. This highlights possible advantages of using localised, rather than global control orders when different dynamics are exhibited across different localised regions.

Although we have shown that the technology has the potential to deal with some key epidemiological problems, there is still an issue regarding the quality of real-world epidemic data that is currently available. When the mixture model was applied to the 2001 Devon FMD data set the results suggested that the AD and species-specific densities did not drive the resistant process, with the mixture model producing results similar to the conventional models fitted in chapter 5. The 2001 Devon FMD data set is one of the most detailed epidemiological data sets currently available, and yet additional knowledge about factors thought to affect the spread of the disease across premises is clearly needed if we wish to study the spatial spread of infectious animal diseases at this scale. The next chapter will comment on all the issues presented in this thesis and discuss possible future work.

Chapter 8

Conclusions and further considerations

In this thesis we have explored the feasibility of applying spatial survival modelling techniques to model inter-premise spread of infectious animal diseases. The Royal Society inquiry into *Infectious diseases in livestock* (Follett et al. 2002) conducted in response to the 2001 UK FMD epidemic resulted in a call for increased funding into this field of research. Clearly the socioeconomic and welfare costs associated with infectious animal diseases can be vast, and ongoing outbreaks such as the H5N1 avian influenza epidemic currently moving across Asia and Europe further highlight the importance of developing new and improved modelling techniques that can be used to help inform effective and efficient strategies to control the spread of contagious animal and epizootic diseases.

In the aftermath of the 2001 UK FMD outbreak there was a lot of skepticism expressed in the veterinary literature about the role that mathematical modelling played in informing British government control policies (BGP). Kitching et al. (2006), Taylor (2003) and Wingfield et al. (2006) all questioned the British government's use of 'unvalidated' mathematical models to justify pre-emptive culling policies. In these articles they express concern about the fact that many of the assumptions used in the development of these

modelling strategies were either misguided or inaccurate - such as having to use out-of-date census counts to measure the size of each premise, the fact that constant transmission rates were assumed across the entire UK area and the absence of accurate contact tracing data. An overriding conclusion from these reports are that models that have not been biologically and epidemiologically validated should not be used to inform control policies.

This has acted as a catalyst for a vital debate on the role of mathematical and statistical modelling in epidemic research, and highlights once again the importance of getting the biology correct from the outset if we are to give any practical weight to inferences derived from these kinds of modelling approaches. The predictive potential of mathematical models can help to provide powerful insights into the dynamics of infectious diseases. Modelling strategies can aid investigations into the effect and extent of various internal and external factors on the spread of epidemics, help to disentangle conflicting information and also focus study on to particular issues. For example, simulation models are particularly useful when investigating the potential advantages and disadvantages of varying control policies, or the effects of different biological assumptions on the eventual course of an epidemic; but they are simply one tool at policy makers' disposal and discretion needs to exercised in the interpretation of results derived from them.

In order to tackle these problems effectively a collaborative effort is required across many different scientific disciplines, from microbiology and pathology to mathematics and economics, since the development of efficacious control strategies is dependent on a sound understanding of disease dynamics at many scales, from the microscopic through to the global (worldwide) level. In addition, it is important to remember the role that human interaction plays in this process, with response and relief strategies greatly aided or hindered by existing economic, cultural and political climates.

A useful mathematical or statistical epidemic model should incorporate information about important biological factors thought to affect the dynamics of the disease in question. We have tried to develop our modelling approach with a view to being able to take the basic principals and apply them to different infectious animal diseases, other than just FMD,

even though this particular scenario serves to drive our study. Our approach incorporates many of the features of previous epidemic models.

The most common form of mathematical model for modelling epidemic data is based around a compartmental framework, i.e. at any time point each individual can be classified into one of a series of distinct categories based on their current disease status. The models can be either deterministic (e.g. Ferguson et al. 2001a,b) or stochastic (e.g. Keeling et al. 2001a) and can incorporate a range of common epidemiological features such as variable and recurrent susceptibility to infection, temporal immunity, latent and exposure periods, carriers and host vectors, heterogeneous mixing of populations and modelling transmission within subsets of a population (e.g. in venereal diseases).

Deterministic models in particular should be treated with caution since although they are often easier to fit than their stochastic counterparts, they do not encompass any random variation in their definition, which is a fundamental concern in epidemic modelling. Stochastic models instead offer a way of being able to incorporate random variation into the model formulation, though the price is usually felt through more complex fitting mechanisms. In the compartmental framework, cellular automata approaches can help to speed this up, but interactions are limited to modelling over discrete lattice structures in space and time and so leads to the potential problem of neighbourhood saturation (see e.g. Mikler et al. 2005).

One advantage of using statistical methodology is that random variation is modelled directly. In addition the effect of covariates on the epidemic process can be directly quantified. In essence we can elicit information about the causes of stochastic variation and the interactions between dependent and independent random variables directly from the data.

In chapter 3 we discussed some relative advantages and disadvantages between these different types of modelling strategies. For infectious disease epidemic modelling at this scale, where we are interested in predicting space-time spread, it is important that a model contains both spatial and temporal structure. This precludes the use of traditional time-

series models (Box and Jenkins 1976, Chatfield 2001) or temporal GLMs (McCullach and Nelder 1989), and similarly purely spatial approaches are also of limited use in these situations (Bailey and Gatrell 1995, Diggle 2003). Spatio-temporal GLMMs (Bernardinelli et al. 1995, Knorr-Held 2000) are potentially much more useful since they are not restricted by standard normality conditions and can incorporate spatial and temporal structure through mixtures of fixed and random effects, or through space-time varying covariates.

These desirable features apply also to survival modelling, except here it is the *time to infection* that is being modelled, rather than counts of individual infections. This allows us to elicit useful information not only about the magnitude of the hazard of infection for each individual over time, but also its shape. From the hazard we can produce spatial and temporal risk maps that can provide information analogous to both the basic reproductive number, R_0, and the epidemic curve (see chapter 3).

These latter measures are commonly used in epidemic modelling to assess the effects of different biological assumptions or control strategies on the course of the epidemic. Since R_0 measures the average number of secondary infections from each primary infection, it is important to reduce this value to less than one in order to prevent epidemic spread. Many approaches, both early (Ferguson et al. 2001a,b, Keeling et al. 2001a) and more recent (Tildesley et al. 2006) have used this approach. In all of these models however the focus is on the effect of response policies on the global epidemic, and an important point noted in Keeling et al. (2001a) was that the size of the neighbourhood for any kind of cull strategy would be situation and disease specific. In particular there was evidence that the disease dynamics changed across different spatial regions in the UK (for example the disease was more aggressive in Cumbria than in Devon for example).

An interesting scenario presents itself when we consider whether a more focussed control strategy could be enforced, based upon using a series of models at smaller spatial scales such as at the county level. Incorporating the differences in disease dynamics exhibited across these smaller regions into localised spatio-temporal hazard maps could be used to influence the choice and extent of control adopted in each area. With this in mind we

developed a series of statistical spatial survival models and applied them to data from the 2001 FMD epidemic in Devon. A key feature of these models was the ability to predict future risk.

From a modelling perspective a further important consideration concerns how to integrate information about factors affecting disease dynamics at different biological scales into the model. Here we are looking at modelling at a large scale (e.g. premise level spread), and in order to do this we have to amalgamate together information from smaller biological scales. An introduction into the sorts of problems facing epidemiologists in this context is given in chapters 1 and 2.

