Statistical tools for spatio-temporal epidemiology, with application to veterinary diseases.

A thesis presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

at Massey University, Palmerston North, New Zealand.

Author:

Kate Richards

Supervisor:

Prof. Martin L. Hazelton (IFS)

Co-supervisor:

Prof. Mark Stevenson

(University of Melbourne)

December 14, 2015

Abstract

In epidemiology we are concerned with disease occurrence and its associated explanatory factors. Through analysis of the patterns in disease spread, in space and/or time, we are able to obtain information about possible risk factors and transmission mechanisms. The main focus in spatial epidemiology has been human health. However, economic costs and the concern about zoonoses has fuelled a growing field of veterinary epidemiology. Veterinary epidemiology has the added complication of the 'human effect'. For a disease to be recorded we require humans to detect and report the disease, and once reported human intervention is generally applied. This can lead to the true level of disease being under-represented with the loss of information impeding modelling and model predictions.

The reliability of statistical analyses depends on the quality of the underlying data. Anomalies could introduce significant bias and lead to inappropriate decision making. Residual analysis is often used to detect anomalous data, but with hierarchical models (common within epidemiology) the highly flexible representation of variation can mask outliers. We propose the use of exceedance probabilities as a tool for identifying and assessing anomalous data in spatio-temporal models for routinely collected areal disease count data. We illustrate this methodology through a case study on outbreaks of foot-and-mouth disease (FMD) in Viet Nam for the time period 2006 to 2008. The exceedance probabilities identify several provinces where the number of infected communes was unexpectedly low. These findings are particularly interesting as these provinces are located along major cattle movement pathways within Viet Nam.

With epidemic data, the primary interest is the understanding of the transmission of the disease and the effectiveness of intervention strategies. While epidemic curves provide an excellent representation of the temporal patterns, we propose the additional use of a new graphical tool, the 'cluster curve' to summarise the changes in spatial clustering through time. The cluster curve is based on the inhomogeneous K-function, and provides a means for summarizing the progression of clustering in infectious disease outbreak data taking into consideration spatial variation in the underlying population.

We look at the application of the cluster curve to two outbreaks of FMD in England (2001) and Japan (2010) and to the 2007 epidemic of Equine Influenza (EI) in Australia. By comparing our knowledge of the actual course of the outbreak with the insight provided by the cluster curve we are able to showcase the effectiveness of our tool. Throughout the progression of the outbreak several time windows obtained small sample sizes. Therefore, we also look at

the inclusion of significance indicators to definitively differentiate between true clustering and noise due to these small sample sizes.

The epidemic outbreaks studied all had intervention methods applied. The impact of intervention strategies was investigated through the simulation (via InterspreadPlus) of five intervention methods on outbreaks of FMD in two geographical regions. Using the cluster curve, we found that intervention methods that created buffer zones were found to have particular characteristics of spatial spread. We found that non-buffer methods were less effective in controlling local spread. This is most likely due to infection transmission prior to clinical signs. This kind of analysis demonstrates the practical importance of having effective tools for describing changes in the spatial patterns of disease during an epidemic outbreak.

Publications arising from this thesis

Richards, K.K., Hazelton, M.L., Stevenson, M.A., Lockhart, C.Y., Pinto, J. and Nguyen, L. (2014). Using exceedance probabilities to detect anomalies in routinely recorded animal health data, with particular reference to foot-and-mouth disease in Viet Nam. *Spatial and Spatio-temporal Epidemiology* **11**, 125-133.

The data in this thesis and the results we publish within contain privacy issues and are strictly confidential.

Acknowledgements

Firstly, I would like to give my greatest thanks to my primary supervisor Prof. Martin Hazelton. Without him pulling me into his office in my first month of being at Massey, Palmerston North and informing me about the honour track program, I would have never contemplated continuing my study after my undergraduate degree. I also can not thank him enough for his support throughout my postgraduate studies and being the best supervisor anyone could ever hope for.

Secondly, I would like to thank my co-supervisor Prof. Mark Stevenson for all his support with finding interesting datasets to base my PhD around and his work to get grants from the FAO to fund the primary years of my PhD.

I would like to send a large thank you to my family, who have been there though out this journey, with words of encouragement, being a sounding board and even a proof reader when required.

Also my greatest thanks and love goes out to my partner, Ian Donohue, who put up with years of long distance, and many months being ignored so I could be where I am today.

To all my friends I thankyou for all your support, specially Alice (PBF) and Rochelle for making my time away from home enjoyable.

To all the statistic postgraduates, this journey would of not have been the same without you. With a special thanks to all those who made the final stages of my PhD the most enjoyable time, even when stress levels were high.

Thanks must be given to Massey University and especially the Institute of Fundamental Sciences, for their support both financially and through the knowledge of all those involved.

I would also like to thank the FAO for the access to their data as well as there financial support. Also to the other Government agencies that have allowed me access to their data.

Finally, I would like to thank my heavenly father for his all the strength and comfort he provides.

List of Abbreviations

ANEMIS	Animal emergency management information systems
CPPP	Cluster Poisson point process
CSR	Complete spatial randomness
D	Depopulation
DED	Depopulation and pre-emptive culling (depopulation)
DIC	Deviance information criteria
DV	Depopulation and vaccination
EI	Equine influenza
EVI	Enhanced vegetation index
FAO	Food and Agriculture Organisation
FMD	Foot-and-mouth disease
HPP	Homogeneous Poisson point process
IP	Infected premise
IPP	Inhomogeneous Poisson point process
MCMC	Markov chain Monte Carlo
NC	No control
NMB	Non-movement ban
OIE	Office Internationale des Epizooties
PPP	Poisson point process
SIR	Standardized incidence ratio
SP	Suspect premise
SPR	Squared Pearson residual
V	Vaccination

Contents

List of Figures

Chapter 1

Introduction

With epidemiology we are interested in the investigation of the occurrence of a disease in regards to explanatory factors (Lawson, 2006). A lot can be learned about risk factors and possible transmission mechanisms through the evaluation of patterns of disease spread in space and time. The analysis of spatio-temporal data in epidemiology requires sophisticated statistical tools for analysis and reliable interpretation. Within this thesis we aim to describe two new tools, and demonstrate their use for various problems in veterinary epidemiology.

The driving force behind spatial epidemiology has been human health. However, with the economic impact of veterinary diseases as well as the concern with zoonoses (infectious disease of animals transferable to humans), the field of spatial veterinary epidemiology is growing. The aim of the science of veterinary epidemiology is to answer and describe the 'four Ws and an H': the what, when, where, why and how. This is the distribution of diseases and animal health populations (what, when, where and how much) and the investigation of the contributing factors for the disease occurrence (the why) (Christensen, 2001).

The basic principles we apply to human health data also apply to animal health data, however, there is the presence of added complications due to human input in the latter case. These complications include the requirement of humans to detect and then report the disease since animals themselves cannot report the sickness, and the fact that humans generally intervene once a disease is found. For an event to be recorded in veterinary epidemiology we first require someone (owner, farm worker, veterinarian, etc.) to detect the disease and then report the disease to the appropriate authorities. This multi-step process can lead to the true level of the disease and its spread being under-represented, making the modelling process more difficult. The intervention of humans with an animal disease is intended to control and/or eradicate the disease, limiting and controlling the spread of the disease. Even through the reduction

in spread is entirely desirable, from the analytical side it can lead to a significant amount of missing information, which can impact on the quality of model predictions (Lawson, 2013).

In this thesis we illustrate the application of our statistical tools on two veterinary diseases, Foot-and-mouth disease (FMD) and Equine influenza (EI). FMD is a is a highly contagious virus that can be transmitted by direct and indirect animal contact as well as airborne spread. It can affect all types of cloven-hoofed animals, principally affecting cattle, sheep and pigs (Bachrach, 1968). We look at the endemic cases in Viet Nam over the period 2006–2008, as well as two epidemic outbreaks: England 2001 and Japan 2010. EI is a highly contagious respiratory disease with two known subtypes identified in horses. Similar to FMD it can be spread by direct and indirect animal contact as well as by airborne spread. We examine the 2007 epidemic in Australia. We discuss FMD and its outbreaks in more detail in Chapter 3 and Equine influenza in Chapter 8.

These datasets include both areal and point process data. Areal data is also known as count data, as cases of disease are aggregated and this count is then associated with a region (Lawson and Zhou, 2005). Point process data is also known as case event data, as we know the locations of individual disease. We provide an overview of these data types, and review methods and models for their analysis, in Chapter 2.

As mentioned above, one of the characteristics of veterinary epidemiology is that it is heavily reliant on clinical symptoms being identified by another party, since animals themselves cannot report sickness. This can lead to problems with data quality and with the disease being undetected or under-reported. For example, it is common knowledge that FMD outbreaks in endemic regions often go under-reported, with this being partially related to either the political development or economic level of the country (Sumption et al., 2008). Anomalies in data records have the potential to introduce significant bias in the results of descriptive analyses and fitted statistical models, skewing results and potentially leading to inappropriate decision making.

In general the statistical methodology for identifying anomalous data is residual analysis (e.g. Gail, 1991). For simple models, such as linear regressions with fixed effects, this type of analysis is straightforward to implement and interpret. In this case a large residual would suggest a data record that is poorly predicted by the model, indicating either a problem with the data point in question or inadequacy of the model as a whole (or both). However, when we are dealing with the kinds of complex models required to handle spatio-temporal epidemiological data, residual analysis becomes more complex. For example, hierarchical models are often employed, which seek to describe the variation in disease occurrence by incorporating random variables at multiple levels (see, for example, Lawson, 2013). The flexibility that these models allow are not without shortcomings, as it may cause possible

'odd' data records to be 'absorbed'.

In Chapter 4 we propose the use of *exceedance probabilities* as a tool for identifying and assessing anomalous data in spatio-temporal models for routinely collected areal disease count data. Exceedance probabilities can be applied at any level of the model to describe the extent to which an individual random term, or combination of random terms, is unusual, in the sense of lying in the extreme tails of the specified distribution. Exceedance probabilities have been used previously to detect anomalous clusters of cases in point process data (e.g. Diggle et al., 2005; Davies and Hazelton, 2013), and have also been used to identity regions with unusually high relative risk when modelling areal count data (e.g. Best et al., 2005; Lawson, 2010). We show wider uses of exceedance probabilities, demonstrating their application to individual stochastic terms in areal count models and also noting their utility for highlighting localities with unexpectedly low (as well as high) reported rates of disease. We illustrate the use of exceedance probabilities as a diagnostic tool for anomalous data through a case study of the Viet Namese FMD data described above.

While exceedance probabilities (and other model diagnostics) are of considerable use with routinely collected data, the problems posed by epidemic outbreaks are rather different. We know (by definition) that the disease pattern in an epidemic is different to routinely recorded endemic cases. The problem in an epidemic is to try and understand the transmission of the disease in the outbreak, and decide on an effective intervention strategy. To help address these issues, there is typically a lot of active data collection during an epidemic.

A common method that epidemiologists use to describe the temporal pattern of a disease is the epidemic curve. An epidemic curve in its simplest form plots the number of new cases over a sequence of time periods. The visual portrayal of this temporal trend can give insight into incubation times, and be used for comparisons of disease frequencies. While epidemic curves can provide important clues about likely mechanisms of disease spread, they provide limited knowledge about factors driving disease transmission. For example, in an infectious disease epidemic, are large numbers of cases observed over a short period of time due to spread from local contact, or simply due to many sporadic cases? One way to assess the characteristics of disease transmission is to look at spatial clustering of cases.

With disease and ill-health, clustering in space and/or time is common place. This is generally because the disease is contagious or point source in nature. Consequently, if we fail to consider and investigate such clustering, any procedures to control and eradicate could be hampered (Carpenter, 2001).

Clustering in point pattern processes can be investigated through second order analysis, for example using the inhomogeneous K-function. In Chapter 5 we present a new graphical tool,

the 'cluster curve', which is based on the K-function and provides a means for summarizing the degree of clustering in infectious disease outbreak data, and how this changes during the course of an epidemic. We propose that this tool/method should be used in conjunction with the epidemic curve to shed more light on the progression and properties of an outbreak.

We develop a variety of enhancements for the basic cluster curve, including methods to highlight statistical significance, and software developments that facilitate its use. In Chapter 5 we apply the cluster curve in a graphical window with a slider to allow for easy changes of spatial resolution. In Chapter 6 we look at extending the cluster curve to include significant indicators to definitively differentiate between true clustering and noise due to small sample sizes within a given time window. We also look at the application of the cluster curve as a web application to make the tool more readily available for use in the veterinary epidemiological field.

We apply our cluster curve to two outbreaks of FMD in England 2001 and Japan 2010 (Chapter 5 & 6) and the 2007 epidemic of EI in Australia (Chapter 8). All three of these epidemics had intervention strategies applied. The English outbreak was controlled with a combination of non-movement bands, all FMD positive farms being culled and any farms within 1.5 km of an infected farm also being culled. The Japanese outbreak used non-movement bands and ring vaccinations. The Australian EI outbreak used zone based movement restrictions and vaccination to take control of the outbreak and regain the country's disease free status. With these disease outbreaks, as we are carrying out post-hoc analysis, we know information on the outbreak spread. However, if we assumed we had no prior knowledge of the outbreak we can use the cluster curve to infer changes in spatial clustering through time and these inform how the disease spreads. By comparing our knowledge of what actually happened with the insight provided by the cluster curve we are able to showcase the effectiveness of this tool.

As the examples above illustrate, many disease outbreaks are not left to run their natural course but instead result in some form of intervention being applied to stop the spread and ideally eradicate the disease. These control methods usually occur as combinations of slaughter, contact reduction, chemical use, modification of host resistance and environment, and management controls. To investigate the impact of these methods on the changes in spatial clustering through time we used InterspreadPlus (Stevenson et al., 2013) to simulate 5 intervention methods on outbreaks of FMD in two geographical regions. We used simulated datasets rather then real world epidemics as we want to compare the effects of strategies without the results being (heavily) influenced by the characteristics of the individual disease outbreaks. For example a particular method 'A' may work better than another 'B' in a real life scenario, but only because A was applied to a much more straightforward outbreak. The application of the cluster curve to these datasets is presented in Chapter 7.

In Chapter 9 we draw together our findings and conclusions from earlier chapters. We also explore potential limitations and possible extensions of our work, and hence map out some avenues for future research.

Chapter 2

Review of methods and models

2.1 Data types

In spatial epidemiology there are two basic types of data: count and case event data. Lawson and Zhou (2005) describe count data as data that occurs when the number of cases of disease are accumulated and the count is associated with a region (also known as areal data), while case event data occurs when the locations of individual disease are known (also known as point process data). It can be said that for data to be classed as case event, the probability of more than one event occurring at any one location should be negligible. However, the classification of data depends ultimately on the objective of the study and the data available. For example, if we were looking at the distribution of a disease across a country, plotting the residential addresses may be impractical and unnecessary, resulting in information overload; instead a more practical solution may be looking at the statistics summarised over an area (such as suburb). Alternatively, if we want to look at the relationship between residential proximity to a pollutant source and ill health, then using the point location of residential addresses is better suited (Lawson and Williams, 2001).

Though the basic principles in disease mapping hold for both types of data, there are differences in the modelling methodology.

2.1.1 Case event data (point pattern)

For case event data we define the area under observation as the study window (W), then within this window we have m disease case events that occur at locations $\boldsymbol{x}_i = (x_{i1}, x_{i2})^T, i = 1, ..., m$ (Lawson, 2006). As mentioned, case event data occurs when the locations of individual cases

of a disease are known. The classification of the point location depends on the spatial scale of the analysis (Lawson and Zhou, 2005). For example, if we were looking at cases occurring in a city, a point could represent a residential address, while if we were looking at the region such as a country, a point may represent a suburb or even a city. The observational scale should be very large in relation to the point scale in order for it to be classed as a point process. This is because point process theory relies on the concept of orderliness, and under this, the basic principle is that there should be minimal probability that more than one event can occur at any point location (Lawson, 2006).

An example of a case event map is shown in Figure 2.1. Our study window is the entire region of Viet Nam, and our point locations represent the centroid coordinates of communes that had at least one reported case of foot and mouth disease in cattle for 2006. To consider the spatial scale, Viet Nam can be broken down into 64 provinces, further into districts, and finally into 11052 communes covering an average of 30 km^2. The spatial resolution of the data is sufficiently fine to be considered a point process.

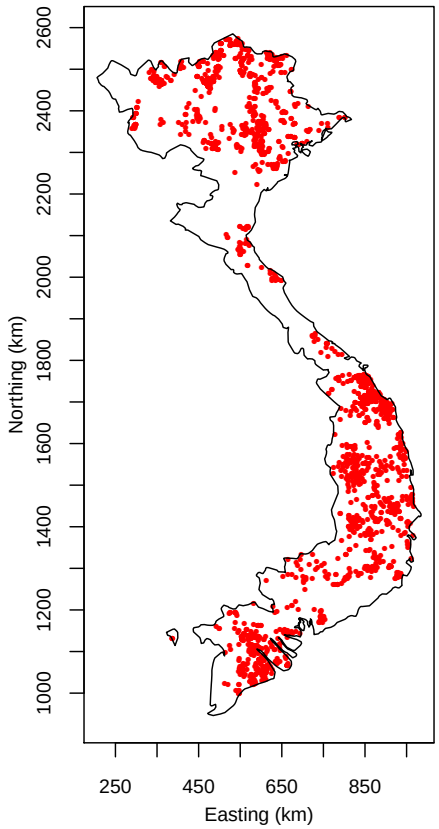


Figure 2.1: Example of case event data: Map of Viet Nam showing the point location of commune-level FMD outbreaks for 2006

2.1.2 Count data (areal)

Count data result from the accumulation of cases of a disease, with these counts being associated with a subregion (Lawson and Zhou, 2005). For count data we once again, as per the case event data, define our study window as W. Within this window we have m bounded subregions. Within these m bounded subregions we have event counts denoted $n_i, i = 1, ..., m$ (Lawson, 2006). The occurrence of count data generally occurs due to confidentiality. In many situations the exact location of a disease, such as a residential address, will not be publicly available due to confidentiality of the resident, and instead a count of infected residents will be available over a small region, such as a street or suburb (Lawson and Williams, 2001). In veterinary epidemiology a farm could be classed as a small area in which the count would represent the number of infected animals present on that farm. However, at a larger spatial scale, and in particular with highly contagious disease (once present on a farm all susceptible animals on the property are assumed to be infected), the count could be considered as the number of infected farms within a wider region (Lawson and Zhou, 2005).