There are many factors, both internal and external to an individual host organism that will determine the extent and magnitude of disease spread. At one end of the scale the within-host spread of a disease is largely dependent its pathogenicity, and this can vary greatly even between competing strains of the same basic pathogen. At the other end of the scale factors such as climate conditions, geographical complexity and varying biosecurity can affect the transmission potential of a disease between hosts and the ability for a pathogen to survive in transit. Many previous methods have attempted to average out factors involved in disease spread, and by modelling at the global scale allows individual-level effects to be soaked up. Individual-level predictive epidemic models are much harder to implement since the effect of individual level heterogeneities can be high, however their use as a means of targeting control policies is potentially much greater. How effective these predictive models are is largely dependent on how well we can identify and incorporate factors associated with the epidemic process. The use of frailty effects can help to account for extra variation from unknown confounding factors, but for predictive purposes this is of limited use if the factors that we do know about do not constitute a good fit to the data. This is further compounded by unreliable data (see chapter 5).

Previous FMD models have assumed species-varying transmission rates, with later papers also including species-varying susceptibility. However these all assumed constant transmission rates over time (though the very recent paper by Savill et al. 2007 attempts to

address this particular problem). In our model we deal with this in a different manner. Firstly we incorporate temporal information on the within-herd spread of the disease into the model through the use of scaled infectivity functions. These were developed from some species-specific deterministic SEIR models (for the viral excretion of an infected herd over time) developed by colleagues working in the VLA, Weybridge (Arnold 2005). Once scaled by the actual number of animals of the corresponding species on an IP, they gave a measure of total viral excretion for a premise over time. By smoothing the viral excretion over space, a measure of the amount of viral load per unit area can be obtained. If the viral load is included in the hazard function, then this results in space-time varying risk (i.e. transmission potential) based on species-specific differences in viral excretion. This is a particularly desirable aspect to our approach, since not only is it adaptive (i.e. it changes when new information arises), but it also allows information on the epidemic process to drive the transmission potential.

Currently, as discussed in chapter 5, there are many factors relating to external conditions (e.g. climate, wind speed and direction, geographical complexities) that we do not have information about. There is the potential that these effects could be built into the model, either directly as covariates in the hazard, or through the form of smoothing function used, though presently an isotropic process for viral spread is assumed and a bivariate normal kernel function used as a spatial smoothing function. A possible way to incorporate anisotropy could be through the correlation matrix between locations, and the use of different bandwidths to result in different levels of spatial smoothing.

The choice of bandwidth is of key importance in spatial smoothing techniques (Bailey and Gatrell 1995). A series of conventional models fitted to a simulated spatial epidemic with no resistance was used to investigate the effects of using different bandwidth and threshold values on the predictions from the model. The results suggested that the choice of threshold was less important than the choice of bandwidth, with smaller bandwidths seeming to provide better predictions. This even held in a situation where a smaller bandwidth was used to generate VL than was used in the simulation. There is a trade-off with this approach in that the bandwidth had to be such that the number of infected

premises in the model fit did not become too small. In this case the predictions began to break down. With regard to the parameters used in the simulation, the best predictive model was one with a smaller bandwidth but with the threshold adjusted such that the number of IPs used in the model fit was similar to the number from censoring with the exact values. This is reasonable in the simulated case since a smaller bandwidth will reduce the amount of censored observations in the model whilst preserving the amount of IPs. In the Devon data set infections are more sparse, and a larger bandwidth is required in order to obtain a reasonable number of IPs. This suggests that the Devon data set may not follow a true localised epidemic process (this will be discussed in more detail below).

The viral load incorporates various desirable epidemiological features. It is biologically motivated, including information on the epidemic process in its definition, it varies over space and time, can be sequentially updated and also builds in species-specific within-herd spread of the virus based on the relative size and proximity of nearby IPs. It also incorporates the need (Keeling et al. 2001a) to include information about the infectiousness of individual premises and how this changes over time, though there is still the additional issue of how this relates to survival time. For example, it is reasonable to expect that two premises with identical characteristics and an equal VL covariate will have the same hazard of infection, although if they are subjected to this viral pressure at different times then their relative hazards will not equal one. This can be partly dealt with by conditioning out the dependence on the past covariate history from the hazard (see chapter 5). However the temporal component (i.e. baseline hazard) is still linked to the absolute time from the beginning of the epidemic. In order to make hazard measurements comparable it is more useful to consider using survival time from *exposure* to the virus rather than absolute time, and since the viral load is a measure of virus per unit area a threshold can be applied, whereby an individual is classified as exposed or unexposed corresponding to the VL acting on that location at that point in time. This way of using the viral load to censor the data set via exposure is a useful way of thinking about epidemic modelling, since it helps to reflect the idea that the risk set of susceptible individuals changes over space as well as time. Moreover, this still retains past temporal structure through the survivor and density functions.

Furthermore, if average cumulative viral load (AV) is used as a covariate in the model, then this also incorporates a 'lag' effect for the decrease in risk due to culling of IPs and removal of infected animals. A downside of the current model is that it does not account for relative changes in transmission over time. That is it does not allow for the distributional form of the hazard for two premises with the same VL to be different if they were exposed at different *absolute* time points. This could potentially arise due to external factors such as movement and trading restrictions being enforced. The effect of the latter is that non-localised infection should be reduced, and this will be felt through the calculation of the VL. Movement restrictions on or around premises are more likely to affect localised transmission potential and a change-point style approach could possibly be used to account for this. (The melding of relative and absolute time is also a potential issue with the mixture model approach - see chapters 6 and 7 - if the survival distribution for one of the groups is believed to be based on a different time scale.)

Some early model results (chapter 5) showed that AV did not capture the dynamics of the disease in Devon, resulting in hugely overpredicted survival times. Fitting to the data set censored via exposure did little to improve matters. Two potential reasons for this were identified: the presence of possible 'spark' or non-localised infections and/or the possibility of premise-varying susceptibility to infection. Incorporating species-specific (SD, CD and PD) and non-specific uninfected animal (AD) densities as surrogates for susceptibility in the hazard also failed to solve the problem of overprediction (though in the latter case the model suggested that the parameter for AD was significantly different to zero).

Considering these two problems in turn, it seemed that resistance to infection was more likely to be the cause of the overprediction. The presence of non-localised infection in the data set was thought to be low, particularly since non-exposed premises are removed from the data set due to censoring via exposure. An advantage of the VL and animal density measures are that information from non-exposed premises can still be included in the model. In addition, removal of unexposed premises and the movement restrictions that were imposed early on in the epidemic means that the risk of any remaining non-localised infections heavily inflating the survival times is thought to be small. Some

simulation studies (chapter 6) indicated that the effect of small numbers of susceptible spark infections on the accuracy of the parameter estimates obtained from a reasonable model is small.

The same set of simulated experiments suggested that the effect on the parameter estimates from including resistant premises was much greater, and some more complex spatial simulation studies conducted in chapter 7 further reinforced these findings. Resistant premises effectively act as outlying and influential observations, since they become noticeable as premises with large viral loads but long survival times. In order to deal with this problem two alternative model formulations were considered, the long-term survivor (Maller and Zhou 1996) and mixture models (McLachlan and Peel 2000). The former splits the data into susceptible and immune proportions, and allows only the susceptible proportion to experience failure. The mixture model is potentially more useful in the FMD case, since it allows the resistant group to experience infection. The important difference between incorporating susceptibility through the hazard function and through a long-term survivor or mixture model is that the latter models allow the *shape* of the survival distribution to be different, rather than just the magnitude.

The spatial simulations conducted in chapter 7 showed that the mixture model and long-term survivor models performed much better than the conventional approach when resistance to infection was present in the data set. Furthermore they both seemed to replicate the results of the conventional model when resistance wasn't present.

The predictive power of the mixture and conventional models was compared in chapter 7.2. The survival approach allows us to obtain individual level predictions of future failure times, and from these various measurements of risk can be obtained. Due to the unconventional nature of our viral load covariate, the predictions had to be obtained by using the posterior samples from our model fit to simulate the future path of the epidemic, since VL is dependent upon the ongoing epidemic process. In simulation studies the mixture approach performed much better than the conventional approach, and did not exhibit the same extent of overprediction. The accuracy of the individual level predictions were

still not great however, and we must take great care in attributing too much significance to individual predicted failure times (Henderson and Keiding 2005, Parkes 1972). Instead we investigated whether some type of spatial or temporal risk map, based on the hazard of infection in some subsequent time period, or the probability of being a top-ranked future infection would be more useful. One issue with the former approach is that the actual hazard is based on the predicted survival time, and so if we are concerned about the accuracy of the individual level predictions then this carries over to the value of the hazard as well. It will give some kind of 'relative' risk in comparison to other premises (though this is not technically a relative risk since it is not controlled by some background control measure, though in some sense the AD and species-specific densities incorporate aspects of the background population since they measure the *intensity* of susceptible animals per unit area).

If the hazards are ranked then the probability of being in the top, r say, future infections may be a more robust measure of risk than the hazard, and the level of ranking of interest can be adapted as seen fit by the analyst. Spatial maps can help to identify areas of potential future risk, whilst individual level predictions can help to distinguish between susceptible and resistant premises within the same area. Obviously a key issue in any of these approaches is ensuring that the mixing is well-defined, and the accuracy of the predictions depends greatly on how well the covariates capture this process. We have shown that in the worst case scenario the mixture model will at least still replicate the conventional results.

We then applied our mixture model approach to a simulated epidemic situation to test whether a targeted control policy could perform better than a more standard contiguous cull strategy. The 'contiguous' cull strategy employed in the simulation was a mixture of contiguity and a ring cull since we did not have information about the actual contiguity matrix used by DEFRA. Instead we applied a high probability of two premises being classed as contiguous if they were within 1km of each other, and a small probability if they were within 2km of each other. Though this perhaps related to more premises being classed as contiguous than in reality, it was certainly less encompassing than a true 3km

ring cull. (Note that although a 3km ring cull was not introduced in Devon in 2001, the dynamics of the simulation were more aggressive than those actually observed, and the extent of control required was reflected by this fact.)

The simulation assumed that all IPs were 'culled' (i.e. removed from the study) after 24hrs and all contiguous UIPs within 48hrs of an infection. In addition there was no residue effect left over from infected premises after culling. The results from the simulation showed that an IP-only cull policy failed to keep the epidemic under control. The contiguous cull worked much better but resulted in large numbers of removals. The target cull policy resulted in similar numbers of IPs as the contiguous cull, but with substantially less premises needing to be culled. This highlights the potential advantages of focussing response strategies at a smaller spatial scale and adapting the aggression of each strategy to the individual-level dynamics. However as mentioned earlier it is vital that the model and covariates provide a good fit to the data.

The problem of defining the mixing was evident when we applied the mixture model to the real data set. It is clear that the VL and uninfected animal densities do not capture the dynamics of the disease in Devon. There are various potential reasons for this. As discussed in chapters 2 and 5, there are many possible confounding factors that we are either unaware of, or do not have information about. For example, we know there are variable incubation periods, difficulties with detecting clinical signs and diagnosing infections, animal movements between and around premises, variable biosecurity and animal husbandry conditions, differences in farming practices, and variable landscape fragmentation and livestock densities.

The approaches that we have discussed in this thesis have the capacity to incorporate information about many of these factors, if available. The survival framework could be adapted to introduce left- and interval-censoring to help account for variable incubation periods for example. Latent infections relating to culled UIPs could potentially be imputed (Deardon et al. 2006) and anisotropy could be introduced in the spatial smoothing function to reflect landscape fragmentation and environmental conditions such as wind

speed and direction. As with any real-life situation, obtaining information about these factors is difficult and so instead we must try to produce as robust a model as possible in the absence of this information. Frailty effects are useful in this respect since they can absorb excess variation from some of these confounding factors, however they are of little practical use if the model does not accurately capture the dynamics of the disease, particularly when using the model to predict. It is in these situations where stochastic simulation based models such as those described in chapter 3 are useful.

Also, we have based our infectivity functions on a deterministic model for within-herd spread. Since we could not produce individual premise level curves for viral excretion we had to make a series of assumptions about how this behaves relative to the size of each premise. Also we have to make assumptions regarding the spatial spread and density of animals around each point (premise) location. This affects the timing and smoothing of excretion rates, as does the additional assumption that the infection spreads to all animals in a herd.

Another possible reason why we are not capturing the dynamics of the disease is that our model may be wrongly defined. We have assumed a parametric form for our survival distributions, and this assumption may be too strong. In addition the proportional hazards model may be inappropriate. An interesting extension of the Keeling et al. (2001a) model, published by Diggle (2005), developed a partial likelihood approach for modelling the probability of infection. The advantage of this type of method was that parameter estimates can be obtained for the effect of covariates on the epidemic process, without the additional (and sometimes restrictive) assumption that the data follow a particular parametric form. A problem with this type of semi-parametric approach (see Cox 1972, Cox and Oakes 1984) is that the lack of parametric dependence means that prediction becomes extremely difficult. There have been recent examples (e.g. Demeris and Sharples 2006) of papers in which survival estimates have been extrapolated from semi-parametric models, and it may be interesting to look at these sorts of approaches in the context of epidemic modelling as a potential way of tackling this issue.

Of course the form of the mixture model is also important. We have assumed two groups and have forced a parametric form to each component in the mixture. This could introduce additional biases not only due to the choice of parametric distribution (as highlighted above), but also due to the number of groups in the model. This latter issue could potentially be dealt with using reversible jump methodology (Richardson and Green 1997).

The final alternative is that the Devon epidemic simply did not follow a localised infective process. This is a possibility, since it can be seen from figure 5.1 that there are areas where there are large numbers of uninfected premises in-between nearby infected premises. Movement restrictions were in place fairly early on in the Devon epidemic and landscape fragmentation was less extensive in Devon than in some other infected regions, and this perhaps indicates that the sources of infection may be due to infected animals being brought into farm premises rather than through a localised mechanism. It would be interesting to try our techniques on a data set from an area such as Cumbria, which exhibited a more traditional wave-like epidemic spread pattern synonymous with localised epidemic processes.