An example of a count data map is shown in Figure 2.2. Our study window is the entire region of Viet Nam, and our count data represent the number of communes that had at least one reported case of foot and mouth disease in 2006 per province. This plot is an areal version of our point data example.

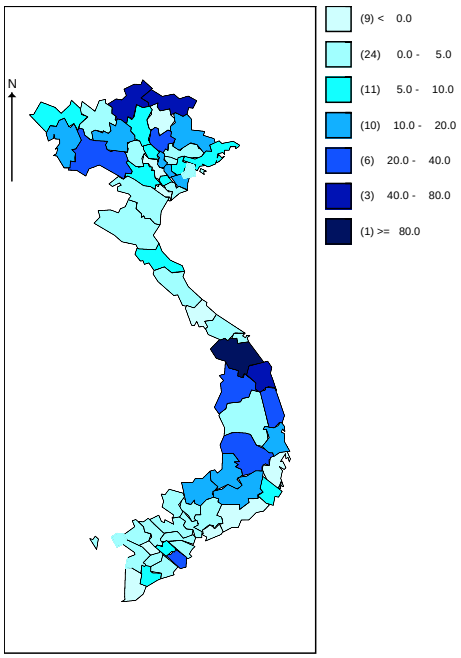


Figure 2.2: Example of count data: Map of Viet Nam showing the number of infected communes with FMD in 2006 per province

In the analysis of the spatial distribution of diseases, it is acknowledged that basic models exist that are generally assumed to work or act as a starting point for further models in both case event and count data. Since count data is usually derived from from case event data, it is logical to first consider models for case event data (Lawson, 2006).

2.2 Point process data

Generally we talk about case event data as a point process or a point pattern. The term point process usually refers to a stochastic process that generates the locations of events of interest within the defined study window, W (Bivand et al., 2008). A point pattern can be considered as a dataset in which the spatial location of all events observed within the study window are recorded (Baddeley and Turner, 2005).

For spatial patterns and processes we use the term 'event' for each location of an observed point to distinguish between that which is of interest and any other arbitrary locations within the window (Gatrell et al., 1996; Diggle, 2000). For these events we record the locations using vector notation as a shorthand for referring to the x and y coordinates. We have a set of locations ($\boldsymbol{x}_1$, $\boldsymbol{x}_2$, etc.) where $\boldsymbol{x}_i$ refers to the location of the ith observed event, with x_{i1} and x_{i2} representing the x and y coordinate, respectively, for that event (Gatrell et al., 1996).

For every point pattern and process we have a defined study window, W. Regardless of the size and shape of this window, consideration needs to be taken in terms of the effect the border or edge could have on the analysis. Edge effects occur because generally our study window is part of a larger region where the underlying process occurs. Therefore, events can occur outside of our window and interact with events inside our window. As the events that occur outside of the window are not observed it is difficult to properly take into account their effect (Diggle, 2003). We can try to account for edge effects by creating a guard area between the border of the original study window and a slightly smaller subregion in which the analysis is carried out, or by adapting the analytical tool to take into account any edge effects (Gatrell et al., 1996).

In general, with spatial stochastic processes, aspects of their behaviour can be classed in terms of its first/second-order properties. Essentially, the first-order properties describe the way in which the expected value, the mean, of the process differs through space, while the second-order properties describe the covariance or correlation between events from the process at different locations in space (Gatrell et al., 1996).

When specifying a point process we generally refer to them based on the underlying intensity,

more specifically, as a homogeneous poisson point process or a heterogenous poisson point process. With a homogeneous point process the underlying intensity is assumed to remain constant over the entire study region, while for a heterogeneous point process the underlying intensity is allowed to vary over the region. This is explained in more detail in the following section.

2.2.1 Spatial Point Process Models

Point pattern data can be modelled as the outcome of a spatial point process model. If $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ denote the locations of the points in some study window W, then it is common to write $N(A)$ for the number of those points which lie within some region A contained within W. The properties of the spatial point process can then be described in terms of N.

The first and second order properties of a single random variable can be described by its mean and variance. The analogue of the mean for a spatial point process is the intensity function $\lambda(\boldsymbol{x})$. Intuitively, this explains how the expected (mean) number of points varies across the window W (Gatrell et al., 1996). More technically, for any region $A \subseteq W$, the intensity function satisfies

$$E[N(A)] = \int_A \lambda(\boldsymbol{x})\, d\boldsymbol{x}. \tag{2.1}$$

It follows that if $A(\boldsymbol{x_0})$ is a very small region centred at the coordinates $\boldsymbol{x}_0$, then the expected number of points in $A(\boldsymbol{x_0})$ is given approximately by

$$E[N(A(\boldsymbol{x_0}))] \approx a_0 \lambda(\boldsymbol{x_0})$$

where a_0 is the area of $A(\boldsymbol{x_0})$.

If the intensity $\lambda(\boldsymbol{x})$ is constant over the study region, the point process is referred to as homogeneous (Marcon et al., 2003). If the intensity varies over the study window the point process is called inhomogeneous or heterogeneous.

A homogeneous Poisson point process (HPP) is a point process where all events are independently and uniformly distributed in the study window. Therefore the location of an event has no affect on another event and the expected number of events in a region for a given size does not depend on the location or shape of the region (Bivand et al., 2008).

With inhomogeneous Poisson point process (IPP) the intensity function varies over the study window, but having adjusted for this, the locations of the points are independent. This is generally the case within real world examples of spatial point patterns. The estimation of the intensity function can be carried out via parametric or non-parametric methods (Bivand

et al., 2008). We describe estimation by kernel smoothing in more detail in the following section.

2.2.2 Estimation of the intensity function for point processes

The expected number of points in a spatial point process over the entire window W is

$$\lambda_W = E[N(W)] = \int_W \lambda(\boldsymbol{x})\, d\boldsymbol{x}.$$

If we define a normalized version of the intensity function by $f(\boldsymbol{x}) = \lambda(\boldsymbol{x})/\lambda_W$ then

$$\int_W f(\boldsymbol{x})\, d\boldsymbol{x} = 1.$$

The function f is a probability density function, describing the distribution of a single point location over the region.

It follows that the intensity function can then be decomposed as

$$\lambda(\boldsymbol{x}) = \lambda_W f(\boldsymbol{x}). \tag{2.2}$$

This suggests that estimation of the intensity function can be handled in two stages. The first stage is to estimate the expected number of points in the pattern. A very natural estimate is $\hat{\lambda}_W = n$, where n is the number of points in the observed point pattern. The second stage is to obtain an estimate $\hat{f}(\boldsymbol{x})$ of the density function. The intensity estimate is then $\hat{\lambda}(\boldsymbol{x}) = \hat{\lambda}_W \hat{f}(\boldsymbol{x}) = n\hat{f}(\boldsymbol{x})$.

A common and flexible method of estimating a density function is the use of kernel density estimation (Silverman, 1986). A 'bump' or kernel is centrally placed over each point within the window W. The kernel itself is defined as a bivariate probability density, such as the bivariate normal density, with the overall estimate being obtained by averaging over the succession of small 'bumps' centered on each case. Therefore, the density estimate in areas with numerous observations will be higher than areas with only a few observations (Seaman and Powell, 1996; Brunsdon and Comber, 2015).

More technically, the kernel density estimate can be defined as:

$$\hat{f}(\boldsymbol{x}) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h}\right) \tag{2.3}$$

where K is the kernel function and h is a bandwidth which controls the effective kernel radius

(Seaman and Powell, 1996).

In certain situations we may wish to apply different bandwidths in the x and y coordinate directions. Equation 2.3 can then be re-specified as:

$$\hat{f}(\boldsymbol{x}) = \hat{f}(x, y) = \frac{1}{nh_x h_y} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_x}, \frac{y - y_i}{h_y}\right) \tag{2.4}$$

where $\boldsymbol{x}_i = (x_i, y_i)$ are the coordinates of the ith data point. In this case the parameters h_x and h_y are the bandwidths in the x and y direction respectively. This results in kernels with an elliptical shape, where the lengths of the axes are controlled by the separate bandwidths.

In the application of kernel estimation one must choose appropriate bandwidths. Small bandwidths tending to create a 'spiky' or undersmoothed distribution while large bandwidths create a 'flattened'or oversmoothed distribution (Brunsdon and Comber, 2015). The selection of a bandwidth can be done by eye as well as by the application of algorithms. Scott (1992) and Bowman and Azzalini (1997) recommend some simple rules to support this selection process (Brunsdon and Comber, 2015).

In practice some of the points will almost certainly lie close to the boundary of the region W, and part of the kernel for any such point will spill over the edge of W, creating a bias. This can be address by various forms of edge correction, see for example Wand and Jones (1995) and Hazelton and Marshall (2009).

2.2.3 Second order properties

With the analysis of point processes it is common to want to investigate the strength of interaction between events, having adjusted for variations due to the intensity function. Does the presence of an event promote or inhibit another event occurring nearby? Is there clustering or competition between events? Such questions can be answered by looking at the second order properties of a point process. The K-function is a common method used in describing these properties (Brunsdon and Comber, 2015; Bivand et al., 2008).

The K-function is a measure of the expected number of events that occur within a radius r of a given event. For a point process with constant intensity $\lambda(\boldsymbol{x}) = \lambda$, the homogeneous K-function was developed by Ripley (1976, 1977) and is defined by

$$K(r) = \lambda^{-1} E[N_r]$$

where N_r is the number of events within a radius r of an arbitrary event (Dixon, 2012;

Brunsdon and Comber, 2015).

If the point process is deemed to follow complete spatial randomness (CSR), where the events are uniformly and independently distributed (i.e. a homogeneous Poisson process), the expected number of events within a radius r is the multiplication of the intensity function and the area of the circle of radius r. The K-function for CSR is therefore defined by

$$K_{CSR}(r) = \pi r^2 \tag{2.5}$$

The K-function under the assumption of CSR (equation 2.5), $K_{CSR}(r)$, is generally used as a benchmark to assess the presence of clustering in point processes, sometimes referred to as the theoretical value. This is, if $K(r) > K_{CSR}(r)$ then the number of events within a radius r of an arbitrary event is greater than expected under the independence assumption and therefore, suggests the presence of clustering for that point process. If $K(r) < K_{CSR}(r)$ then there are less points than expected within the radius r suggesting competition amongst events of the point process, indicating a more regular pattern of events than expected under independence (Dixon, 2012; Bivand et al., 2008).

The K-function can be estimated by:

$$\hat{K}(r) = \hat{\lambda}^{-1} \sum_i \sum_{j \neq i} \frac{k_{ij}}{n} \tag{2.6}$$

where the estimated (constant) intensity $\hat{\lambda}$ is specified as

$$\hat{\lambda} = \frac{n}{|W|}$$

in which $|W|$ denotes the area of the entire window, W. In equation 2.6 k_{ij} is an indicator variable that takes a value of 1 if the distance between event i and j is less than or equal to r and 0 otherwise (Brunsdon and Comber, 2015; Francois and Raphael, 1999).

This estimation of the K-function, however, does not take into consider the effect of unobserved points on events occurring along the outside of the borders of the window, known as edge effects. There are numerous methods to correct for edge effects, including the use of buffer zones and toroidal duplication. A common method used to account for edge effects is that proposed by Ripley (1976). This method is widely used for its unbiased and robust nature (Francois and Raphael, 1999). The estimated K-function $\hat{K}(r)$, based on Ripley (1976), is defined as:

$$\hat{K}(r) = \hat{\lambda}^{-1} \sum_i \sum_{j \neq i} \frac{k_{ij}}{n w_{ij}} \tag{2.7}$$

where w_{ij} are weights applied to the K-function to take into consideration edge effects. These

weights are calculated by taking the proportion of the area for a circle, centered at $\boldsymbol{x}_i$ with a radius given by the distance between the two events, x_i and x_j, that is inside the window W (Brunsdon and Comber, 2015; Bivand et al., 2008).

An alternative method is the translation edge correction proposed by Ohser (1983). This is defined as

$$w_{ij} = \frac{1}{area(W \cap (W + (x_j - x_i)))}$$

and is available as one of the in built edge correction functions within spatstat (Baddeley et al., 2015). This is the correction method we will use later in chapter 5.

In these definitions of the K-function we are assuming that the underlying intensity is constant over the window, as occurs for homogeneous point processes. In epidemiology it is unlikely that the underlying intensity is constant and therefore, in the following section we will consider the inhomogeneous K-function.

Baddeley et al. (2000) proposed an extension of the homogeneous K-function for use with inhomogeneous point processes. For a given dataset with intensity function $\lambda(\boldsymbol{x})$ the inhomogeneous K-function can be specified as:

$$\hat{K}_{inhomo}(r) = |W|^{-1} \sum_i \sum_{j \neq i} w_{ij}^{-1} \frac{k_{ij}}{\lambda(\boldsymbol{x}_i)\lambda(\boldsymbol{x}_j)}. \tag{2.8}$$

With real world data the true intensity is generally not known and must be estimated using the methods described earlier. Consequently the inhomogeneous K-function is redefined as:

$$\hat{K}_{inhomo}(r) = |W|^{-1} \sum_i \sum_{j \neq i} w_{ij}^{-1} \frac{k_{ij}}{\hat{\lambda}(\boldsymbol{x}_i)\hat{\lambda}(\boldsymbol{x}_j)} \tag{2.9}$$

The inhomogeneous case of the K-function is a generalisation of the homogeneous case. Equation 2.9 can be reduced to equation 2.7 if $\lambda(\boldsymbol{x}) = \lambda$. Therefore, similar to the homogeneous case, πr^2 can be used as a theoretical baseline under the assumption of independence between points. Specifically, this is the value that would be obtained if the process was truly an inhomogeneous Poisson process, so that there are no interdependencies between the locations of points. Once again when $\hat{K}_{inhomo}(r)$ takes values greater than πr^2 it suggests a greater level of grouping of events than is expected under the independent assumption. Lower values of $\hat{K}_{inhomo}(r)$ suggest more regularity (Bivand et al., 2008).

2.2.4 Methods of cluster detection

With the analysis of point patterns we are generally concerned with investigating if the events appear to be grouped (clustered), if they are randomly distributed across the region (complete spatial randomness, CSR), or if the presence of one event reduces the likelihood of another occurring nearby (regular).

Similar to most spatial techniques in epidemiology, methods for cluster detection were developed for application to human health. The Centers for Disease Control (1990) drafted a series of guidelines for the investigation of clusters for health events. However, the use of techniques to investigate spatial clustering in veterinary medicine is less common, but on the rise (Ward and Carpenter, 2000).

In this section we will focus our attention on clustering techniques in veterinary epidemiology for case event data, under the key classification of spatial, temporal and space-time clustering.

Spatial clustering

In the analysis of spatial clustering the focus is on the location of events over the region, with the temporal component for the events either unknown or ignored. Common methods used in the analysis of spatial clustering include; nearest neighbour, Cuzick-Edwards, spatial scan, and K-function methodologies (Carpenter, 2001).

Nearest neighbour The nearest neighbour index was developed by Clark and Evans (1954) and was originally developed to be used in plant ecology (Ward and Carpenter, 2000). The index is generally used as a statistical tool for exploratory or descriptive analysis (Carpenter, 2001). The statistic can be explained as the ratio of the mean euclidean distance between nearest-neighbour points for a given area ($\overline{D}_{observed}$) and the mean distance that is expected given the same area and a series of randomly distributed points ($\overline{D}_{random}$). More specifically:

$$\begin{aligned} R &= \frac{\overline{D}_{observed}}{\overline{D}_{random}} \\ \overline{D}_{observed} &= \frac{\sum_i d_i}{n} \qquad (2.10) \\ \overline{D}_{random} &= 0.5\sqrt{\frac{|W|}{n}} \qquad (2.11) \end{aligned}$$

where d_i is the distance of the i^{th} point to its nearest neighbour, n the number of points and $|W|$ is the area of the study window. If R is approximately zero the test would suggest that

the points are clustered. A value of around one would suggest a random distribution, and a value of approximately 2.15 (maximum value R can take) would suggest an entirely regular distribution (Ward and Carpenter, 2000).

If we want to test the significance of the distribution not being randomly distributed (i.e. $0 \leq R < 1$) we can calculate the standard deviation for the mean nearest neighbour distance by

$$\sigma_{\overline{D}_{random}} = \frac{0.26136}{\sqrt{N(\frac{N}{A})}}.$$

With this we can then calculate a Z score statistic:

$$z = \frac{(\overline{D}_{observed} - \overline{D}_{random})}{\sigma_{\overline{D}_{random}}}$$

A large positive Z score (beyond say the 0.95 quantile of the standard normal distribution) suggests that the data is regularly distributed. If the Z score takes a large negative value (say more negative than the 0.05 standard normal quantile) then this suggests that the data is clustered, while a value around 0 suggest the data is randomly distributed (Salman, 2003).

The major disadvantage of this method is that it does not account for spatial variation in the population at risk. Rather it assumes that it is uniformly randomly distributed, which is generally not the case in real world examples (Salman, 2003).

Cuzick-Edwards The Cuzick-Edwards method for cluster detection is a variation of the nearest neighbour method, created by Cuzick and Edwards (1990). Their method can be used with case-control data. Unlike the nearest neighbour method the Cuzick-Edwards method looks at the relative distance rather than the actual distance between the points (Salman, 2003). For each case, their statistic works by calculating the number of its k nearest neighbours that are also cases. This is then summed over all cases and its significance evaluated by also calculating an expected count. The result is a test statistic for a one tailed nonparametric test where the presence of large values would suggest that clustering is present (Tango, 2010). As with the evaluation of the nearest neighbour statistic, the significance of the test can be calculated through a Z score.

One of the major limitations for the nearest neighbour statistic was its assumption of a randomly distributed underlying population. However, this is not the case for the Cuzick-Edwards method. This method takes into consideration the underlying population at risk by comparing the locations of cases and controls, inherently considering the clustering that may naturally occur in the population under investigation (Durr and Gatrell, 2004). The limitation for this model, however, is that the results can be highly sensitive to the value of

k chosen (Tango, 2010).