With regards to non-localised infections, the contact-tracing data available for the movement of animals between different farm premises provided by the Cattle Traceability System (CTS) has information on cattle movements into and out of premises. Although this does not help gauge movement of other livestock such as pigs and sheep, it may perhaps still give a better indication of the movement of animals around the UK in general, and provide useful links between different areas of the country, helping to better identify dangerous contacts.

The models that we have explored in this thesis provide some potentially useful frameworks in order to view and model important biological and epidemiological aspects of infectious animal disease epidemics. In situations where the movement and location of susceptible animal herds is known and can be controlled, the viral load approach to censoring the data set via exposure is useful, since not only does it exclude premises that are 'not-exposed' to the virus, but it also allows survival times to be compared relative to exposure.

Extensions to the basic survival models can be used to adapt to different situations, and in particular the mixture and long-term survivor models have shown great potential in simulation studies to deal with the issue of resistance or immunity to infection. Furthermore the mixture or competing risks approaches can be used to model multiple causes of failure.

Accounting for additional aspects of the epidemic process is difficult however, especially since reliable data is not always available due to the logistical constraints of large scale data collection. A further issue is that of sequential updating of the model as an epidemic progresses. A key feature of the VL is the ability to be able to incorporate new information about the epidemic process, and the hazard maps shown in chapter 7 were produced by fitting a new model at each time point. An interesting future project could be to see whether models could be updated using prior information from models fitted at previous time points. One problem with this is that the accuracy of each new model is dependent on the accuracy of the models at previous time points, which could result in serious forecasting error. Methodology such as particle filters (Doucet et al. 2001) may potentially be of interest here.

Appendix A

Fitting non-standard likelihoods in WinBUGS

None of the models used throughout this thesis fall into the list of standard probability distributions in WinBUGS. Therefore an alternative fitting mechanism is needed. Fortunately we can use the so-called 'zeros' trick (Spiegelhalter et al. 2003).

Consider that our data, $T_1, \ldots, T_n$, is a random sample from a non-standard distribution, where each observation contributes L_i to the likelihood. If we introduce a set of n zero-valued Poisson latent observations, Q_i, with mean θ_i, then Q_i contributes $e^{-\theta_i}$ to the likelihood. The correct likelihood contribution for T_i can then be obtained by setting $\theta_i = -\log(L_i)$. (Note that a similar approach can be used if we introduce a set of n Bernoulli random variables instead, each with value one and parameter p_i, such that $p_i = \frac{L_i}{C}$, where C is a constant that ensures $p_i < 1$.)

This approach is used for all of the models in this thesis, but as an illustrative example consider the discrete time Weibull model with time-dependent covariate given in section 5.4,

with likelihood:

$$
\begin{aligned}
L(\alpha, \beta_0, \beta_1) \;=\; & \prod_{i=1}^{n} \Bigg(\Bigg\{ \left[1 - \exp(-\lambda_{i(t_i-1)}[t_i^{\alpha} - (t_i - 1)^{\alpha}])\right] \\
& \times \exp\left(- \sum_{j=1}^{t_i - 1} \lambda_{i(j-1)}[j^{\alpha} - (j-1)^{\alpha}] \right) \Bigg\}^{\delta_i} \\
& \times \left[\exp\left(- \sum_{j=1}^{t_i} \lambda_{i(j-1)}[(j)^{\alpha} - (j-1)^{\alpha}] \right) \right]^{1-\delta_i} \Bigg), \qquad \text{(A.1)}
\end{aligned}
$$

where $\lambda_{it} = \exp(\beta_0 + \beta_1 X_{it})$ and X_{it} is a time-dependent covariate. The WinBUGS code for (A.1) is then:

```
model
{
      C<-100

      for(i in 1:N)
      {
           for(j in 1:t[i])
           {
                lambda[i,j]<-exp(beta0+beta1*x[i,j])
                h[i,j]<-lambda[i,j]*(pow(j,alpha)-pow(j-1,alpha))
           }
           for(j in (t[i]+1):maxt)
           {
                lambda[i,j]<-0
                h[i,j]<-0
           }
           S[i]<-exp(-sum(h[i,]))
           f[i]<-(1-exp(-h[i,t[i]]))*exp(h[i,t[i]]-sum(h[i,]))
           zeros[i]<-0
           theta[i]<-(-1)*(delta[i]*log(f[i])+(1-delta[i])*log(S[i]))+C
           zeros[i]~dpois(theta[i])
      }

      beta0~dnorm(0,0.01)
      beta1~dnorm(0,0.1)
      alpha~dgamma(0.1,0.1)
}
```

Here C is simply a constant to ensure that $\theta_i > 0$ (since it is a Poisson mean). The matrix

of covariate values, $\boldsymbol{X}$, that was passed to WinBUGS from R was formulated such that column 1 of $\boldsymbol{X}$ corresponded to covariate values at time 0 and so on. Hence although the WinBUGS code states `lambda[i,j]<-exp(beta0+beta1*x[i,j])` this actually corresponds to $\lambda_{i(j-1)}$ in real terms.

An alternative and perhaps more attractive way to do this for large-scale models is to use the WinBUGS Development Interface, known as WBDev (Lunn 2005). This allows non-standard distributions and/or complex logical expressions to be hard-wired directly into the WinBUGS package by coding them in component Pascal. Moreover the hard work has already been done and direct knowledge of component Pascal is not necessary, since example code is provided that can be adapted to allow the user to insert their own functions at pre-defined places within the code. This allows complex model formulations to run much more efficiently.

Appendix B

Initial value generation

The models were reasonably sensitive to the choice of initial value passed to WinBUGS. To illustrate the method of generation, let T be the response variable with hazard function, $h(t)$ of a similar form to the models described in chapter 5. Suppose also that we have an additional m-vector of covariates, $\boldsymbol{x} = (x_1, \ldots, x_m)$. To generate initial values, $\boldsymbol{\Psi}_{\text{ini}} = (\alpha_{\text{ini}}, \boldsymbol{\beta}_{\text{ini}})$, for each chain consider the following steps:

1. Sample a value of α_{ini} from a positive distribution.
2. Produce a scatter plot of the data and obtain a rough estimate of the range of infection times for individuals with $\boldsymbol{x} \approx \mathbf{0}$. Denote this range (t_{L0}, t_{U0}).
3. Calculate $E(t_{L0})$ and $E(t_{U0})$ in terms of parameter β_0, providing upper and lower limits (denoted b_{L0} and b_{U0} respectively).
4. Sample $\beta_{0\text{ini}}$ from a $U(b_{L0}, b_{U0})$ distribution.
5. Set $i = 1$.
6. Let $\boldsymbol{x}_{i-} = \mathbf{0}$, where $\boldsymbol{x}_{i-}$ is the vector $\boldsymbol{x}$ with x_i variable removed.
7. Produce a scatter plot of the data and obtain a rough estimate of the range of infection times for individuals with $x_i \approx \bar{x}_i \pm c, c > 0$. Denote this range (t_{Li}, t_{Ui}).

8. Calculate $E(t_{Li})$ and $E(t_{Ui})$ (using $\beta_{0\text{ini}}$ and $\bar{x}_i \pm c$) in terms of parameter β_i, providing upper and lower limits (denoted b_{Li} and b_{Ui} respectively).

9. Sample $\beta_{i\text{ini}}$ from a $U(b_{Li}, b_{Ui})$ distribution.

10. Set $i = i + 1$ and repeat steps 6 to 9 until $i > m$.

This was the basic method used to generate the initial values for the regression parameters in each of the models fitted throughout this thesis. A similar method was used to generate the mixing parameters, $\boldsymbol{\gamma}$.

Bibliography

Aalen, O. (1988), "Heterogeneity in survival analysis," *Statistics in Medicine*, 47, 1121–1137.

Acha, P. N. and Szyfres, B. (2003), *Zoonoses and communicable diseases common to man and animals*, vol. I-III, Pan American Health Organisation, 3rd ed.

Ahmed, E. and Agiza, H. (1998), "On modeling epidemics. Including latency, incubation and variable susceptibility," *Physica A*, 253, 347–352.

Aitken, M. and Clayton, D. (1980), "The fitting of exponential, Weibull and extreme value distributions to complex survival data using GLIM," *Applied Statistics*, 29, 156–163.

Alexandersen, S., Quan, M., Murphy, C., Knight, J., and Zhang, Z. (2003a), "Studies of quantitative parameters of virus excretion and transmission in pigs and cattle experimentally infected with foot-and-mouth disease virus," *Journal Of Comparative Pathology*, 129, 268–282.

Alexandersen, S., Zhang, Z., Donaldson, A., and Garland, A. (2003b), "The pathogenesis and diagnosis of foot-and-mouth disease," *Journal of Comparative Pathology*, 129, 1–36.

Anderson, R. and May, R. (1991), *Infectious diseases of humans*, Oxford University Press.

Anderson, T. (1971), *The Statistical Analysis of Time Series*, Wiley.

Arnold, M. (2005), "A spatial analysis of the 2001 foot-and-mouth disease epidemic," .

Athreya, K. and Ney, P. (1972), *Branching Processes*, Berlin: Springer Verlag.

Ayele, W., Neill, S., Zinsstag, J., Weiss, M., and Pavlik, I. (2004), "Bovine tuberculosis: an old disease but a new threat to Africa," *The International Journal of Tuberculosis and Lung Disease*, 8, 924–937.

Bailey, N. T. (1975), *The mathematical theory of infectious diseases*, Charles Griffin and Company Ltd., London and High Wycombe, 2nd ed.

Bailey, T. C. and Gatrell, A. C. (1995), *Interactive Spatial Data Analysis*, Harlow: Longman Scientific & Technical.

Banerjee, S. and Carlin, B. P. (2004), "Parametric spatial cure rate models for interval-censored time-to-relapse data," *Biometrics*, 60, 268–275.

Basu, A. (1983), *Identifiability*, Wiley Interscience, no. 4, pp. 2–6.

Berke, O. (2005), "Exploratory spatial relative risk mapping," *Preventive Veterinary Medicine*, 71, 173–182.

Berkson, J. and Gage, R. P. (1952), "Survival curve for cancer patients following treatment," *Journal of the American Statistical Association*, 47, 501–515.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., and Songhini, M. (1995), "Bayesian analysis of space-time variation in disease risk," *Statistics in Medicine*, 14, 2433–2443.

Bernoulli, D. (1760), "Essai d'une nouvelle analyse de la mortalité causée par la petite vérole et des avantages de l'inoculation pour la prévenir," *Mémoires de l'Académie Royale des Sciences - Histoire Année 1760, Paris*, 1–45.

Besag, J. and Kooperberg, C. (1995), "On conditional and intrinsic autoregressions," *Biometrika*, 82, 733–746.

Besag, J., York, J., and Mollié, A. (1991), "Bayesian image restoration with two applications in spatial statistics (with discussion)," *Annals of the Institute of Statistical Mathematics*, 43, 1–59.

Boag, J. (1949), "Maximum likelihood estimates of the proportion of patients cured by cancer therapy," *Journal of the Royal Statistical Society. Series B (Methodological)*, 11, 15–53.

Box, G. E. and Jenkins, G. (1976), *Time Series Analysis: Forecasting and Control*, Holden-Day Inc., San Francisco, revised ed.

Breslow, N. (1974), "Covariance analysis of censored survival data," *Biometrics*, 30, 89–99.

Breslow, N. and Clayton, D. (1993), "Approximate inference in generalized linear mixed models," *Journal of the American Statistical Association*, 88, 9–25.

Carlin, B. and Banerjee, S. (2003), *Hierarchical multivariate CAR models for spatiotemporally correlated survival data (with discussion)*, Oxford: Oxford University Press, pp. 45–63.

Celeux, G. (1997), "Discussion on 'On Bayesian analysis of mixtures with unknown number of components' (by S. Richardson and P.J. Green)," *Journal of the Royal Statistical Society. Series B (Methodological)*, 59, 775–776.

— (1998), "Bayesian inference for mixtures: the label switching problem," in *COMPSTAT 98*, eds. Payne, R. and Green, P., Heidelberg:Physica, pp. 227–232.

Celeux, G., Chauveau, D., and Diebolt, J. (1996), "Stochastic versions of the EM algorithm: an experimental study in the mixture case," *Journal of Statistical Computation and Simulation*, 55, 287–314.

Celeux, G., Hurn, M., and Robert, C. (2000), "Computational and inferential difficulties with mixture posterior distributions," *Journal of the American Statistical Association*, 95, 957.

Chatfield, C. (2001), *Time Series Forecasting*, Chapman and Hall/CRC.

— (2004), *The Analysis of Time Series: An Introduction*, Chapman and Hall/CRC, 6th ed.

Chen, M.-H., Ibrahim, J. G., and Sinha, D. (1999), "A new Bayesian model for survival data with a surviving fraction," *Journal of the American Statistical Association*, 94, 909–919.

Chowell, G., Rivas, A., Hengartner, N., Hyman, J., and castillo Chavez, C. (2006), "The role of spatial mixing in the spread of foot-and-mouth disease," *Preventive Veterinary Medicine*, 73, 297–314.

Chung, Y., Dey, D. K., Kim, M., and Kim, C. (2005), "Bayesian model choice in exponential survival models," *Communications in Statistics - Theory and Methods*, 34, 2311–2330.

Clayton, D. (1978), "A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence," *Biometrika*, 65, 141–151.

Collett, D. (2003), *Modelling Survival Data in Medical Research*, Chapman and Hall, London, 2nd ed.

Commenges, D. (1999), "Multi-state models in epidemiology," *Lifetime Data Analysis*, 5, 315–327.

Congdon, P. (2001), *Bayesian Statistical Modelling*, Wiley.

— (2003), *Applied Bayesian Modelling*, Wiley.

Cox, D. (1972), "Regression models and life tables (with discussion)," *Journal of the Royal Statistical Society. Series B (Methodological)*, 34, 187–220.

Cox, D., Donnelly, C. A., Bourne, F. J., Gettinby, G., McInerney, J. P., Morrison, W. I., and Woodroffe, R. (2005), "Simple model for tuberculosis in cattle and badgers," *Proceedings of the National Academy of Sciences of the United States of America*, 102, 17588–17593.

Cox, D. and Oakes, D. (1984), *Analysis of Survival Data*, Chapman and Hall, London.

Cressie, N. (1993), *Statistics for Spatial Data*, Wiley.