Spatial scan The spatial scan statistic is also known as the circular spatial scan statistic as the method identifies possible clusters by applying a series of circular windows of different radii to a dataset (Tango, 2010). This method can locate as well as test the significance of each cluster (Carpenter, 2001). Application of this tool is carried out via SatScan software (Kulldorff and Information Management Services, 2009). For a given data set a series of circular windows of varying radius are applied, from zero to a predefined maximum. The end result is the study area is covered in a variety of circles in all different sizes and locations, with each circle representing a possible cluster. The second stage is the test for significance. This can be done through the use of the likelihood ratio test statistic which ranks all likely clusters (Tango, 2010).

The original version of this method had reduced power if the clusters were non-circular in shape. This issue was (partially) addressed through a generalization of the methodology to include elliptical clusters (Kulldorff et al., 2006). The method has limitations if the cluster size is small and clusters occur in multiple locations (Durr and Gatrell, 2004).

K-function The use of the K-function for spatial cluster detection is becoming increasingly popular. We have previously described the K-function in detail (section 2.2.3) for both the homogeneous and inhomogeneous cases. The use of the K-function in cluster detection is different to the methods previously mentioned, as it is more a method for describing a pattern.

Temporal clustering

Clustering does not only occur in space, but can also occur in time. Methods commonly used to investigate temporal clustering in veterinary epidemiology include the Ederer-Myer-Matel (EMM) test and a temporal version of the scan statistic.

Ederer-Myer-Mantel (EMM) The Ederer-Myer-Mantel (EMM) is one of the most frequently used methods for detecting clusters in time (Carpenter, 2001). It was developed by Ederer et al. (1964) for the use in testing if leukemia cases come in clusters. Their approach is based on cell occupancy theory, more specifically their goal was to calculate the maximum frequency of cases within disjoint subintervals of time and evaluate if these are significantly large. For this method the dataset is broken down into m subintervals in time. The statistic is

$$M = \max(n_1, ..., n_m), \qquad n = n_1 + ... + n_m$$

where n_i is the number of cases in the i^{th} interval and n the overall sample size. Large values of M suggest the presence of clustering in time for the disease in question (Tango, 2010). To test if a value is large we can calculate an estimate for the expected number of cases given a randomly distributed population. If M exceeds the expected values to a sufficient degree then clustering in time is suggested (Carpenter, 2001). The significance can be tested using a χ^2 test with one degree of freedom (Tango, 2010).

This method is not sensitive to changes in the underlying populations over the study area, but can be biased by changes in the population over time. If the data is sparse this method is not recommended (Ward and Carpenter, 2000).

Temporal scan statistic The temporal scan test is equivalent to the spatial scan statistic for time. The scan statistic for use in temporal clustering was first proposed by Naus (1965). Their method involves identifying clusters by "scanning" the whole time series with a predetermined window radius, d. To obtain the best results this radius should be selected to match the natural clustering width. In the application of this method, difficulties were experienced in the specification of the window (Ward and Carpenter, 2000), as this is required to be chosen without looking at the data. This process of choosing a window size is normally done via multiple testing of plausible sizes. Naus's method was later extended by Nagarwalla (1996). Nagarwalla (1996) extension accounts for multiple testing by making the window variable, where the width d does not have to be decided prior. The test statistic is based on a likelihood ratio test:

$$\lambda = \sup_{d,k \geq n_0} \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k} \left(\frac{T}{d}\right)^k \left(\frac{T}{T-d}\right)^{n-k}$$

where n is the total number of points in the study, k the number of points in the window, n_0 a predetermined lowest number of cases needed for a cluster, and T the length of the total study region. For this method, detection of a hot-spot, would suggest the presence of temporal clustering in the study period and Monte Carlo methods can be used to simulate a p-value (Tango, 2010). The spatial scan statistic proposed by Kulldorff and Nagarwalla (1995) and Kulldorff (1997) (presented earlier) can also be used to identify temporal clusters (Tango, 2010).

Space-time clustering

When clustering is found in space or time, the next step is to test whether these occur simultaneously. This is commonly done in veterinary epidemiology by Barton's method, the Knox method, the Mantel method, or the space-time scan method (Carpenter, 2001).

Barton The Barton test, proposed by David and Barton (1966), is comparable to an analysis-of-variance test. Time is included in this method by being treated as a covariate. The test works by looking at the ratio between the within-cell variance to the overall variation, more specifically comparing the spatial distribution of cases within the same time interval with the spatial distribution of cases in other time intervals. A disease is suggested to be clustered in space and time if the variation within-cells is less than that of between the cells (Carpenter, 2001; Tango, 2010).

This method has a few weaknesses in terms of its diagnostic power, the first concerning the ability to detect small diameter clusters and the second; the presence of overlapping clusters. Generally the distances that are most interesting are those that are small, but these appear less significant. Large distances can overshadow the statistics of smaller ones, which can lead to the masking of possible clusters. If clustering occurs in different spatial locations but in the overlapping time intervals the analysis can also be distorted (Tango, 2010).

Knox The Knox method for detecting space-time clustering was developed by Knox and Bartlett (1964) and has been widely used since the mid-1960s. Their method works by examining the numbers of "close" pairs observed for both space and time (Tango, 2010). The choice of length for which a pair is considered to be close (different for space and time) must be determined prior to the investigation for clustering. Generally these are chosen based on epidemiological knowledge, such as the duration of incubation for the disease or the likely distance of transmission. The determination of this distance can be a weakness for this method of cluster detection as it is often difficult to justify the chosen "closeness" threshold (Durr and Gatrell, 2004). Once the threshold has been chosen a 2x2 contingency table is created by dividing the $\frac{n(n-1)}{2}$ (n being the total number of cases) pairs of cases into 'close in space only', 'close in time only', 'close in neither space or time', and 'close in both time and space'.

If the test statistic is large it would suggest the presence of clusters in space and time. To determine what is 'large' we can create an estimated expected value by multiplying the proportion of pairs of cases close in space with the proportion of pairs close in time, giving us the expected proportion of cases close in space and time. This is then multiplied by the number of pairs of cases to give the number of pairs expected to be close in both space and time. When the observed value is significantly greater than the expected value it implies that there is clustering in space and time (Carpenter, 2001).

Mantel The Mantel method can be said to be a generalisation of the Knox method, as the Knox method can be derived from Mantels method when treated as a special case. The

major advantage of the Mantel method over Knox's approach, is that there is no longer a need to pre-specify the distance threshold for space and time to be considered close. Instead the actual data values are used and the correlation is assessed between the space and time distances for all pairs (Carpenter, 2001; Durr and Gatrell, 2004).

Mantel (1967) suggested applying a reciprocal transformation of distances so the influence of large distances would be reduced and smaller distances would have a greater influence. However, this can cause problems when the distances between cases are very small, so small arbitrary constants are sometimes added to the distances. However, the choice of these constants can affect results (Tango, 2010).

Space-time scan statistic The space-time statistic operates in a similar way to the spacial-scan statistic but this time the window is operated in a cylindrical shape where the height of the window represents the temporal component of the cluster (Kulldorff, 2001; Carpenter, 2001). Once again the statistic radius moves across the study window with a varying radius of zero to a defined maximum. Once a cluster is identified, the significance is tested via conditional likelihood ratios, with the largest ratios being the most likely clusters. Similar to the spatial-scan statistic, the test is carried out through the use of the SaTScan software (Kulldorff, 2015). If clusters are identified via this software it would suggest that the disease is clustered within the space and time (Tango, 2010).

Summary of Spatial, Temporal and Space-Time Clustering Methods

It is very common to see large amounts of significant clusters in space and time during a veterinary epidemic outbreak, generally because of the contagious nature of the diseases. Spatial, temporal and space-time cluster methods can identity the presence of clusters, but do not provide any real insight into how (if at all) the spatial pattern of clustering is evolving through time. Such knowledge is important in understanding the progression of an outbreak, and also in determining the efficacy of intervention strategies. With that in mind, in chapter 5 we will begin to look at methods that specifically focus on determining how patterns of spatial clustering change through time.

2.3 Areal data

As previously mentioned there are many occasions in which the point location of an outbreak, even if known, is unavailable and instead the data comes in the form of counts assigned to

small areas. Because of how common these data are, considerable research has occurred in the methods of analysis for areal data (Lawson and Williams, 2001).

2.3.1 Explanatory variables

Remote sensing

Since the early 1970s scientists have been using remotely sensed data to investigate the earth's biotic and abiotic components (Beck et al., 2000). With the capabilities of modern technology these remotely sensed variables can be included in modelling health data (Glass et al., 2000). They provide the opportunity for a greater understanding of disease mechanisms as well as the implementation of control methods (Glass et al., 2000; Graham et al., 2004). For example; climate can have a significant effect in disease transmission, as viruses as well as parasites can be sensitive to climatic conditions (Hay et al., 1996; De La Rocque et al., 2004).

Remotely sensed images are most commonly available as a 2D array of squares (Hay et al., 1998) and can provide explanatory variable information such as vegetation type, land cover use, land surface temperature, soil moisture, slope, presence of bodies of water, rainfall, humidity and elevation (Rinaldi et al., 2006; Graham et al., 2004).

There are two broad categories of satellites; geostationary and polar-orbiting. Geostationary satellites orbit at the equator, they travel at the speed of the earth's rotation so they remain fixed focusing on a section of the earth. Polar-orbiting satellites orbit the earth repeatedly, with every orbit passing over a different section of the rotating earth (Hay et al., 1996). Depending on what we are wanting to investigate also determines the spatial resolution we require from satellite images. Climatic variables only require a coarse resolution as they generally influence disease epidemiology in a similar manner over large areas, while land cover information needs a medium to fine spatial resolution as detail varies down to a small area estimate (Graham et al., 2004). A complication with satellite images, is that cloud cover can affect the amount of information we can obtain (Hay et al., 1996).

2.3.2 Modelling

Spatial models

First we consider a basic model for areal disease mapping by focusing our attention on a purely spatial scale. As we are generally looking at the counts of deaths per specific area it is common to assume a Poisson distribution (Lawson et al., 2003).

Wakefield et al. (2000) provide a good example of a basic three level hierarchical model based on the aggregation of the underlying individual risk level:

$$Y_i \sim Poisson(e^{S_i} E_i), \quad i = 1, 2, 3, ..., n$$

$$S_i \sim p(.|\theta)$$

$$\theta \sim \pi()$$

where: Y_i is the observed number, E_i is the expected number of cases for area i (under the assumption of uniform risk), S_i is the log relative risk in area i, $p(.|\theta)$ is an appropriate second stage prior distribution for S_i, and π is the prior distribution for the parameter vector θ. Their model is suitable when the disease is rare and it is based on the assumptions that the individual risk level varies randomly within the area and risk associated with a particular area acts proportionally on the baseline risk for each area.

For many outbreaks the time scale of the events is also known. This information can be included in the modelling process.

Spatiotemporal models

When the time component is known, the response becomes the number of cases in each area in each time period. For modelling purposes we denote Y_{it} as the number of cases in area i in time period t. The main focus with spatial temporal modelling is estimating the Poisson mean which varies with i and t. It is generally assumed that

$$Y_{it} \sim Poisson(E_{it}\theta_{it})$$

where θ_{it} is the unknown true relative risk and E_{it} is the expected number of cases under the assumption of uniform risk. For a basic spatial temporal model the relative risk logarithm can be specified as:

$$\log(\theta_{it}) = u_i + v_i + \tau_t + \gamma_{it}$$

with u_i being the spatially correlated extra variation, v_i the uncorrelated extra variation, τ_t the temporal variation and γ_{it} the space-time interaction. Commonly τ_t and γ_{it} are assumed to follow random walks, $\tau_t \sim N(\tau_{t-1}, \sigma_\tau^2)$ and $\gamma_{it} \sim N(\gamma_{i,t-1}, \sigma_\gamma^2)$ respectively, which allows for a smooth variation in time (Unkel et al., 2012).

Veterinary models

The majority of initial research and modelling was driven by human health data. Though the basic principles remain for veterinary data, there are added complications. These complications include the use of human intervention to control an outbreak, and that the animals themselves cannot report sickness. Control methods used to limit and control spread can have dramatic effect, altering the progression of the disease. Even through the reduction in spread is desirable from an epidemiological point of view, from the analytical side it can lead to a significant amount of missing information, which can later impact on the quality of model predictions. Since the disease reporting process first requires someone (owner, farm worker, veterinary, etc) to detect the disease and then report the disease to the appropriate authorities, the true level of the disease and its spread can be under-represented. This can therefore add difficulty in the accuracy in modelling the disease (Lawson, 2013).

An example of spatial veterinary modelling is provided by Stevenson et al. (2005) where they use Bayesian Poisson models to describe the geographical distribution of Bovine spongiform encephalophathy (BSE). The disease itself is not contagious and relatively rare so the observed number of cases in each area (O_i) was assumed to follow an independent Poisson distribution with the average number of cases (μ_i) equal to the product of the expected number of cases (E_i) and an estimated area-level relative risk. Their model can be expressed by

$$\log(\mu_i) = \log E_i + (\alpha + \beta_1 x_{1i} + ... + \beta_m x_{mi}) + U_i + S_i$$

where E_i is estimated by

$$n_i \left(\frac{\sum_{i=1}^{178} O_i}{\sum_{i=1}^{178} n_i} \right)$$

with n_i being the total cattle population in the ith area and 178 being the total number of areas. The model includes m area-specific fixed effects $(\beta_1, \ldots, \beta_m)$ associated with explanatory variables $x_1, \ldots, x_m$, and spatially correlated and non-spatially correlated terms S_i and U_i respectively. They applied flat priors for the intercept α and regression coefficients $\beta_1, ..., \beta_m$. They applied a normal prior to U_i, while S_i was assigned a conditional intrinsic Gaussian autoregressive (CAR) prior.

CAR priors are commonly used in Bayesian analysis of spatial data (Hodges et al., 2003). They were first proposed by Besag (1974) and made popular in disease mapping by Besag et al. (1991). CAR priors allows the posterior estimates for a region to take into consideration the neighbouring regions (Hodges et al., 2003; Jin et al., 2005).

An example of spatiotemporal veterinary modelling is provided by Branscum et al. (2008). They used a flexible Bayesian Poisson regression model to describe the annual provincial

occurrences of FMD of Turkey from 1996 to 2003. They defined their model as

$$Y_{i,t} \sim Poisson(\mu_i)$$

$$\log(\mu_i) = g_{i,t} + \beta x_{i,t} + \eta_i$$

where $Y_{i,t}$ is the yearly counts of cases in province i for year t and $x_{i,t}$ for the explanatory variables with its corresponding regression coefficient βs. The function $g_{i,t}$ models the longitudinal trend specified by a Gaussian process. They used a Gaussian process because it provided the ability to have a wide variety of temporal shapes. Similar to the spatial model example above, a CAR prior was applied to the spatial process η_i.

Further examples of veterinary disease mapping models for FMD are described in sections 3.3.3 and 3.4.3.

2.3.3 Model selection and assessment

There is an increasing array of complex models available to today's biostatisticians and epidemiologists. However, the most complex model is not necessarily the best model. We therefore require a method, or methods, to assess model fit and complexity to determine which provides the 'best' representation of the data.

One of the main methods for model selection is the use of information criteria. The three most popular criteria are the Bayes Information Criterion (BIC), the Akaike Information Criterion (AIC) and the Deviance Information Criterion (DIC).

DIC is an extension of AIC, and for simple one-stage models they are identical, however, in cases of hierarchical and latent variable models, differences occur. This is because while AIC uses the actual number of parameters, DIC uses the number of 'effective' parameters (Ntzoufras, 2009).

The reason for using these criteria rather than the deviance directly, is that the deviance does not take into consideration the number of parameters in the model. Including more parameters in the model will allow the model to more closely represent the data, but at the same time adds complexity to the model. The above mentioned information criteria, are attempts to take into consideration this increasing complexity by penalizing it (Lawson, 2013).

For Bayesian modelling the DIC method is frequently used, hence, we will focus our attention on this form of information criteria. For more information on AIC and BIC see (Lawson,

2013).

DIC

DIC was proposed by Spiegelhalter et al. (2002*a*) for the use in comparing complex hierarchical models. Their idea was based on the principle: DIC = 'goodness of fit'+'complexity' (Spiegelhalter et al., 2014). More specifically it can be defined by:

$$DIC(M) = D(\bar{\theta}_M, M) + 2p_M$$

where p_M is the number of 'effective' parameters for model M specified by

$$p_M = \overline{D(\theta_M, M)} - D(\bar{\theta}_M, M)$$

where $\overline{D(\theta_M, M)}$ is the posterior mean deviance for the model and $\bar{\theta}_M$ is the posterior mean of the parameter vector θ in model M. When we are evaluating the DIC, the smaller the value, the better the model fit (Ntzoufras, 2009; Lawson, 2013).

The use of DIC as a model selection tool is widely exercised in biostatistics and ecology (Spiegelhalter et al., 2014), where hierarchical models are common.

One of the limitations of DIC is that it assumes that the posterior mean provides a good summary description of the central location of the posterior distribution. However, when the posterior distribution is not symmetrical or unimodal, problems in using the DIC have been reported (Ntzoufras, 2009; Lawson, 2013).

In recent years several adjustments have been made to improve the DIC statistic including a proposed alternative measure of complexity and allowing for prediction (Spiegelhalter et al., 2014).

2.3.4 Residual analysis

The analysis of residuals, specifically the plotting of the residuals, play a vital part in the assessment of a model's goodness of fit (Lawson, 2013; Dunn and Smyth, 1996).

There are many types of residuals available, the choice of which will depend on the models applied. A few examples of methods include raw and deviance residuals, Pearson residuals, raw and scaled residuals for Poisson point processes (Baddeley et al., 2005) and a variety of Bayesian residuals (Carlin and Louis, 2000). In Chapter 4 we use the squared Pearson

residual method to identify outliers and assess goodness of fit in a model for areal counts, therefore we will focus on this method of residual analysis.

Pearson residuals

Detection of outliers and assessment of model fit, by the Pearson residual, occurs via the comparison of the observed values and those predicted by the model. The statistic provides evidence of a models lack-of-fit by taking large values (Hosmer et al., 1997).