Crowder, M. (2001), *Classical Competing Risks*, Chapman and Hall, CRC.

Deardon, R., Brooks, S. P., Grenfell, B. T., Keeling, M. J., Tildesley, M. J., Savill, N. J., Shaw, D. J., and Woolhouse, M. E. (2006), "Inference for individual-level models of infectious diseases in large populations," Submitted.

DEFRA (2001), "Comparisons with 1967," website.

— (2002a), "Foot-and-mouth disease 2001: Lessons to be learned inquiry - Chairman, Dr Iain Anderson CBE," website.

— (2002b), "Origin of the UK Foot and Mouth disease epidemic in 2001," website.

— (2004), "Animal Health and Welfare: FMD Data Archive - Introduction," website.

Demeris, N. and Sharples, L. (2006), "Bayesian evidence synthesis to extrapolate survival estimates in cost-effectiveness studies," *Statistics in Medicine*, 25, 1960–1975.

Dempster, A., Laird, N., and Rubin, D. (1977), "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *Journal of the Royal Statistical Society. Series B (Methodological)*, 39, 1–38.

Diebolt, J. and Robert, C. P. (1994), "Estimation of finite mixture distributions through Bayesian sampling," *Journal of the Royal Statistical Society. Series B (Methodological)*, 56, 363–375.

Diggle, P., Chetwynd, A., Häggkvist, R., and Morris, S. (1995), "Second order analysis of space-time clustering," *Statistical Methods in Medical Research*, 4, 124–136.

Diggle, P., Zheng, P., and Durr, P. (2005), "Nonparametric estimation of spatial segregation in a multivariate point process: bovine tuberculosis in Cornwall, UK," *Applied Statistics*, 54, 645–658.

Diggle, P. J. (1990), *Time Series: A Biostatistical Introduction*, Oxford Science Publications.

— (2003), *Statistical Analysis of Spatial Point Patterns*, Arnold, 2nd ed.

— (2005), "A partial likelihood for spatio-temporal point processes," Tech. rep., Johns Hopkins University, Department of Biostatistics Working Papers. Working Paper 75.

Donaldson, A. (1983), "Quantitative data on airborne foot-and-mouth-disease virus - its production, carriage and deposition," *Philosophical Transactions Of The Royal Society Of London Series B - Biological Sciences*, 302, 529–534.

— (1987), "Foot-and-mouth disease: the principal features," *Irish Veterinary Journal*, 41, 325–327.

Donaldson, A. and Alexandersen, S. (2001), "The relative resistance of pigs to infection by natural aerosols of foot-and-mouth disease virus," *Veterinary Record*, 148, 600–602.

Donaldson, A., Alexandersen, S., Sorensen, J., and Mikkelsen, T. (2001), "Relative risks of the uncontrollable (airborne) spread of FMD by different species," *Veterinary Record*, 148, 602–604.

Donaldson, A., Gibson, C., Oliver, R., Hamblin, C., and Kitching, R. (1987), "Infection of cattle by airborne foot-and-mouth disease virus: minimal doses with O1 and SAT2 strains," *Research in Veterinary Science*, 43, 339–346.

Doran, R. J. and Laffan, S. W. (2005), "Simulating the spatial dynamics of foot and mouth disease outbreaks in feral pigs and livestock in Queensland, Australia, using a susceptible-infected-recovered cellular automata model," *Preventive Veterinary Medicine*, 70, 113–152.

Doucet, A., Freitas, N. D., and Gordon, N. (eds.) (2001), *Sequential Monte Carlo methods in practice*, Springer.

Ducrot, C., Abrial, D., Calavas, D., and Carpenter, T. (2005), "A spatio-temporal analysis of BSE cases born before and after the reinforced feed ban in France," *Veterinary Research*, 36, 839–853.

Durand, B. and Mahul, O. (2000), "An extended state-transition model for foot-and-mouth disease epidemics in France," *Preventive Veterinary Medicine*, 47, 121–139.

Durr, P. and Froggatt, A. (2002), "How best to geo-reference farms? A case study from Cornwall, England," *Preventive Veterinary Medicine*, 56, 51–62.

Durr, P., Tait, N., and Lawson, A. (2005), "Bayesian hierarchical modelling to enhance the epidemiological value of abattoir surveys for bovine fasciolosis," *Preventive Veterinary Medicine*, 71, 157–172.

Ebrahimi, N., Gelfand, A. E., Ghosh, M., and Ghosh, S. K. (1997), "Bayesian analysis of change point hazard rate models," Tech. Rep. 9708, University of Connecticut.

Efron, B. (1977), "The efficiency of Cox's likelihood function for censored data," *Journal of the American Statistical Association*, 72, 557–565.

Escobar, M. D. and West, M. (1995), "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Farewell, V. (1977), "A model for a binary variable with time censored observations," *Biometrika*, 64, 43–46.

Ferguson, N. M., Donnelly, C. A., and Anderson, R. M. (2001a), "The foot-and-mouth epidemic in Great Britain: pattern of spread and impact of interventions," *Science*, 292, 1155–1160, published online 12 April 2001.

— (2001b), "Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain," *Nature*, 413, 542–548.

Follett, B., Allen, P., Bateson, P., Black, D., Brown, F., Eddy, R., Leather, S., Levin, S., Linklater, K., Longfield, J., McConnell, I., McLean, A., McMichael, A., Mumford, J., Weiss, R., and Westergaard, J. (2002), "Inquiry into Infectious Diseases in Livestock," Tech. rep., Royal Society.

Fuentes, M. and Kuperman, M. (1999), "Cellular automata and epidemiological models with spatial dependence," *Physica A*, 267, 471–486.

Fuks, H. and Lawniczak, A. (2001), "Individual-based lattice model for spatial spread of epidemics," *Discrete Dynamics in Nature and Society*, 6, 191–200.

Garner, M. and Lack, M. (1995), "An evaluation of alternate control strategies for foot-and-mouth disease in Australia: a regional approach," *Preventive Veterinary Medicine*, 23, 9–32.

Gelfand, A. and Smith, A. (1990), "Sampling based approaches to calculating marginal densities," *Journal of the American Statistical Association*, 85, 398–409.

Gelfand, A. and Vounatsou, P. (2003), "Proper multivariate conditional autoregressive models for spatial data analysis," *Biostatistics*, 4, 11–25.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis*, Chapman and Hall/CRC, 2nd ed.

Geman, S. and Geman, D. (1984), "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Gerbier, G., Bacro, J., Pouillot, R., Durand, B., Moutou, F., and Chadœuf, J. (2002), "A point pattern model of the spread of foot-and-mouth disease," *Preventive Veterinary Medicine*, 56, 33–49.

Gilks, W., Richardson, S., and Spiegelhalter, D. (eds.) (1996), *Markov Chain Monte Carlo In Practice*, Chapman and Hall, London.

Gilks, W. and Wild, P. (1992), "Adaptive-rejection sampling for Gibbs sampling," *Applied Statistics*, 41, 337–348.

Gloster, J., Blackall, R., Sellers, R., and Donaldson, A. (1981), "Forecasting the airborne spread of foot-and-mouth disease," *Veterinary Record*, 108, 370–374.