The Pearson residual r is defined by

$$r = \frac{(O - E)}{\sqrt{E}}$$

where O is the observed and E is the expected number of cases in some area (and for some given time interval) based on the model in question. If the aim is simply to identify the worst fitting data, irrespective of whether the fitted model overestimates or underestimates the observed count, then the squared Pearson residual r^2 can be used. Areas where the model fails to fit are represented by large values. To quantify what qualifies as a large value, we can use the fact that r^2 has an approximate chi-squared distribution. Squared residuals r^2 exceeding high quantiles (e.g. the 0.95 quantile) of this distribution suggest a poor fit. The reason behind this poor fit can be the inadequacy of the model or the presence of outliers.

Exceedance probabilities

The use of exceedance probabilities are beneficial in assessing the spatial behaviour of a model as well as unusual clusters or disease aggregation (Lawson, 2013). Diggle (2005) argues that the mapping of predictive exceedance probabilities is more relevant than mapping predictive estimates, as the latter is highly variable, which can result in possible over-interpretation. A basic example of an exceedance probability can be expressed as $q_i^c = Pr(\theta_i > c)$ for the relative risk θ_i in some area. This probability is an estimate of how frequently the relative risk exceeds the null risk value, $\theta_i = c$, and can be regarded as an indicator of 'how unusual' the risk is in that unit (Lawson, 2013).

The general focus when using exceedance probabilities, has been detecting areas that have a high probability of being in the upper tail, suggesting a possible heightened risk. The probability of exceeding an upper threshold can also help in decision making. For example Diggle et al. (2007) use exceedance probabilities to determine areas where the relative risk exceeds a policy intervention threshold, therefore highlighting areas in which authorities need

to intervene.

As we discuss in Chapter 4, exceedance probabilities can also be used to look for areas that have a high probability of being in the lower tail, suggesting less cases then expected. Identification of such areas may reflect under-reporting of disease, a problem that is quite common in veterinary epidemiology, particularly in counties with less well developed veterinary health and surveillance systems.

Chapter 3

Disease epidemiology and datasets

3.1 Foot and Mouth Disease

Foot and mouth disease (FMD) is a virus that can be transmitted by direct and indirect animal contact as well as airborne spread. It can affect all types of cloven-hoofed animals, principally affecting cattle, sheep and pigs (Bachrach, 1968). The virus causes fever and affects the epithelial tissue causing vesicular lesions on hard wearing body parts such as the mouth, snout, feet and occasionally teats. The lesions normally cause a loss of appetite, lameness, reduction in productivity, with a fatality in approximately 2% of the adult cases and 20% for neonates (Madin, 2011).

The incubation period for the disease can range from 2 to 14 days from infection to clinical signs, and within this time the animals can become infectious to others. The susceptibility of animals to the disease by different transmission methods varies with species, with cattle being more susceptible to infection via inhalation and less likely to become infected by ingestion. The survival of the disease in different forms of transmission varies with climate. The virus can survive for long periods of time in low temperature with a neutral pH and only short periods in regions of high temperature (Madin, 2011; Davies, 2002).

The classification of FMD falls into seven known serotypes with multiple subtypes. The serotypes are A, O, C, Asia-1, and South African Territories (SAT) 1, 2, and 3 where the subtypes are due to virus mutation (Davies, 2002; Grubman, 2004).

FMD has been acknowledged as the most significant restriction to international trade in animals and animal products (Grubman, 2004). It is listed on the World Organisation for Animal Health (OIE) A list. This means that it is considered to have the potential to spread

rapidly with great public health or socio-economic consequences (Davies, 2002).

The first written record of cases of FMD are estimated to have occurred in 1514. Since the first outbreak, FMD has been recorded in all livestock inhabited regions of the world except for New Zealand. Since the start of the 20th century, the risk of re-emergence of FMD in many countries is of grave concern. Because of this, institutes to investigate methods of disease control have been established (Grubman, 2004).

3.1.1 Control methods

Due to FMD having a high rate of mutation, laboratory diagnosis can be difficult. This also causes complications in vaccine creation.

The methods applied to control FMD are usually based around three procedures; movement bans, vaccination, and depopulation. In regions where the disease is endemic a control strategy of vaccination is usually applied (Haydon et al., 2004), while in regions that usually have a status of 'FMD free' more drastic measures are generally applied.

The first step in any control method is the confirmation of diagnosis. This is crucial as other diseases, such as swine vesicular disease (SVD), vesicular stomatitis, and vesicular exanthema in swine and cattle, can cause similar symptoms. The confirmation of diagnosis is currently done via antigen capture enzyme-linked imunosorbent assay (ELISA). A positive result can be confirmed by the laboratory within 3-4 hours of samples received. However, a negative response can take up to four days to confirm as cultures must be carried out. Reverse transcription polymerase chain reaction (RT-PCR) can also be used for speedy detection of the disease but has less sensitivity and is more labour intensive (Grubman, 2004).

The time to detection usually will have an affect on the strategies applied. All programmes will usually require some sort of movement restrictions. This is normally paired with a campaign of slaughter and/or vaccination on all infected farms as well as those surrounding an infected premise (Davies, 2002).

An example of control methods applied to eradicate the disease is the 2001 outbreak in England. For this outbreak a national movement ban (NMB) for all susceptible animals was executed. For the eradication procedure a method of slaughtering, burning and burial of susceptible animals was chosen. This was applied to all infected farms as well as any farms in a 1.5km radius of an infected farm (pre-emptive culling) (Lawson and Zhou, 2005; Bessell et al., 2010).

3.1.2 Global impact

There are many global impacts that come along with FMD in a country. Whether it is the economic costs of eradication, the economic cost of export market restrictions, or the lasting impact that can result from animal culling.

An example of the economic impact that an outbreak brings is shown in the 2001 outbreak of FMD in England. This outbreak was estimated to have cost between US$12.3 billion and US$13.8 billion. This is made up of approximately 36% in lost tourism and US$4.2 billion in compensation payed out to the agriculture and food chain industry (Grubman, 2004).

Standards established by the OIE (Office Internationale des Epizooties) classify each country into one of three disease states. These are: FMD free without vaccination; FMD free with vaccination; and FMD present. The income from exporting animal products can be affected by a country's FMD status. Generally, countries with a status of FMD free without vaccination are able to reach higher levels of export pricing (Haydon et al., 2004).

The aim of trying to eradicate the disease can have dramatic effects long after the disease has been eliminated. An example of this occurred in the 2010 FMD outbreak in Japan. During the outbreak a seed bull farm contracted FMD and was later culled. This farm produced 90% of the Miyazaki prefecture sperm used in cattle production, therefore having a damaging affect on the livestock industry for at least 5 years after the initial event (Nishiura and Omori, 2010).

3.2 Viet Namese outbreak

3.2.1 Background

FMD is endemic in Viet Nam as well as the surrounding Southeast Asian countries: Laos, Cambodia, Thailand, Malaysia, Myanmar and the Philippines (Le et al., 2010). In Viet Nam three (out of the seven known) serotypes have been recorded (O, A, Asia-1). Serotype O is the most common, and is predominantly observed in cattle, buffalo and pigs (Cocks et al., 2009). FMD has been recorded in Viet Nam since 1999 when immunological and molecular methods were applied for the confirmation of FMD (Le et al., 2010). Originally documented as mainly occurring in the south of Viet Nam, it is now widespread (Sharma and Baldock, 1999).

Viet Nam has a largely agriculturally based economy and social structure, hence, livestock

are of great importance. FMD in Viet Nam greatly affects many levels of agriculture. These include the direct impact on livestock, a decrease in yield due to loss of productivity, the loss of stock via death (more common in the young and old) and the economic cost of treatment. The presence of FMD can also have indirect effects in other areas as cattle and buffalo are used as draught power (Cocks et al., 2009).

There is a significant amount of trade in cattle and buffalo between Viet Nam and Laos, Thailand, and People's Republic of China (Gleeson, 2002). By no means all of this is regulated, with unmonitored movement of cattle across borders quite prevalent. This may have an important impact on the spatial epidemiology of FMD in Viet Nam.

3.2.2 Description of data from 2006–2008

Our dataset, available from the Viet Namese Department of Animal Health, consists of 2734 reported cases of FMD in buffalo, cattle and pigs for Viet Nam over the time period 1st January 2006–31st December 2008. For every reported case we have an estimated date of onset, geographical area (described below), and the serotype of the outbreak. Once an animal in a herd was confirmed to have FMD all other animals in the herd were assumed to also be infected. Any other cases from the same herd during an outbreak were assumed to be the same strain as the first case. All three endemic serotypes (O, A, and Asia-1) were observed, although the number of occurrences of type A was low in comparison to O and Asia-1.

Viet Nam is divided administratively into provinces, and then further subdivided into counties, communes and districts. There are 64 provinces, with an average size of just over 5000km^2, and 11052 communes with an average area close to 30km^2. These geographical divisions are illustrated in Figure 3.1, which displays the provincial structure of Viet Nam (left-hand panel); a zoomed section thereof (middle panel); and the division of a single province into communes (right-hand panel).

Spatially, the FMD cases are identified by commune. As an illustration, Figure 3.2 shows the geographical pattern of disease for each of the years 2006, 2007 and 2008. Specifically, each point represents the centroid of a commune in which at least one case of FMD in cattle was observed during the year in question.

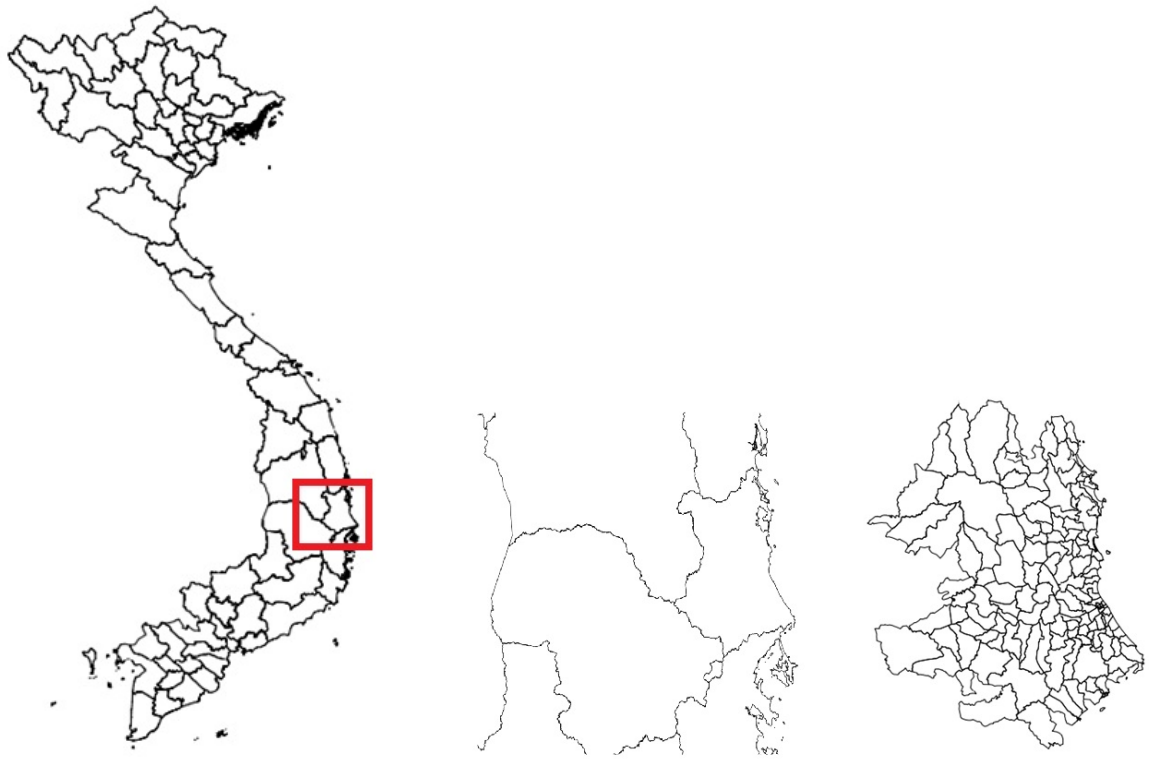


Figure 3.1: Spatial aggregation of Viet Nam. Left: province, middle: zoomed section by province, right: communes.

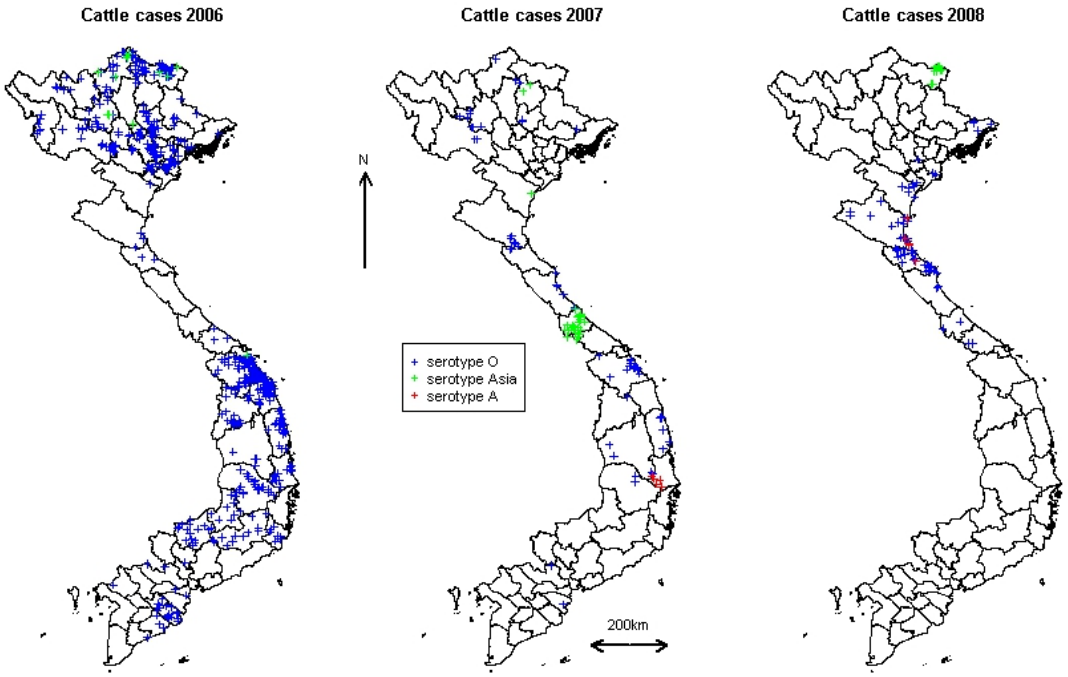


Figure 3.2: Points represent the centroid of communes in which FMD is observed in cattle for each of the years 2006, 2007, and 2008. Blue represents serotype O, green serotype Asia-1 and red serotype A.

3.3 English outbreak

3.3.1 Background

The 2001 outbreak in England marked the first major outbreak since 1968 in the United Kingdom (Sobrino and Domingo, 2001). This was followed by another outbreak in southern England in 2007 (Ellis-Iversen et al., 2011). The 2001 epidemic was diagnosed in Essex and was traced back to Northumberland. The epidemic infected over 2000 premises and lasted 241 days before eradication (Lawson and Zhou, 2005). The 2007 outbreak occurred in Surrey and only infected eight premises before its rapid and effective elimination (Ellis-Iversen et al., 2011).

3.3.2 Description of the data

The outbreak

On the 19th February 2001 in an abattoir in Essex, FMD was diagnosed in pigs. The following day, it was identified as the sub-type O Pan Asiatic strain of the virus. Veterinarians were able to trace the outbreak back to a pig farm in the village of Haddon on the Wall, in Northumberland in the north of England. Sheep farms surrounding the pig farm became infected, and due to the movement of sheep around the UK, the virus spread rapidly. On the 23rd of February a NMB for all susceptible species was implemented, after the transmission became spatially localised (50% of new infected premises (IPs) were within 3KM of an infectious IP and around 80% were within 10km). The peak of the outbreak occurred 33 days after it was discovered and by the 3rd September 2001, the outbreak had infected 2000 premises. At the time of the UK epidemic, FMD outbreaks also occurred in other EU countries: one IP in Ireland, two IPs in France and 26 IPs in the Netherlands. To eradicate FMD a strategy of slaughtering, burning and burial of all FMD-susceptible livestock on IPs within 24 hours, and on farms within a 1.5km radius of an IP within 48 hours, was adopted by the UK regulatory authority (Lawson and Zhou, 2005; Bessell et al., 2010).

The data

The dataset available for the 2001 epidemic in England contained information on the status of each farm within central Northern England, on each day of the 241 day outbreak (145 day epidemic), illustrated in Figure 3.3. A farm is given a status 1 if it was diagnosed with FMD, and the day of diagnosis recorded. If a farm has not been diagnosed it is given status 0. For farms with a status 0 but a date of onset less then 145 means that these were pre-emptively culled for preventative purposes. The dataset also contained descriptive variables: type of farm (cattle, sheep, goat or mixed) and the total number of susceptible animals on each farm. For every farm, the centroid coordinates were also known, enabling point process modelling.

The outbreak is shown in Figure 3.3, where each point represents the centroid of a farm, with blue farms where no disease and no culling occurred, green farms where no disease was found but culling occurred, and red farms where cases occurred.