Gloster, J., Freshwater, A., Sellers, R., and Alexandersen, S. (2005), "Re-assessing the likelihood of airborne spread of foot-and-mouth disease at the start of the 1967-1968 UK foot-and-mouth disease epidemic," *Epidemiology and Infection*, 133, 767–783.

Haganaars, T., Donnelly, C., and Ferguson, N. (2006), "Epidemiological analysis of data for scrapie in Great Britain," *Epidemiology and Infection*, 134, 359–367.

Harris, T. (1963), *The Theory of Branching Processes*, Berlin: Springer-Verlag.

Hastings, W. (1970), "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, 57, 97–109.

Henderson, R. (1969), "The outbreak of foot-and-mouth disease in Worcestershire. An epidemiological study: with special reference to spread of the disease by wind-carriage of the virus," *Journal of Hygiene, Cambridge*, 67, 21–33.

Henderson, R. and Keiding, N. (2005), "Individual survival time prediction using statistical models," *Journal of Medical Ethics*, 31, 703–706.

Henderson, R., Shimakura, S., and Gorst, D. (2002), "Modeling spatial variation in leukemia survival data," *Journal of the American Statistical Association*, 97, 965–972.

Hougaard, P. (1986a), "A class of multivariate failure time distributions," *Biometrika*, 73, 671–678.

— (1986b), "Survival models for heterogeneous populations derived from stable distributions," *Biometrika*, 73, 387–396.

— (1999), "Multi-state models: A review," *Lifetime Data Analysis*, 5, 239–264.

— (2000), *Analysis of Multivariate Survival Data*, Springer-Verlag, New York.

Hugh-Jones, M. (1972), "Epidemiological studies on the 1967-1968 foot-and-mouth disease epidemic: attack rates and cattle density," *Research in Veterinary Science*, 13, 411–417.

Hugh-Jones, M. and Wright, P. (1970), "Studies on the 1967-1968 foot-and-mouth disease epidemic: the relation of weather to the spread of the disease," *Journal of Hygiene, Cambridge*, 68, 253–271.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001), *Bayesian Survival Analysis*, Springer.

Kalbfleisch, J. D. and Prentice, R. L. (1973), "Marginal likelihoods based on Cox's regression and life model," *Biometrika*, 60, 267–278.

— (2002), *The Statistical Analysis of Failure Time Data*, Wiley, 2nd ed.

Kao, R. (2001), "Landscape fragmentation and foot-and-mouth disease transmission," *Veterinary Record*, 148, 746–747.

Kaplan, E. and Meier, P. (1958), "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, 53, 457–481.

Keeling, M. J., Tildesley, M. J., Savill, N. J., Woolhouse, M. E., Shaw, D. J., Deardon, R., Brooks, S. P., and Grenfell, B. T. (2006), "FMD control strategies - Comment," *Veterinary Record*, 158, 707–708.

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. (2001a), "Dynamics of the 2001 UK foot and mouth epidemic: Stochastic dispersal in a heterogeneous landscape," *Science*, 294, 813–817.

— (2001b), "Dynamics of the 2001 UK foot and mouth epidemic: Stochastic dispersal in a heterogeneous landscape - Supplementary material," .

Kermack, W. and McKendrick, A. (1927), "Contributions to the mathematical theory of epidemics," *Proceedings of the Royal Society. Series A*, 115, 700–721.

Kitching, R., Thrusfield, M., and Taylor, N. (2006), "The use and abuse of mathematical models: an illustration from the 2001 foot and mouth disease epidemic in the United Kingdom," *Revue Scientifique et Technique de l'Office International des Épizooties*, 25, 293–311.

Klein, J. P. and Moeschberger, M. L. (1997), *Survival Analysis: Techniques for Censored or Truncated data*, Springer, New York.

Knorr-Held, L. (2000), "Bayesian modelling of inseparable space-time variation in disease risk," *Statistics in Medicine*, 19, 2555–2567.

Knorr-Held, L. and Besag, J. (1998), "Modelling risk from a disease in time and space," *Statistics in Medicine*, 17, 2045–2060.

Lawson, A. and Zhou, H. (2005), "Spatial statistical modeling of disease outbreaks with particular reference to the UK foot and mouth disease (FMD) epidemic of 2001," *Preventive Veterinary Medicine*, 71, 141–156.

Lawson, A. B. (2001), *Statistical Methods in Spatial Epidemiology*, Wiley.

Lee, E. T. and Wang, J. W. (2003), *Statistical Methods for Survival Data Analysis*, Wiley, 3rd ed.

Li, Y. and Ryan, L. (2002), "Modeling spatial survival data using semiparametric frailty models," *Biometrics*, 58, 287–297.

Lunn, D. (2005), *WinBUGS Development Interface*.

Maller, R. and Zhou, X. (1996), *Survival Analysis with Long-Term Survivors*, Wiley.

McCullach, P. and Nelder, J. (1989), *Generalized Linear Models*, Chapman and Hall, London, 2nd ed.

McGilchrist, C. and Aisbett, C. (1991), "Regression with frailty in survival analysis," *Biometrics*, 47, 461–466.

McLachlan, G. and Basford, K. (1988), *Mixture Models: Inference and Applications to Clustering*, New York, Marcel Dekker.

McLachlan, G. and Peel, D. (2000), *Finite Mixture Models*, Wiley.

Meester, R., de Koning, J., de Jong, M. C., and Diekmann, O. (2002), "Modeling and real-time prediction of classical swine fever epidemics," *Biometrics*, 58, 178–184.

Menach, A. L., Legrand, J., Grais, R. F., Viboud, C., Valleron, A.-J., and Flahault, A. (2005), "Modeling spatial and temporal transmission of foot-and-mouth disease in France: identification of high-risk areas," *Preventive Veterinary Medicine*, 36, 699–712.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953), "Equations of state calculations by fast computing machine," *Journal of Chemical Physics*, 21, 1087–1091.

Mikler, A. R., Venkatachalam, S., and Abbas, K. (2005), "Modeling infectious disease using global stochastic cellular automata," *Journal of Biological Systems*, 13, 421–439.

Morley, P. and Chang, J. (2004), "Critical behavior in cellular automata animal disease transmission model," *International Journal of Modern Physics*, 15, 149–162.

Murray, J. (2003), *Mathematical Biology I - An Introduction*, Springer, 3rd ed.

Murray, J., Stanley, E., and Brown, D. (1986), "On the spatial spread of rabies amongst foxes," *Proceedings of the Royal Society - B*, 229, 111–150.

Nelder, J. and Wedderburn, R. (1972), "Generalised linear models," *Journal of the Royal Statistical Society. Series A (General)*, 135, 370–384.

Ng, S. and McLachlan, G. (1998), "On modifications to the long-term survivor mixture model in the presence of competing risks," *Journal of Statistical Computation and Simulation*, 61, 77–96.

Oakes, D. (2001), "Biometrika Centenary: Survival analysis," *Biometrika*, 88, 99–142.

Oakes, D. and Jeong, J. (1998), "Frailty models and rank tests," *Lifetime Data Analysis*, 4, 209–228.

OIE (2005), "Old Classification of Diseases Notifiable to the OIE," website.

Parkes, C. (1972), "Accuracy of predictions in survival in later stages of cancer," *British Medical Journal*, 2, 29–31.

Pearson, K. (1894), "Contributions to the theory of mathematical evolution, II: skew variation," *Philosophical Transactions of the Royal Society of London A*, 185, 343–414.

Peto, R. (1972), "Contribution to the discussion of a paper by D.R. Cox," *Journal of the Royal Statistical Society. Series B (Methodological)*, 34, 205–207.

R Development Core Team (2005), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Richardson, S. and Green, P. (1997), "On Bayesian analysis of mixtures with an unknown number of components (with discussion)," *Journal of the Royal Statistical Society. Series B (Methodological)*, 59, 731–792.

— (1998), "Corrigendum: On Bayesian analysis of mixtures with an unknown number of components," *Journal of the Royal Statistical Society. Series B (Methodological)*, 60, 661.

Roeder, K. and Wasserman, L. (1997), "Practical Bayesian density estimation using mixtures of normals," *Journal of the American Statistical Association*, 92, 894–902.

Samuel, A. R. and Knowles, N. J. (2001), "Foot-and-mouth disease virus: cause of the recent crisis for the UK livestock industry," *Trends in Genetics*, 17, 421–424.

Sanson, R., Liberona, H., and Morris, R. (1991), "The use of a geographical information system in the management of a foot-and-mouth disease epidemic," *Preventive Veterinary Medicine*, 11, 309–313.

Sanson, R. and Morris, R. (1994), "The use of survival analysis to investigate the probability of local spread of foot-and-mouth disease: an example study based on the United Kingdom epidemic of 1967-1968," *The Kenya Veterinarian*, 18, 186–188.

Sanson, R., Morris, R., Wilesmith, J., and Mackay, D. (2000), "A re-analysis of the start of the United Kingdom 1967-1968 foot-and-mouth disease epidemic to calculate transmission probabilities," in *Proceedings of the Ninth ISVEE, Breckenridge, CO (abstract #256)*.

Savill, N., Shaw, D., Deardon, R., Tildesley, M., Keeling, M. J., Woolhouse, M., Brooks, S., and Grenfell, B. (2007), "Effect of data quality on estimates of farm infectiousness trends in the UK 2001 foot-and-mouth disease epidemic," *Journal of the Royal Society Interface*.

Segel, L. (1972), "Simplification and scaling," *SIAM Rev.*, 14, 547–571.

Sellers, R. (1971), "Quantitative aspects of the spread of foot and mouth disease," *Veterinary Bulletin*, 41, 431–439.

— (2006), "Comparison of different control strategies for foot-and-mouth disease: a study of the epidemics in Canada in 1951/52, Hampshire in 1967 and Northumberland in 1966," *Veterinary Record*, 158, 9–16.

Shimakura, S. E. (2003), "Statistical methods for spatial survival data," Ph.D. thesis, Lancaster University, Lancaster.

Sirakoulis, G. C., Karafyllidis, I., and Thanailakis, A. (2000), "A cellular automaton model for the effects of population movement and vaccination on epidemic propagation," *Ecological Modelling*, 133, 209–223.

Smith, N. H., Gordon, S. V., de la Rua-Domenech, R., Clifton-Hadley, R. S., and Hewinson, R. G. (2006), "Bottlenecks and broomsticks: the molecular evolution of *Mycobacterium bovis*," *Nature Reviews Microbiology*, 4, 670–681.

Sorensen, J., Mackay, D., Jensen, C., and Donaldson, A. (2000), "An integrated model to predict the atmospheric spread of foot-and-mouth disease," *Epidemiology and Infection*, 124, 577–590.

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2003), *WinBUGS User Manual, Version 1.4*.

Stegeman, A., Elbers, A. R., Smak, J., and de Jong, M. C. (1999), "Quantification of the transmission of classical swine fever virus between herds during the 1997-1998 epidemic in The Netherlands," *Preventive Veterinary Medicine*, 42, 219–234.

Stephens, M. (1997a), "Discussion on 'On Bayesian analysis of mixtures with unknown number of components' (by S. Richardson and P.J. Green)," *Journal of the Royal Statistical Society. Series B (Methodological)*, 59, 768–769.

— (1997b), "Bayesian methods for mixtures of normal distributions," DPhil thesis, University of Oxford, Oxford.

— (2000a), "Label switching in mixture models," *Journal of the Royal Statistical Society. Series B (Methodological)*, 62, 795–809.

— (2000b), "Bayesian analysis of mixture models with unknown number of components - an alternative to reversible jump methods," *The Annals of Statistics*, 28, 40–74.

Streftaris, G. and Gibson, G. J. (2004a), " Bayesian inference for stochastic epidemics in closed populations," *Statistical Modelling*, 4, 63–75.

— (2004b), " Bayesian analysis of experimental epidemics of foot-and-mouth disease," *Proceedings of the Royal Society: Biological Sciences*, 271, 1111–1117.

Sturtz, S., Ligges, U., and Gelman, A. (2005), "R2WinBUGS: A Package for Running WinBUGS from R," *Journal of Statistical Software*, 12, 1–16.

Taylor, N. (2003), "Review of the use of models in informing disease control policy development and adjustment. A report for DEFRA," website.

Therneau, T. M. and Grambsch, P. M. (2000), *Modeling Survival Data - Extending the Cox Model*, Springer.

Thompson, D., Muriel, P., Russell, D., Osborne, P., Bromley, A., Rowland, M., Creight-Tyte, S., and Brown, C. (2002), "Economic costs of the foot and mouth disease outbreak in the United Kingdom in 2001," *Revue Scientifique et Technique de l'Office International des Épizooties*, 21, 675–687.

Tildesley, M. J., Savill, N. J., Shaw, D. J., Deardon, R., Brooks, S. P., Woolhouse, M. E., Grenfell, B. T., and Keeling, M. J. (2006), "Optimal reactive vaccination strategies for a foot-and-mouth outbreak in the UK," *Nature*, 440, 83–86.

Titterington, D., Smith, A., and Makov, U. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York, Wiley.

Vaupel, J., Manton, K., and Stallard, E. (1979), "The impact of heterogeneity in individual frailty on the dynamics of mortality," *Demography*, 16, 439–454.

Vernon, M. C., Webb, C. R., and Heath, M. F. (2005), "Preliminary analysis of the contact structure of the UK cattle herd," 25th International Sunbelt Social Network Conference, California, USA.

West, M. and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models*, Springer, 2nd ed.

Whitehead, J. (1980), "Fitting Cox's regression model to survival data using GLIM," *Applied Statistics*, 29, 268–275.

WHO (2004), "World Health Organization Disease Watch - Focus: Tuberculosis," website.

— (2006), "World Health Organization - Tuberculosis factsheet," website.

Wilesmith, J., Stevenson, M., King, C., and Morris, R. (2003), "Spatio-temporal epidemiology of foot-and-mouth disease in two counties of Great Britain in 2001," *Preventive Veterinary Medicine*, 61, 157–170.

Wingfield, A., Miller, H., and Honhold, N. (2006), "FMD control strategies," *Veterinary Record*, 158, 706–707.

Yoon, H., Park, C.-K., Nam, H.-M., and Wee, S.-H. (2005), "Virus spread pattern within infected chicken farms using regression model: the 2003-2004 HPAI epidemic in the Republic of Korea," *Journal of Veterinary Medicine Series B -Infectious Diseases And Veterinary Public Health*, 52, 428–431.