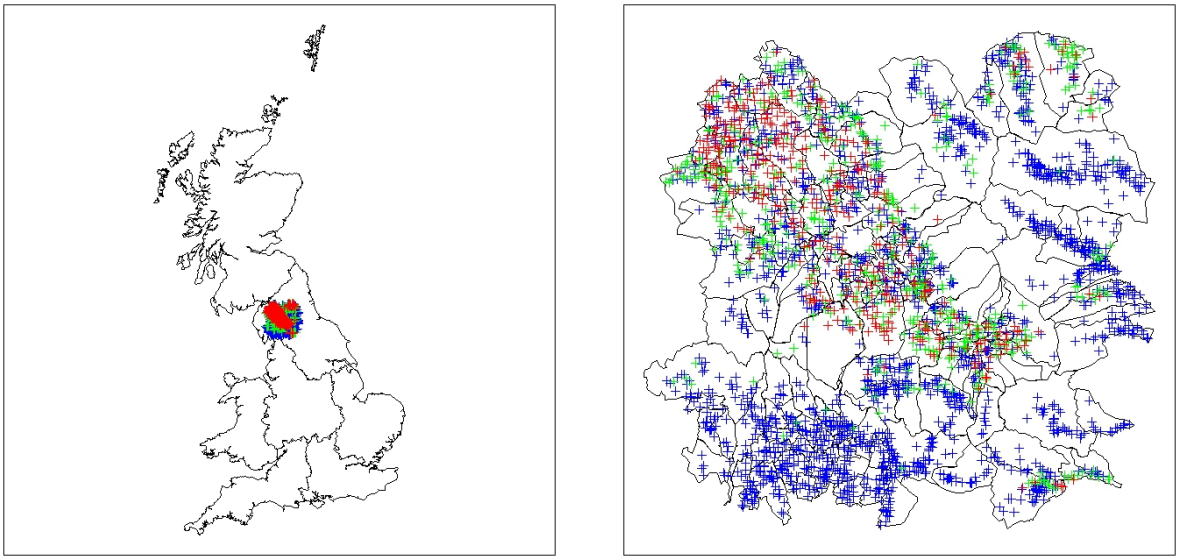


Figure 3.3: Point process plot of farms in the UK epidemic of FMD. Blue indicates farms where no disease and no culling occurred, green are the farms where no disease was found but culling occurred, and red are the farms where cases occurred.

3.3.3 Survey of models and analysis

During the 2001 epidemic in England various modelling techniques were applied. These include: binomial likelihood, survival modelling, mixed effects logistic regression, epidemiological network, stochastic models, Bayesian spatial SIR models, as well as models to look at geographic and topographic determinants.

Lawson and Zhou (2005) applied a binomial likelihood model to the epidemic. They ignored the culling that took place in preference to provide a simple descriptive model of the disease progression. The model seeks to represent the disease status of each farm (either infected or FMD free) in terms of geographical location. Specifically, their model is defined by equation 3.1.

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 x_i + \beta_2 y_i + \beta_3 x_i y_i + v_i + \xi_j \tag{3.1}$$

where p_{ij} is the probability of infection for the i^{th} farm in the j^{th} time period. The variables x_i and y_i represent respectively the easting and northing coordinates of the centroid of the farm, $\beta_1 x_i + \beta_2 y_i + \beta_3 x_i y_i$ a fixed spatial trend, v_i is a random uncorrelated heterogeneity term and ξ_j a temporal trend term. This model showed that there appeared to be a significant negative east to west trend, but no significant north to south trend was found. The interaction term in the model was also found to be significant, indicating a non-linear spatial trend. Although this model identified possible patterns in the spread of the outbreak, not including the effect

of culling that occurred on all IPs was a major limitation.

An alternative method proposed by Lawson and Zhou (2005) is the survival-marked process model. This is a form of survival analysis where in the model they considered the individual farm data with the date of infection as a temporal endpoint. They tried a variety of models, varying in complexity, and settled on the model shown in equation 3.2. The final model was chosen based on the deviance information criteria (DIC).

$$\begin{aligned} y_i &\sim \mathsf{Poisson}(\lambda\{d_i\}) \\ \log(\lambda_i) &= \alpha_0 + \alpha_1 d_i + \delta_i \\ d_i &\sim \mathsf{Weib}(r, \mu_i) \\ \log(\mu_i) &= \beta_0 + \beta_1 x_{ci} + \beta_2 y_{ci} + \beta_3 x_{ci} y_{ci} + v_i \end{aligned} \tag{3.2}$$

In equation 3.2, y_i is the count of disease cases within the i^{th} farm and d_i is the time until disease for the i^{th} farm. The centroid coordinates for farm location are x_{ci}, y_{ci} and δ_i as the farm type. In this model v_i is an uncorrelated spatial frailty term. During the 2001 epidemic all farms within 1.5km radius of an infected farm were pre-emptively culled. This led to a loss of information for these farms as exact information on infection status and time are unknown. This censoring was included in the model by conventional survival analysis methods. The results of this analysis showed that farm type (δ_i) was significant, particulary the presence of cattle on a farm. The identified limitations of the model include the presence of large amounts of censored observations due to the control methods, a lack of a complex dependence for survival or a method of taking into account the number of cases given the number of cases that have already occurred.

A mixed effects logistic regression was applied by Bessell et al. (2010) to a refined dataset of cases and controls (infected and non-infected, respectively). They excluded all farms that contained only pigs, as these had minimal involvement in the outbreak after the initial cases. They followed the methods of Hosmer and Lemeshow (2000) where the variables are first applied in univariate logistic regression models, to test significance, then significant variables were included in a multivariate regression model. They looked at predictor variables that fell into four main categories:

1. The distance to a seed IP (the Euclidian distance to the nearest farm that was infected before the non-movement ban)
2. Regional data (county that the farm belongs to)
3. Farm data (the type of livestock on the farm)
4. The farm's neighbourhood (the number of cattle and sheep on the neighbouring farms

within 10km radius).

The variables that had a significance level less than 0.001 where included in the multivariate generalised linear mixed model (GLMM). Once again these variables were tested for significance taking into consideration all plausible interactions. From these results, all variables included in the multivariate analysis except sheep density were found to be significant, hence, sheep density was removed from the final model. For the final model they applied transformations to distance to seed, farms's neighbourhood and the farm area to linearize the effects of these variables on the logit scale. They tested autocorrelation and included random effects by applying a grid of hexagons over the susceptible population. The results of this analysis found that the epidemic spread differently in different parts of the UK and that the distance to an infectious seed as well as cattle density are major risk factors for a farm contracting the disease. There is also an increase in risk with an increase in farm land size. One of the limitations expressed by the authors was that the animal density (number of livestock on the holding) was not included as a predictor.

The use of case-control methodology has also been used in the analysis of geographic and topographic determinants in the spread of FMD, this is discussed by Bessell et al. (2008). Here they mainly focused on disease transmission during the period of local spread that followed the non-movement ban (NMB). They looked into the effect that access to farm, road distance between farms, the presence of a forest, presence of intervening roads, the elevation change and the presence of linear features such as rivers and railways have on the spread of the disease. They found that the presence of features such as rivers and railways acted as barriers for transmission.

3.4 Japan

3.4.1 Background

Foot-and-mouth disease is not endemic in Japan. Before an outbreak in the spring (March) of 2000, Japan had been free from the disease since 1908. The epidemic affected four farms and was eradicated through movement control, surveillance and culling. Japan was able to regain its FMD free status by the end of September 2000 (Sugiura et al., 2001). In April 2010 a major epidemic occurred, the first in a decade, where 292 infected farms were detected (Hayama et al., 2012). Both epidemics where caused by serotype O virus (Nishiura and Omori, 2010).

3.4.2 Description of the data

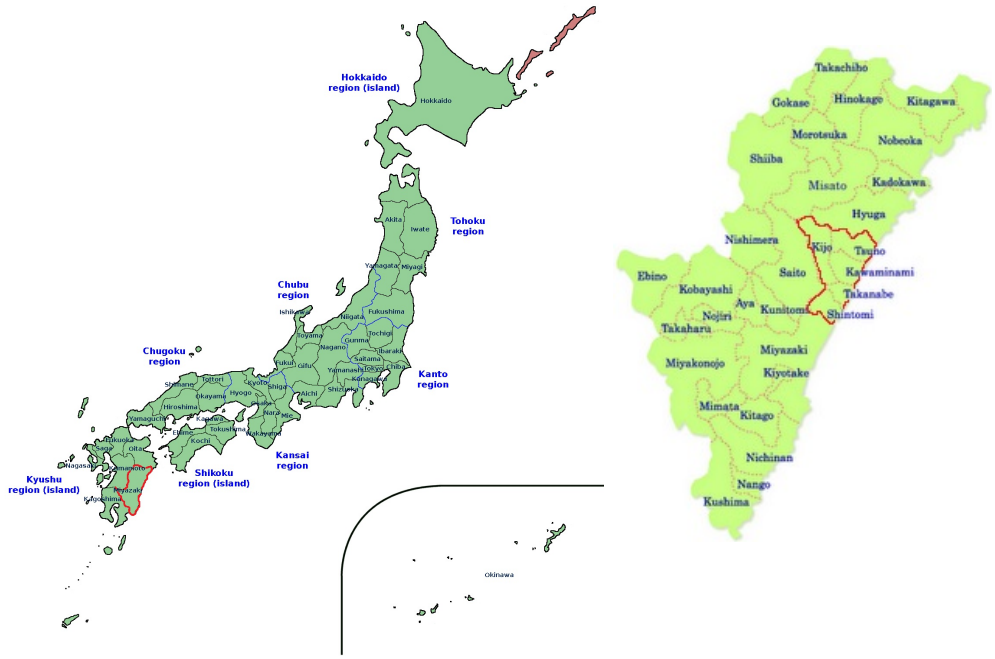
The outbreak

The 2010 epidemic of FMD in Japan was first spotted in water buffalo in late March. On the 31st March, in the town of Tsuno, a farmer reported to the local veterinary practitioner that one of his buffalo had a reduced milk yield as well as a fever. The veterinary service then passed this information onto the Miyazaki Prefectural Livestock Hygiene Service Centre (LHSC). When the officer attended the farm, four buffalo were identified with fever and/or diarrhoea. As these symptoms were not strongly suggestive of FMD it was though that the buffalo had a non-specific diarrhoea disease (however on the 23rd April these were later confirmed to be FMD positive). On the 9th April in a fattening farm, near the location of the water buffalo, the LHSC was notified about a cow having oral ulcers. As only one cow was showing symptoms it was thought to be Ibaraki disease. However on the 16th April the LHSC was notified of two more cows showing symptoms on the same fattening farm so samples were taken for further investigation. On the 20th April, day 20, FMD was confirmed and preventative measures were put into practice. These were established in accordance with the 'Specific Domestic Animal Infectious Disease Quarantine Guidelines' and included culling, movement restriction, surveillance and disinfection (for more details see Nishiura and Omori (2010)). On the 18th May (day 48) a state of emergency was declared. People (farmers and non-farmers) were asked to cooperate with disinfection procedures put in place and all mass gathering or events were cancelled or postponed. A method of ring vaccination over a 10km radius was implemented on the 22nd May and by the 30th June vaccination on all 'at risk' farms was completed (Nishiura and Omori, 2010).

This epidemic lasted from late March until the 10th July 2010 and involved 292 farms being confirmed to be FMD positive. It resulted in nearly 290,000 susceptible animals being culled (Knowles et al., 2012; Hayama et al., 2012).

The data

The dataset available for the 2010 epidemic in Miyazaki, Japan, contained information on the centroid coordinates for all infected farms, the date of outbreak, the date of inspection as well as the date of slaughter. We also have information on whether the farm contained beef or dairy cattle, pigs, goats or sheep, and the population of each present. The prefecture structure of Japan is shown in Figure 3.4 and the spatial distribution of data is shown in Figure 3.5, as in previous figures the point represents the centroid location of the farm.



(a) Japan (b) Miyazaki

Figure 3.4: (a) Map of Japan showing prefectures with the Miyazaki prefecture outlined in red (Wikimedia.org, 2015). (b) Miyazaki prefecture showing regions with our spatial window outlined in red (blogspot.co.nz, 2010).

3.4.3 Survey of models and analysis

After the 2010 epidemic of FMD in Japan, Hayama et al. (2012) used univariate and multivariate methods to examine possible risk factors. Similar to the regression analysis for the 2001 FMD outbreak in the UK, possible risk factors were assessed by a univariate analysis, with significant factors then included in the multivariate model. The risk factors that were investigated are: the susceptible species present on the farm, the herd size, the number of days from the reported date at source farm to the clinical signs, the number of days from clinical onset to the completion of culling at source farm, the Euclidean distance between a source farm and a neighbouring farm, and the amount of time a neighbouring farm was downwind from a source farm. In the univariate analysis two variables were found to be significant, species present on the farm and herd size. These were then applied to a multivariate logistic regression model. The results found that large scale pig farms posed the greatest risk to the surrounding farms, and medium to large scale cattle farms had a higher risk of becoming infected by local spread. The findings are consistent with the earlier findings of Nishiura and Omori (2010).

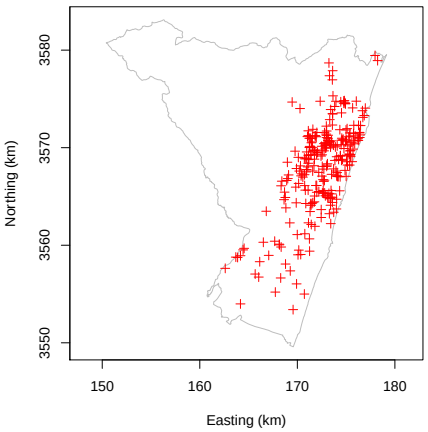


Figure 3.5: Miyazaki 2010 FMD outbreak spatial distribution

3.5 Interspread

3.5.1 Description of Interspread

InterSpread is a software package that uses large, complex and very flexible stochastic simulation models to predict the spread of an infectious disease. The software has the ability to include the influence of many different factors on the spread of infection (Keeling, 2005). The original InterSpread was designed to provided a framework for modelling the spread of infectious diseases (Stevenson et al., 2013). It was created based on the research of R.S Morris and coworkers, particulary the Ph.D. work of Sanson (1993) (Keeling, 2005). The 2001 epidemic of foot and mouth disease in England provided the opportunity for InterSpread to be used as a support tool during an outbreak. After this a revised version of InterSpread was created, called 'InterSpread Plus' (Stevenson et al., 2013).

To create a model in InterSpread one must specify data files including information on the farms, contact locations, epidemic history, and the spatial window of the study area. The disease spread parameters such as the number of simulations, infectivity period and a transmission method also needs to be specified. These parameters can then be specified to create outbreaks mimicing certain diseases. A major benefit of the use of InterSpread is the ability to specify details on a variety of other parameters, including intervention methods such as movement restrictions, vaccination and depopulation (Stevenson et al., 2013). This therefore, allows users to look at the effects that different intervention strategies have on the progression of an outbreak.

Stevenson et al. (2013) provide a helpful flowchart of how InterspreadPlus simulates outbreaks for a given disease (Figure 3.6).

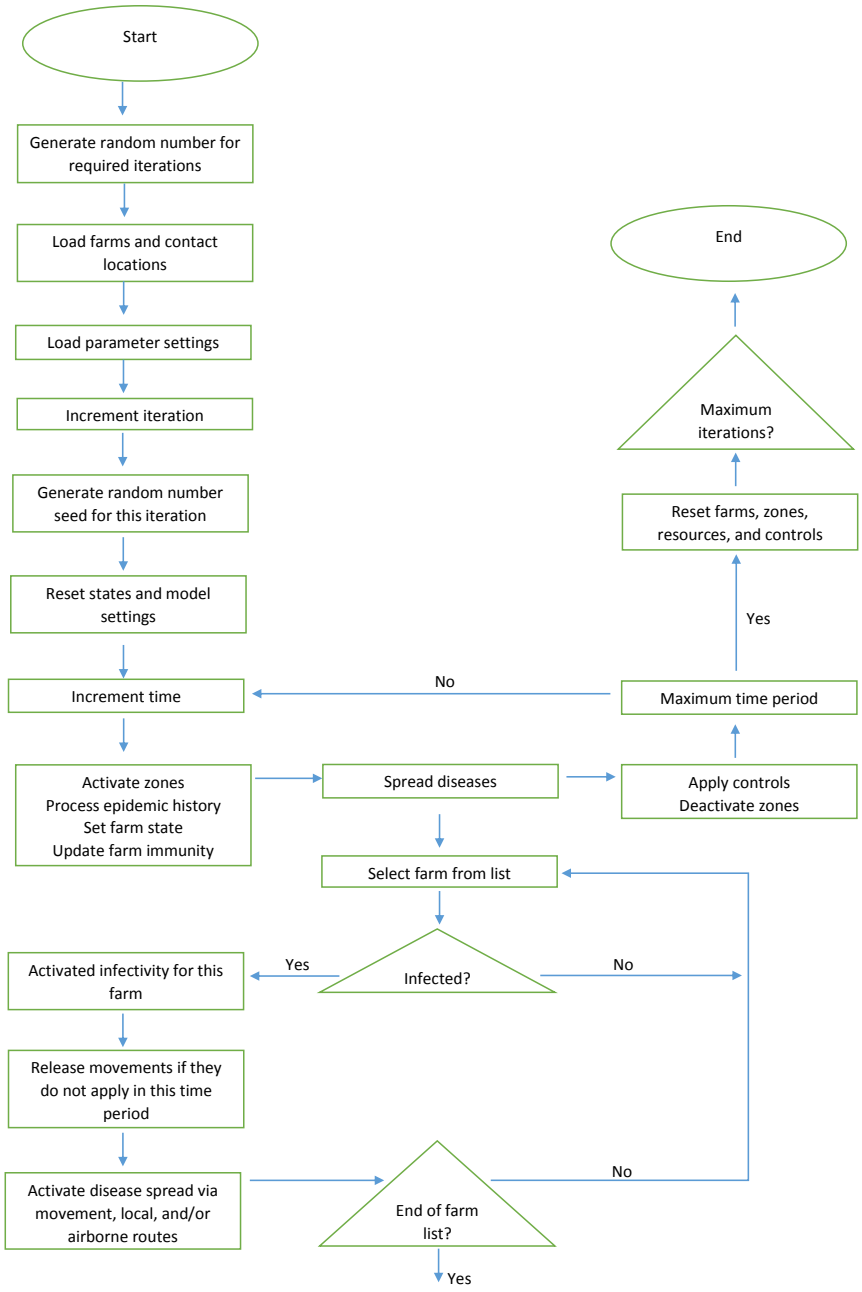


Figure 3.6: Overall simulation flowchart in InterspreadPlus , reproduced from Stevenson et al. (2013).

There are many examples of the use and beenfits of InterspreadPlus. As mentioned previously, during the 2001 UK epidemic of FMD InterSpread was used to be predict the spread of

infection (Keeling, 2005). The information obtained, on how local transmission occurred, from the analysis of this outbreak was then used to develop an InterSpread Plus model to investigate how FMD could possibly spread if introduced in New Zealand (Sanson et al., 2006). As well as predicting how an outbreak may spread within a country, it can also be used to investigate factors that influence the progression of a outbreak for an infectious disease. For example InterSpread Plus was used by Bokund et al. (2012) to investigate the influence on livestock markets on the spread of FMD. They found that animal markets have an effect on the size, duration and cost of the epidemic.

3.5.2 Use of Interspread to simulate our FMD outbreaks

Great Britian simulation

For our work in chapter 7 we investigate the effects of different intervention strategies applied to simulated outbreaks of FMD in Northern England and Southern Scotland. The outbreaks were created through the use of Interspread, with a spatial window covering the counties of Cumbria, Northumberland, Dumfries and Galloway and Scottish borders (all counties of Great Britain and the four selected shown in Fig 3.7). From now on we will refer to our four selected counties as the 'border counties' with the resulting spatial window shown in Figure 3.8.

Our epidemics had a maximum length of 100 days, with the majority of the simulated datasets lasting a lot less than this due to the intervention methods applied. The animals present on the farm could be cattle, deer, goats, pigs, sheep, or a mixture of these. For each farm we know the population of each species and the total farm population, as well as the centroid coordinates. The beginning of the outbreak for all simulations was fixed, with four initial farms infected on days 1,5,6,6 which were later detected on days 15,16,17,15, respectively. Based on this epidemic history one of five intervention strategies (no intervention, vaccination, depopulation, vaccination and depopulation, and depopulation and preemptive culling) was applied and a dataset created. The spatial distribution of the dataset for each intervention strategy is shown in Figure 3.9. We see the infected farms in red, farms where FMD was not detected but depopulation occurred in green, farms where FMD was not detected but vaccination occurred in blue, and farms with no intervention and no disease detected in grey. The application of intervention strategies will be explained in more detail in chapter 7.

(a) Great Britain (b) Selected counties

Figure 3.7: Map of Great Britian showing counties (left) and the four selected counties (right) (Baxter, 2014).

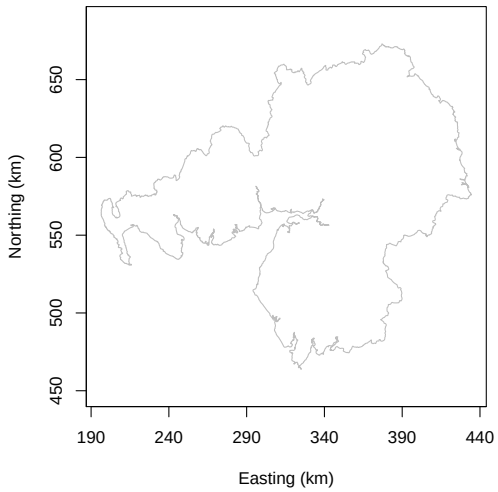
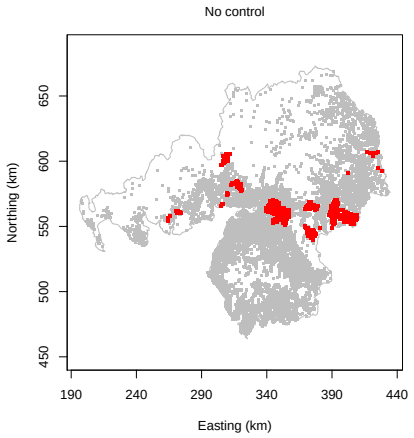


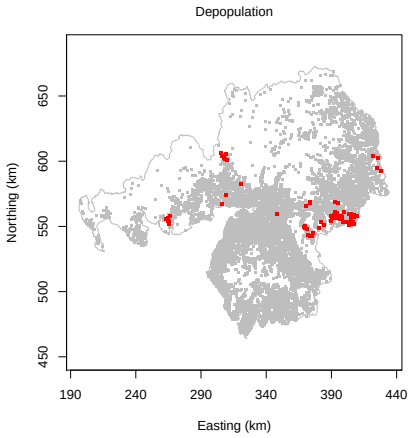
Figure 3.8: Border counties spatial window

Japanese simulation

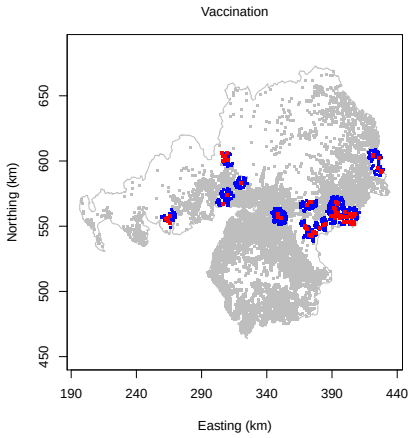
Similar to the Border counties simulated data sets, we created a series of outbreaks through the Interspread software. The spatial window was defined as Miyazaki region (Figure 3.10). In the creation of this simulated data we once again have a maximum length of 100 days, with the majority of the simulated datasets lasting a lot less then this due to the intervention methods applied. The animals present on the farm could be beef or dairy cattle, pigs, goats, wildpigs, sheep, or a mixture of these. For each farm we know the population of each species and the total farm population, as well as the centroid coordinates. The beginning of the outbreak for all simulations was fixed, with 13 initial farms infected on days 1,2,3,4,5,5,6,6,6,7,7,7,7 which were later detected on days 15,12,16,16,16,16,15,17,13,16,17,16,17 respectively. Based on this epidemic history one of five intervention strategies (no intervention, vaccination, depopulation, vaccination and depopulation, and depopulation and preemptive culling) was applied and a dataset created. These datasets are shown in Figure 3.11 where we can see all susceptible farms in grey and farms that have been simulated to be diseased based on a FMD epidemic in red. We see the un-diseased farms affected by the intervention methods, blue for vaccination and green for depopulation. When no intervention strategies are applied the epidemic is allowed to run it natural course, which involves almost all farms becoming infected.



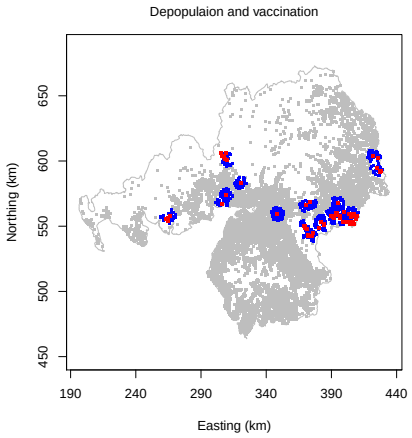
(a) No intervention



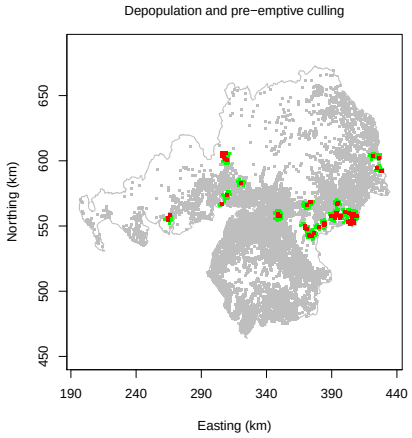
(b) Depopulation



(c) Vaccination



(d) Depopulation and vaccination



(e) Depopulation and pre-emptive culling

Figure 3.9: Image plots of simulated FMD outbreaks in Border counties : (a) No intervention (b) depopulation (c) vaccination (d) depopulation and vaccination (e) depopulation and pre-emptive culling. Grey represents farms with no FMD and no intervention, blue where no FMD but vaccination applied, green where no FMD but depopulation occurred and red where FMD is present.

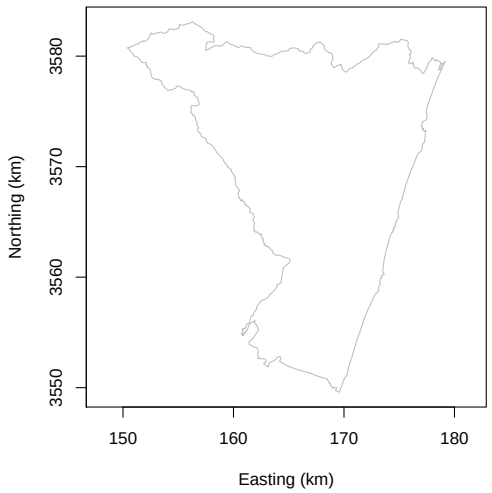
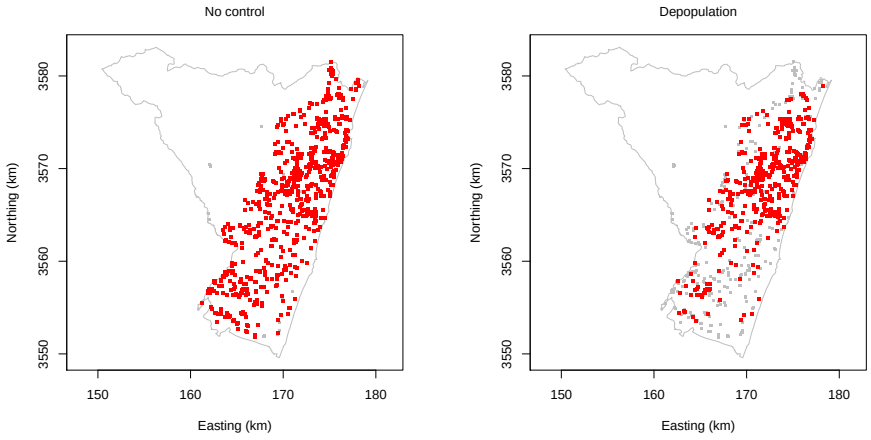
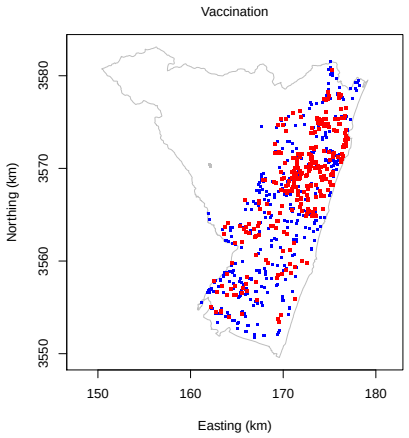


Figure 3.10: Miyazaki spatial window

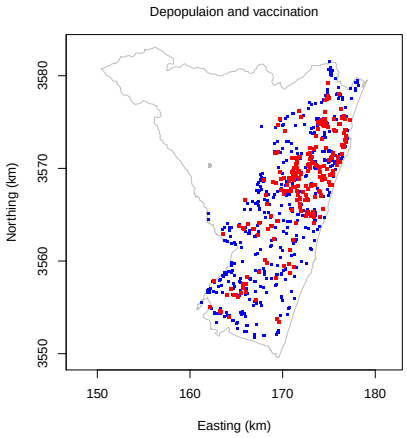


(a) No intervention

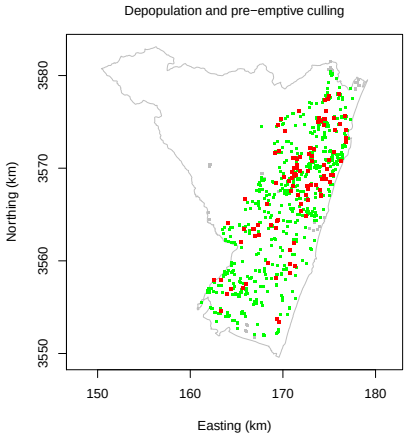
(b) Depopulation



(c) Vaccination



(d) Depopulation and vaccination



(e) Depopulation and pre-emptive culling

Figure 3.11: Image plots of simulated FMD outbreaks in Miyazaki (Grey all farms, Red FMD positive farms) : (a) No intervention (b) depopulation (c) vaccination (d) depopulation and vaccination (e) depopulation and pre-emptive culling Grey represents farms with no FMD and no intervention, blue where no FMD but vaccination applied, green where no FMD but depopulation occurred and red where FMD is present.

Chapter 4

Exceedence probabilities for detecting anomalies in animal health data.

4.1 Introduction

Advances in technology have provided better opportunities for the recording, processing and storing of disease event information. This, in theory, should provide the necessary information needed to carry out detailed analyses of the factors influencing the spatio-temporal distribution of disease in animal populations. However, the reliability of such analyses depends on the quality of the data. Anomalies in data records have the potential to introduce significant bias in the results of descriptive analyses and fitted statistical models, skewing results and potentially leading to inappropriate decision making.

Generally the method used to identify anomalous data is by residual analysis (e.g. Gail, 1991). Where a large residual indicates a data point that is poorly predicted by the model in question, this therefore suggests either a problem with the point in question or significant inadequacies of the model as a whole (or both). For simple models, such as linear regressions with fixed effects, this type of analysis is straightforward to implement and interpret. However, residual analysis becomes more difficult as the complexity of the data types and models increase. For example, until quite recently there was no generally accepted definition of a residual for a spatial point process (Baddeley et al., 2005).

The use of hierarchical statistical models in spatio-temporal epidemiology present such a

problem. Such models seek to describe the variation in disease incidence or prevalence by incorporating random variables at multiple levels (see, for example, Lawson, 2013). This in principle means that we can define residuals at the corresponding levels of the hierarchy. It is often the case that the flexibility that these models provide can mask the presence of truly 'odd' records and, superficially at least, make the model appear to be fitting well.

In this chapter we aim to promote the use of *exceedance probabilities* as a tool for identifying and assessing anomalous data in spatio-temporal models for areal disease data. Exceedance probabilities can be applied at any level of the model to describe the extent to which an individual random term, or combination of random terms, is unusual, in the sense of lying in the extreme tails of the specified distribution. Exceedance probabilities have been used previously to detect anomalous clusters of cases in point process data (e.g. Diggle et al., 2005; Davies and Hazelton, 2013), and have also been used to identify regions with unusually high relative risk when modelling areal data (e.g. Best et al., 2005; Lawson, 2010). Our aim is to showcase the wider uses of exceedance probabilities, demonstrating their application to individual stochastic terms in areal models and also noting their usefulness for highlighting areas with unexpectedly low (as well as high) reported rates of disease.

We illustrate the use of exceedance probabilities as a diagnostic tool for anomalous data through a case study involving outbreaks of foot-and-mouth disease (FMD) in Viet Nam for the period 2006-2008.

The material within this chapter is based upon the paper "Using exceedance probabilities to detect anomalies in routinely recorded animal health data, with particular reference to foot-and-mouth disease in Viet Nam" (Richards et al., 2014).

4.2 Data

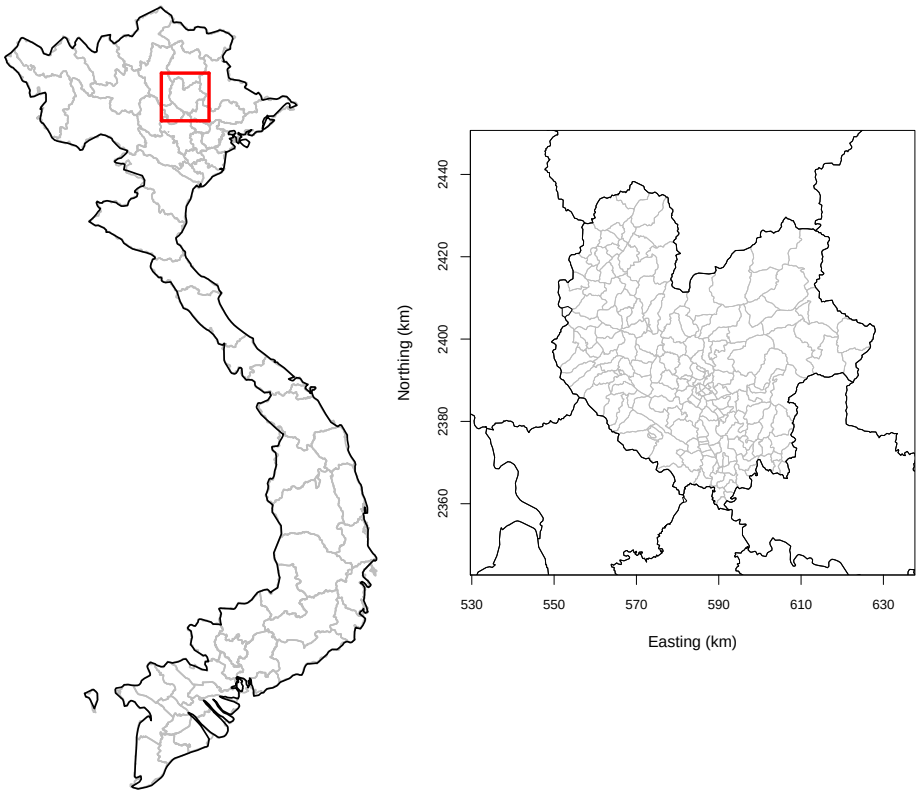
4.2.1 Viet Nam FMD endemic

The full dataset is explained in detail in section 3.2. For modelling purposes the dataset from the Vietnamese Department of Animal health was refined to only included cases that occurred in cattle.

The resulting dataset consisted of 2734 reported commune-level outbreaks of FMD for the time period 1 January 2006 to 31 December 2008. For every reported case we have an estimated date of onset, geographical area (described below), and the serotype of the outbreak. Any other cases from the same commune during the outbreak were assumed to be the same strain

as the first case. All three endemic serotypes (O, A, and Asia-1) were observed, although the number of occurrences of type A was low compared with serotypes O and Asia-1.

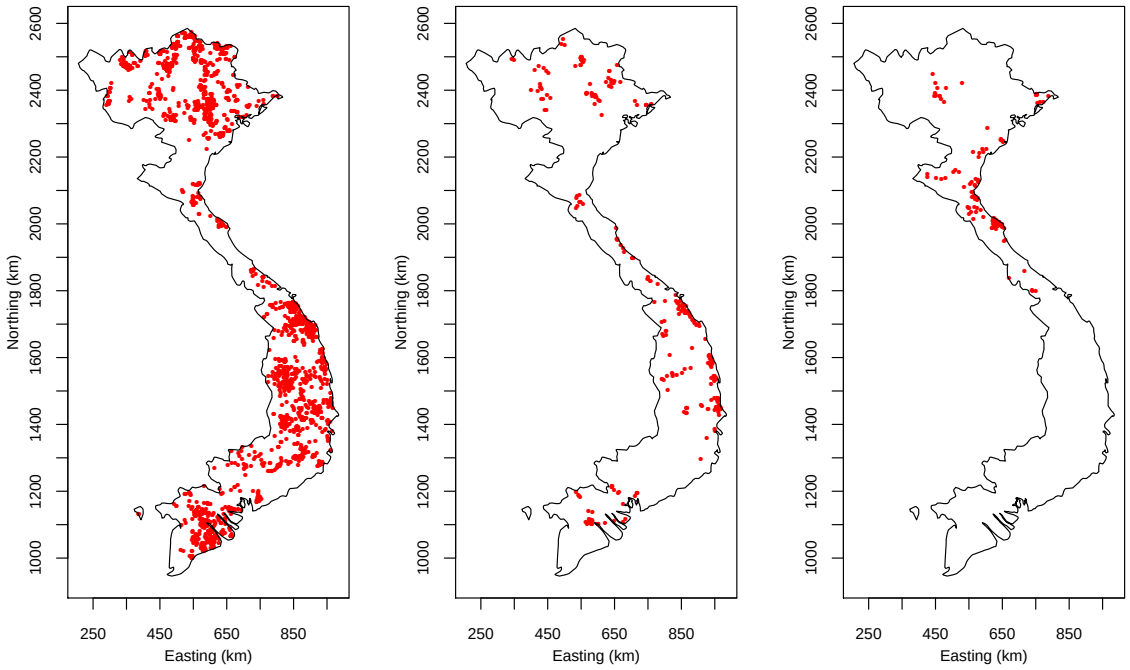
Viet Nam is divided administratively into provinces, and then further subdivided into districts and communes. There are 64 provinces, with an average size of just over 5000 km^2, and 11052 communes with an average area close to 30 km^2. These geographical divisions are shown in Figure 4.1, which displays the provincial structure of Viet Nam (Figure 4.1a) and the division of a single province into communes (Figure 4.1b).



(a) Provincial boundaries (b) Commune boundaries

Figure 4.1: Map of Viet Nam showing provincial (left) and commune boundaries (right). The location of Thai Nguyen province is indicated by the box on the provincial map. Commune boundaries for Thai Nguyen province are shown on the right.

For the following analyses, FMD cases were identified spatially by their commune location. As an illustration, Figure 4.2 shows the geographical pattern of disease for each of the years 2006, 2007 and 2008. Specifically, each point represents the centroid of a commune in which a least one case of FMD in cattle was observed during the respective year.



(a) 2006 (b) 2007 (c) 2008

Figure 4.2: Map of Viet Nam showing the point location of commune-level FMD outbreaks, for: (a) 2006, (b) 2007, and (c) 2008.

4.2.2 Explanatory variables

In addition to the disease data, we also have information on the covariates cattle density, enhanced vegetation index (EVI) and elevation (in metres).

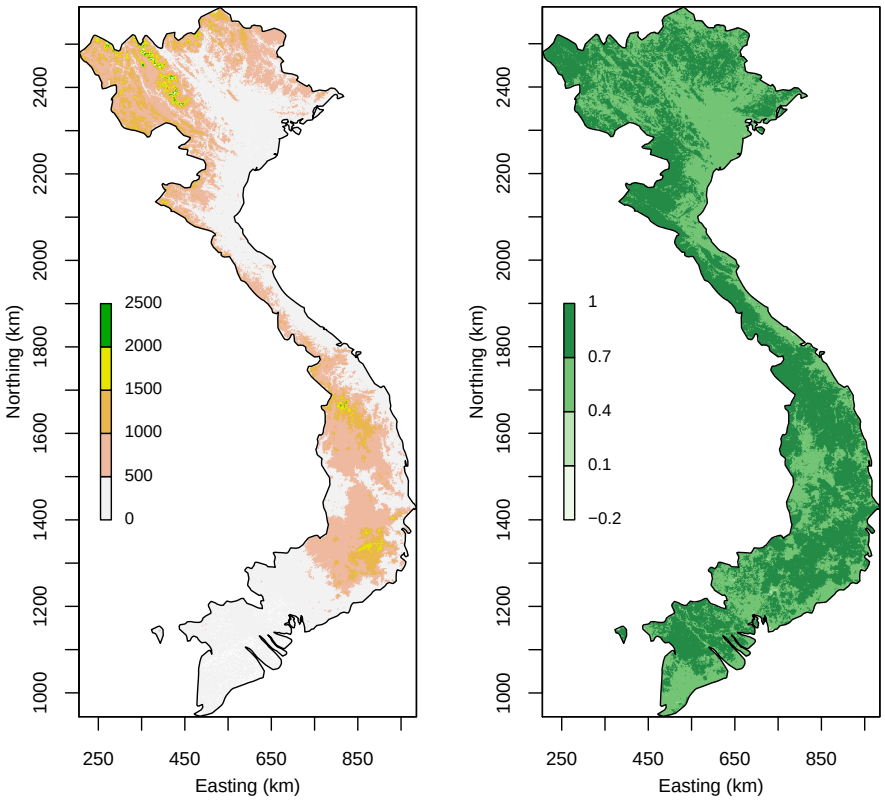
Cattle density is specified for each year at the provincial level only, and is measured as the number of cattle per square kilometre. It is based on annual mean cattle population estimates, routinely recorded by the Vietnamese Department of Animal Health.

Elevation data (Figure 4.3a) was obtained from the NASA Shuttle Radar Topographic Mission (SRTM) 90 metre Digital Elevation Database version 4.1 (CGIAR-CSI, 2013).

EVI is an optimised index of the greenness of the Earth's surface. We chose the use of EVI, over the use of Normalised Difference Vegetation index (NDVI), because it has an improved sensitivity in high biomass regions and better properties for monitoring vegetation cover through a de-coupling of the canopy background signal and a reduction in atmosphere influences. The EVI data (Figure 4.3b) were obtained as monthly time series with a 1km spatial resolution from MODIS data via the NASA Terra satellite. These were sourced from NASA's Land Processes Distributed Active Archive Centre. MODIS data are produced in the

sinusoidal projection (MODLAND Sinusoidal Grid) and made available as 460 tiles covering the Earth's surface, with each tile measuring 10° × 10°. For the analysis of Viet Nam we used the information obtained from six tiles, covering 8° and 24° N and 102° and 110° E. These were obtaied from MODIS Terra Vegetation Indices Monthly L3 Global 1 km SIN grid (MOD13A3, version 5).

To obtain our estimates for EVI and elevation, the data first had to be sorted, reprojected to longitude-latitude and converted to GeoTIFF format using the MODIS reprojection tool (USGS Earth Resources Observation and Science Center, 2011). This was then processed using the `raster` package (Hijmans and van Etten, 2013) in R (R Development Core Team, 2012) to provide estimates for EVI and elevation for each province of Viet Nam.



(a) Elevation (b) NDVI

Figure 4.3: Image plots of Viet Nam showing: (a) elevation (expressed in metres) and (b) Enhanced Vegetation Index for January 2006.

4.3 Modelling the Data

4.3.1 Model Building

The variables in our dataset were recorded with varying degrees of spatial and temporal resolution. While we have disease incidence data at the commune level on a daily basis, we only have annual values by province for the explanatory variable; cattle density. This latter variable represents the 'lowest common denominator' in terms of data resolution, constraining us to work at a province-by-year level of spatio-temporal aggregation.

Therefore, the response variable was defined to be the number of infected communes per province per year, where a commune was classed as infected for any given year if there was at least one reported case of FMD within that 12 month period. The explanatory variables are cattle density; average elevation per province; and annual minimum and maximum EVI values by province. These latter variables were computed by spatial averaging of the monthly EVI raster images over each province, and then selecting the smallest and largest monthly values. Our intention here was that these variables would represent (crude) predictors of FMD frequency (for example, susceptible animal populations at risk are more likely to be kept in areas of the country where minimum EVI is relatively high). All these covariates were centered and scaled to unit standard deviation.

For our initial model of FMD in cattle, we employed a binomial logistic regression model. Here the response y_{ik} is the number of infected communes in province i for year k and is defined as:

$$y_{ik} \sim \mathsf{Binom}(p_{ik}, n_i) \qquad i = 1, \ldots, 64; \quad k = 1, 2, 3; \tag{4.1}$$

where n_i is the number of communes in province i, and p_{ik} is the probability of an FMD infection in any given commune in that province during year k. The notation $y \sim \mathsf{Binom}(p, n)$ indicates that a random variable y follows a binomial distribution with probability parameter p and number of trials parameter n, in the usual way.

The probability p_{ik} that a commune was identified as FMD-positive was modelled on the logit scale. One model that we considered was as follows:

$$\text{logit}(p_{ik}) = \alpha_i + \gamma_k + \delta \mathsf{CD}_{ik} + \theta_1 \mathsf{El}_i + \theta_2 \mathsf{EVlo}_{ik} + \theta_3 \mathsf{EVhi}_{ik} \tag{4.2}$$

where $\text{logit}(p) = \log[p/(1-p)]$. The explanatory variables were defined as follows:

- $\mathsf{CD}_{\mathsf{ik}}$ is the cattle density in province i for year k;
- El_{i} is the average elevation in province i;
- $\mathsf{EVlo}_{\mathsf{ik}}$ is the minimum EVI value for province i in year k; and
- $\mathsf{EVhi}_{\mathsf{ik}}$ is the maximum EVI value for province i in year k.

The parameters α_i and γ_k describe the main effects of province and year respectively, while coefficient δ describes the effect of cattle density. The parameters $\theta_1, \theta_2, \theta_3$ control the effects of elevation, minimum EVI and maximum EVI respectively.

As expected, we found that model (4.2), and other fixed effects models, fitted the data very poorly. This is largely due to the amount of unexplained variation within the model. We therefore added a random effect term on the logit scale to account for this heterogeneity. The most complex model of this type that we considered was

$$\text{logit}(p_{ik}) = \alpha_i + \gamma_k + \delta \mathsf{CD}_{ik} + \theta_1 \mathsf{El}_i + \theta_2 \mathsf{EVlo}_{ik} + \theta_3 \mathsf{EVhi}_{ik} + \varepsilon_{ik} \tag{4.3}$$

where $\{\varepsilon_{ik}; i = 1, \ldots, 64,\ k = 1, 2, 3\}$ is a set of independent $N(0, \sigma^2)$ (i.e. normally distributed with zero mean and variance σ^2) random effects. Equation (4.3) describes an extremely flexible mixed effects logistic regression model, where the random effects allow for additional variation between the responses at every combination of province and year. We also considered more parsimonious models, for example where ε_{ik} in Equation 4.3 was replaced by ε_i (only province-specific random effects).

Variants on models (4.1) and (4.3) were also considered that included serotype effects. In those cases, the response becomes y_{ijk} for the count of infected communes in province i for serotype j in year k, with the logit of the corresponding commune infection probability, $\text{logit}(p_{ijk})$, incorporating a fixed effect for serotype. For the serotyope specific version of (4.3), the random effect ε_{ijk} is specified by combination of province, serotype and year.

The models defined above, for example by equations (4.1) and (4.3), are hierarchical, because they incorporate random variation on both the response and logit levels. Such models are conveniently fitted within the Bayesian paradigm using Markov chain Monte Carlo (MCMC) methods (e.g. Gilks et al., 1996). We implemented this approach using the WinBUGS software (Lunn et al., 2000). Cauchy priors were applied to fixed effect parameters following the scheme described by Gelman et al. (2008). A uniform prior on the interval $(0, 5)$ was applied to the random effect variance σ^2.

Convergence of the MCMC runs for each model was confirmed using standard diagnostics, such as trace plots and Gelman-Rubin convergence plots. The trace plots are a visual representation of the variable value against the iteration number. When convergence has occurred we expect to see constant fluctuations around a value. We can evaluate this for every variable in the model. As we used multiple (three) chains for our analysis, convergence can be evaluated through the use of the Gelman-Rubin convergence statistic. This is available as an in built function in WinBUGS based on Brooks and Gelman (1998). We used these methods to confirm convergence for our variables. This process was slow for our models but occurred for all variables. The posterior means, computed from post convergence iterations, were used as the point estimates for the parameters.

4.3.2 Preliminary Assessment of the Fitted Models

We compared models using the Deviance Information Criterion (DIC) proposed by Spiegelhalter et al. (2002*b*). DIC is described in more detail in section 2.3.3. The DIC calculates goodness of fit while penalising for complexity, with the 'best' model the one with the lowest DIC value.

The DIC statistics for our model variants provided strong support for the most flexible model as defined by Equation (4.3).

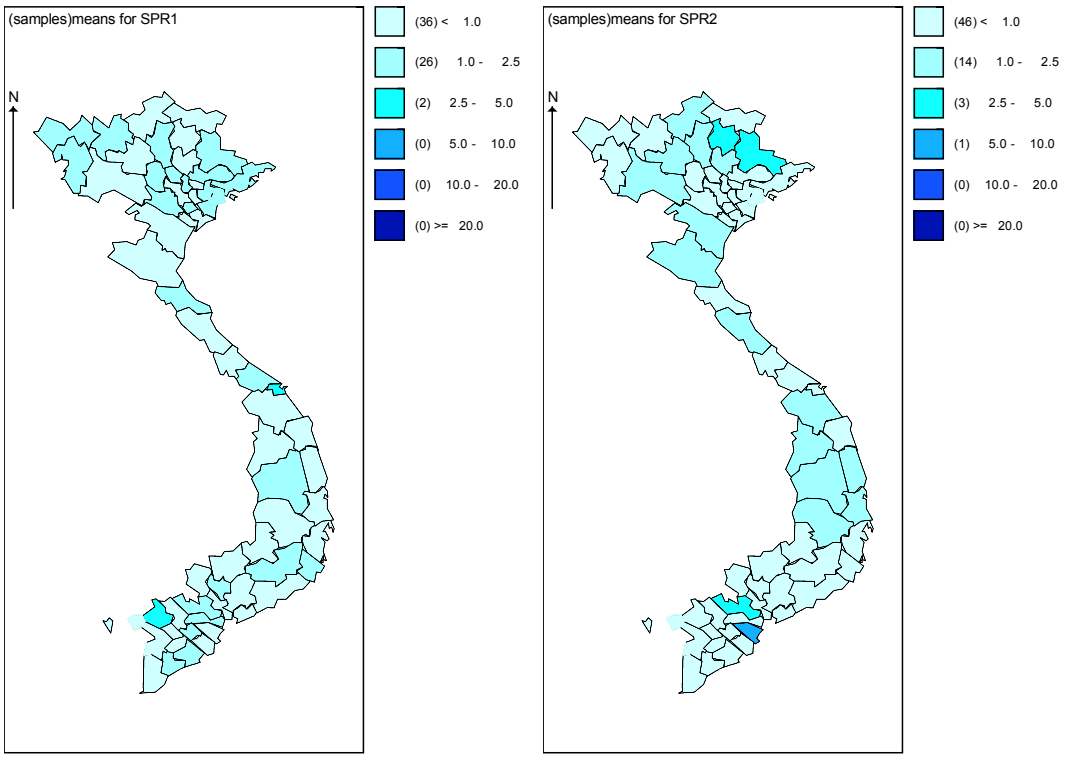
A standard residual analysis for this type of fitted model is to examine the Pearson residuals. The Pearson residual for province i and year k is given by

$$r_{ik} = \frac{y_{ik} - n_i \hat{p}_{ik}}{\sqrt{n_i \hat{p}_{ik} + 0.01}} \tag{4.4}$$

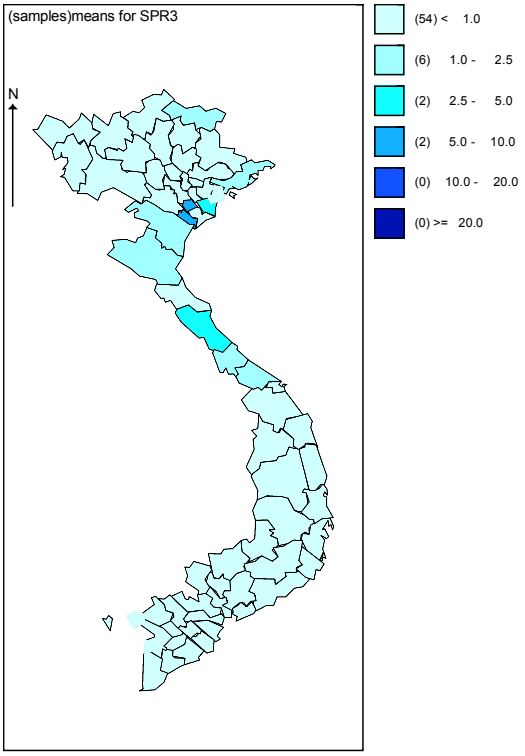
where $\hat{p}_{ik}$ is the posterior mean estimate of p_{ik}. The addition of the small additive constant (0.01) to the denominator in Equation (4.4) prevents division by zero, and applies an upper bound (of $10n_i$) for extreme values for r_{ik} in cases where the *a posteriori* expected number of infected provinces, $\mathsf{E}[y_{ik}] = n_i \hat{p}_{ik}$, is very small.

Identification of anomalous data was investigated by examining the squared Pearson residuals (SPR), r_{ik}^2. If the model provides a good fit to the data then these squared residuals should follow an approximate chi-squared distribution on one degree of freedom. Values $r_{ik}^2 = 5$ and $r_{ik}^2 = 10$ correspond to nominal tail probabilities of approximately 0.025 and 0.001 respectively. Given that there are $64 \times 3 = 192$ residuals, it follows that only values $r_{ik}^2 > 10$ would raise any significant concern about the presence of outliers.

The SPR for the model depicted in Equation (4.3) is shown in Figure 4.4. We see that none of the squared Pearson residuals for our model exceeded ten, suggesting that at face value the model appears to be fitting well. However, the Pearson residuals only describe model fit on the scale of the response variable, y. Therefore, this model could appear to be fitting well because of the high level of flexibility offered by the random effects. We explore this possibility in the next section by computing exceedance probabilities for the random effect terms.



(a) 2006 (b) 2007



(c) 2008

Figure 4.4: Square Pearson residual plots: (a) 2006, (b) 2007, (c) 2008.

4.4 Exceedance Probabilities for Random Effects

4.4.1 Defining the Exceedance Probabilities

The term ε_{ik} in Equation (4.3) models the otherwise unexplained heterogeneity in the data. A case (i.e. province-year combination) with an extreme value for this random effect is unusual and may warrant further investigation. We need then some way of quantifying the 'oddness' of each ε_{ik}. Exceedance probabilities provide a statistically principled way of doing so.

From the definition of our model, $\varepsilon_{ik} \sim N(0, \sigma^2)$ *a priori* for all $i = 1, \ldots, 64$, $k = 1, 2, 3$. It follows that -1.645σ and 1.645σ are respectively the 0.05 and 0.95 quantiles for this distribution, while -2.326σ and 2.326σ are the 0.01 and 0.99 quantiles. These can be thought of as threshold values. If we have strong evidence from the posterior estimates that some random effect exceeds, for example, 2.326σ then this indicates that this case has a substantially greater chance then expected of being infected with FMD, given its characteristics.

For any given random effect ε, we can measure our confidence that it exceeds some threshold τ, and hence lies in the extreme tails of the distribution. We do this by computing the posterior probability of the event $\varepsilon > \tau$ (for an upper tail threshold) or $\varepsilon < \tau$ (for a lower tail threshold). We term these *exceedance probabilities*. For example, if $\boldsymbol{y}^{\mathsf{T}} = (y_1, \ldots, y_{64})$ denotes the full set of disease data, then $\mathsf{P}(\varepsilon < -2.326\sigma \mid \boldsymbol{y})$ is the 1% lower tail exceedance probability, and $\mathsf{P}(\varepsilon > 1.645\sigma \mid \boldsymbol{y})$ is the 5% upper tail exceedance probability.

Exceedance probabilities are easily calculated in practice from the (post convergence) output from an MCMC run. Specifically, if we monitor the sampled values for ε_{ik} then the proportion of these which exceed τ provides the (approximate) exceedance probability $\mathsf{P}(\varepsilon_{ik} > \tau \mid \boldsymbol{y})$. For thresholds that depend on other model parameters, such as $\tau = -2.326\sigma$, we must replace those unknowns with posterior estimates. For example, in practice we calculate the approximate 1% lower tail exceedance probability $\mathsf{P}(\varepsilon < -2.326\hat{\sigma} \mid \boldsymbol{y})$ where $\hat{\sigma}$ is the posterior mean of σ.

The use of exceedance probabilities is an effective method of quantifying the extent to which cases are unusual, though they do not provide any direct explanation as to the cause of any observed anomalies. Nonetheless, a large upper tail exceedance probability indicates an excessive number of infected communes, and could indicate an unusually severe provincial outbreak (compare with Diggle et al., 2005, for an analogous interpretation with point process disease data). Conversely, a large lower tail exceedance probability points to a shortfall in the expected level of FMD. This might be due to unobserved epidemiological factors, but could also be explained by under-reporting of the disease.

4.4.2 Application of Exceedance Probabilities to the Viet Nam Data

As we are interested in the possibility of detecting areas of possible underreporting we focus our attention on the lower tail exceedance probabilities for ε_{ik}. Figure 4.5 shows the lower 5% tail exceedance probabilities for each of the three years 2006, 2007 and 2008. While the plots for the latter two years are unremarkable, the plot for 2006 is interesting. The clustering of provinces with unexpectedly low incidence of FMD is a characteristic of the data, not an artefact of our model, since we do not include any representation of spatial correlation. With the existence of a single province (or a few scattered provinces) with extreme negative random effects we would need to be wary of over-interpretation. After all, some cases must lie in the lower tail by chance alone. However, the observed spatial clustering of extreme provinces in 2006 strongly suggests some systematic effect. One plausible interpretation is that the true level of FMD in the highlighted provinces was under-reported.

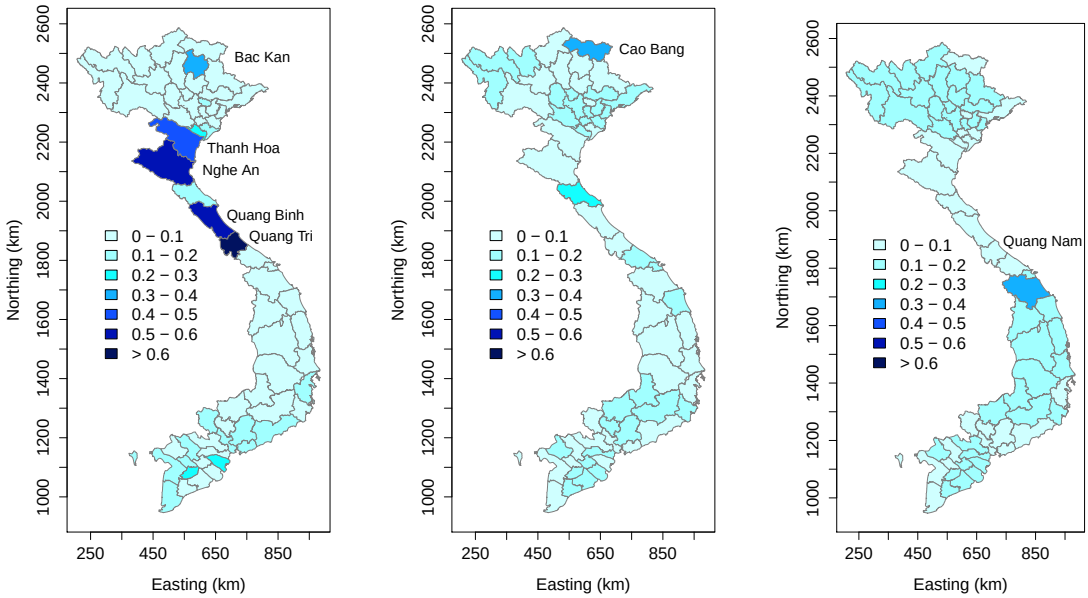


Figure 4.5: Random effect exceedance probabilities for the 5% lower tail for 2006 (left), 2007 (middle) and 2008 (right). Provinces Cao Bang, Bac Kan, Thanh Hoa, Nghe An, Quang Binh, Quang Tri and Quang Nam (listing from north to south) are flagged as having relatively high exceedance probabilities.

We investigated further by looking at results from models incorporating serotype effects. Our findings for serotype O (the most common strain of FMD reported in Viet Nam for the period 2006 to 2008) are the most noteworthy. In Figure 4.6 we display the exceedance probabilities by year for the lower 10% tail. The spatial pattern of extreme provinces for 2006 reflects that for the model without serotype, but we also see at least a suggestion of persistence into 2007.

We also note that a single province in the far north of the country, Cao Bang, is highlighted as unusual in both 2007 and 2008.

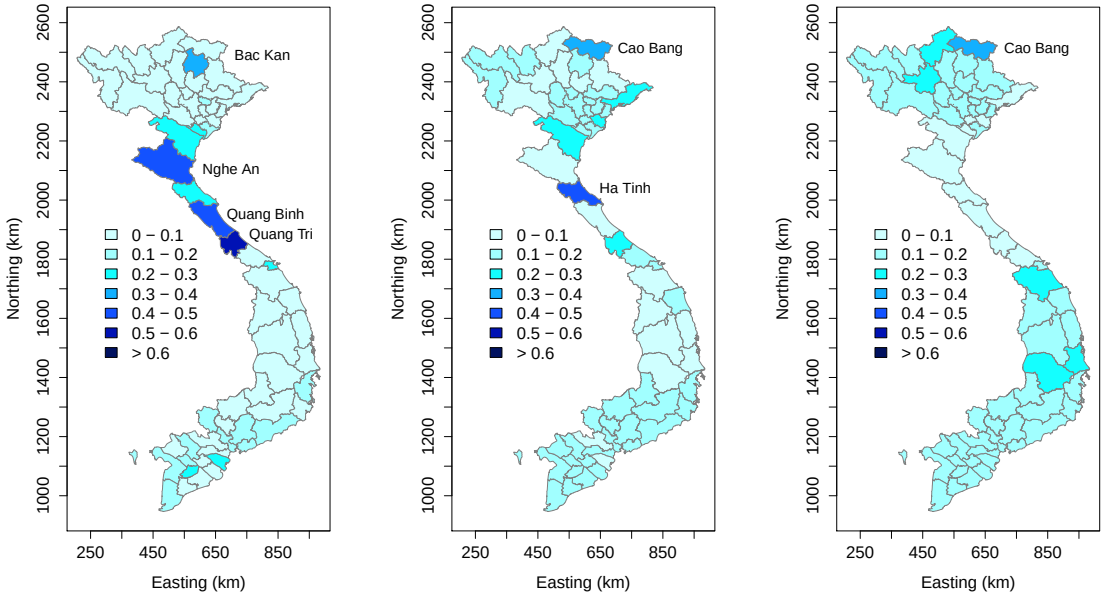


Figure 4.6: Random effect exceedance probabilities for the 10% lower tail for FMD serotype O only, for the years 2006 (left), 2007 (middle) and 2008 (right).

4.5 Discussion

When fitting epidemiological models to data, it is good practice to conduct some form of residual analysis to assess the quality of the model fit to each of the data points. A systematic pattern of large deviations from the predicted values might simply point to weaknesses in the method in which the model represents the true spatio-temporal epidemiology, but it is also possible that lack of fit to a small number of data may reflect problems with the data quality that are not represented in the model. Problems with data quality in veterinary epidemiology are not uncommon. For example, in a study carried out on the evaluation of foot-and-mouth disease outbreak reporting in mainland South-East Asia from 2000-2010 it was found that 78% of reports did not obtain longitude and latitude coordinates, no serotype was recorded for 43% of cases, and only 44% of reports since 2005 contained information on the number of cases, population at risk, and the number of fatalities (Madin, 2011).

In this chapter we have used exceedance probabilities to identify and assess possible anomalous data records in a dataset providing details of the endemic occurrence of FMD in Viet Nam from January 2006 to December 2008. In 2006, when there were relatively large numbers of

communes identified as FMD-positive, there were unusually low levels of disease in a sequence of provinces (Nghe An, Quang Binh, Qunag Tri, and Thanh Hoa) bordering Laos. See Figure 4.7. Interestingly, the provinces Nghe An, Quang Binh, Quang Tri have been identified as entry points into Viet Nam for cross-border movement from Laos of large ruminants and pigs. We also note that the northerly province of Cao Bang is an exit point into China for the same species. In more detail, a study by Cocks et al. (2009) suggested that the main pathways for movement of these animals are as displayed in Figure 4.7.

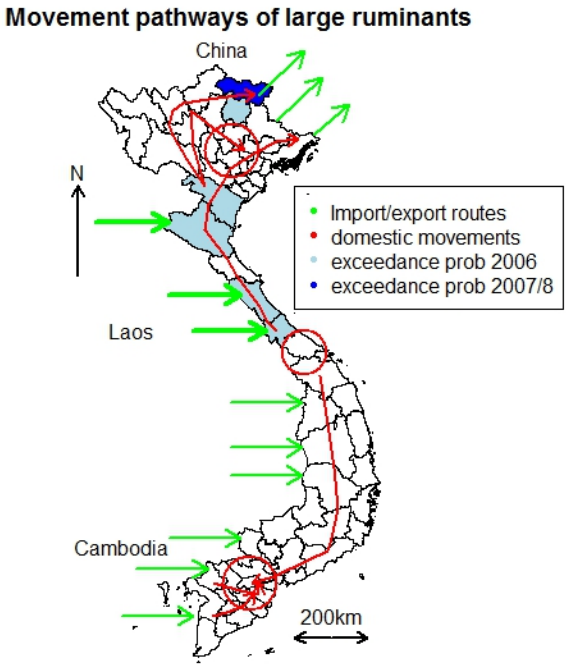


Figure 4.7: Main movement pathways of large ruminants involving Viet Nam including imports, domestic movements and exports. Green arrows indicate major entry/exit points. Shaded provinces are those with high exceedance probabilities, indicating unexpectedly low rates of FMD infection.

The presence of seemingly low disease probability in areas with high animal density and movement patterns warrants further investigation. Of course, it may be that this aspect of the data could be explained by unobserved covariates in a straightforward manner. We acknowledge that one of the limitations of this study is the relatively small number of explanatory variables available, and their fairly crude spatio-temporal resolution. Alternatively, it is possible that the unexpectedly low counts arise from under-reporting of FMD incidence. Regardless of the explanation, the ability of exceedance probabilities to draw out this aspect of the results, which is essentially hidden in a plot of squared Pearson residuals, illustrates their utility as a

model diagnostic. In the absence of other plausible explanations for our findings, we recommend that disease surveillance and disease event recording procedures in the Department of Animal Health offices in those provinces with unusually low levels of disease, are reviewed.

Exceedance probabilities can be helpful in the analysis of any type of hierarchical spatio-temporal model, and can therefore be employed in a myriad of infectious disease modelling problems. What is more, they can be computed in a straightforward manner from the output of a MCMC algorithm, almost regardless of the complexity of the model itself. For example, we considered refinements of the model in Section 4.3 in which we included a spatial CAR prior (e.g. Waller et al., 1997) to account for otherwise unexplained dependence between neighbouring provinces. This was, however, not strongly pursued as we found negligible unexplained spatial correlation in our models as a whole (the spatial clustering on the low disease provinces in 2006 notwithstanding), but nonetheless, computation of the exceedance probabilities was little more complex than for the independence model.

Chapter 5

Spatial clustering through time: an extension to the epidemic curve

5.1 Introduction

Epidemiologists use epidemic curves to describe the temporal pattern of disease onset in populations of individuals at risk. While epidemic curves can provide important clues about likely mechanisms of disease spread, they simply indicate counts of cases as a function of time and hence provide limited insight into factors driving disease transmission. For example, in an infectious disease epidemic, are there large numbers of cases observed over a short period of time due to spread from local contact, or simply due to lots of sporadic cases? A critical issue in understanding the characteristics of disease transmission is to look at spatial clustering of cases.

An epidemic curve in its simplest form plots the number of new cases for each time period. The main purpose of such plots is to visually portray any temporal trends and the type of outbreak. When disease occurrence is low, epidemic curves are useful in the decision making process regarding control methods and when they should be applied. In some circumstances they can also be used to estimate incubation times and to compare disease frequency in different populations (Salman, 2003). The epidemic curve can also be used to investigate clustering of disease over time. However, the clustering of disease can be subtle and complex so results from simple graphical techniques may be difficult to interpret and therefore, conclusions may be invalid (Ward and Carpenter, 2000).

With disease and ill-health, clustering in space and/or time is common place. This is generally because the disease is contagious or point source in nature. Consequently, if we fail to consider

and investigate such clustering, any procedures to control and eradicate could be hampered (Carpenter, 2001).

Clustering in point pattern processes can be investigated through second order analysis, for example using the inhomogeneous K-function. In this chapter we present a graphical tool that provides a means for summarizing the degree of spatial clustering in infectious disease outbreak data, and to document how this changes during the course of an epidemic. Our tool is not aimed at detecting clusters in space and time, but rather at showing how the pattern of spatial clustering changes with time. We propose that this approach should be used in conjunction with the epidemic curve to shed more light on the progression and structure of an outbreak. We illustrate our methodology using simulated data and later apply it to epidemics of foot-and-mouth disease (FMD) from England and Japan.

5.2 Models

To illustrate the motivation for and the use of our clustering tool we will look at four artificial disease datasets. The first pair are drawn from an underlying spatially homogeneous population and the second pair from an inhomogeneous population. For each underlying population we generate two datasets. The first is a Poisson point process (PPP) with the second being a cluster Poisson point process (CPPP). All the datasets were created using the Space-Time Point Pattern simulation, visualisation and analysis (stpp) package in R (Gabriel et al., 2014).

In section 2.2 we introduced spatial models for the homogeneous and inhomogeneous spatially distributed datasets. As we will be investigating the changes in spatial clustering through time, here we will discuss spatial-temporal models for homogeneous and inhomogeneous spatially distributed datasets.

We continue to work with Poisson point processes, for which we define the spatial-temporal intensity as $\lambda(\boldsymbol{x}, t)$ for location $\boldsymbol{x}$ at time t. Following Gabriel et al. (2013) we assume separability, and consequently decompose the intensity as

$$\lambda(\boldsymbol{x}, t) = \lambda_s(\boldsymbol{x})\lambda_t(t) \tag{5.1}$$

where λ_s and λ_t are separate intensity functions over space and time respectively. The number of points N falling in an area A over the time interval (t_1, t_2) is distributed according to

$$N \sim \text{Pois}(\mu)$$

where

$$\mu = \int_{t_1}^{t_2} \int_A \lambda_s(\boldsymbol{x})\lambda_t(t) d\boldsymbol{x} dt. \tag{5.2}$$

5.2.1 Homogeneous Spatio-Temporal Models

For spatio-temporal homogeneous models we require $\lambda_s(\boldsymbol{x})$ and $\lambda_t(t)$, from equation 5.2, to be constant (i.e. not dependent on x and t respectively). This can occur for example with an endemic disease where the population at risk is uniformly distribution over space. Such a model is rather unrealistic (since real world populations tend to highly non-uniform) but it provides an easily interpretable test case.

5.2.2 Inhomogeneous Spatio-Temporal Models

The inhomogeneous spatial temporal models are an extension to the homogeneous case where $\lambda_s(\boldsymbol{x})$ is allowed to vary. If we keep $\lambda_t(t)$ constant then the model can be thought of as representing an endemic disease with spatial variation in the underlying population at risk. For our test models we define the spatial component of the intensity as

$$\lambda(\boldsymbol{x}) \propto 0.01 + 0.2N(\boldsymbol{x}|\boldsymbol{\mu}_1, \Sigma) + 0.1N(\boldsymbol{x}|\boldsymbol{\mu}_2, \Sigma) \tag{5.3}$$

where $N(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma)$ denotes a bivariate normal density function with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ evaluated at $\boldsymbol{x}$. We set $\boldsymbol{\mu}_1 = (34, 54)^T$, $\boldsymbol{\mu}_1 = (35, 52)^T$, and

$$\Sigma = \begin{pmatrix} 4 & 0 \\ 0 & 3 \end{pmatrix}.$$

The visual portrayal of λ is shown in Figure 5.1. Our function mimics an example of an urban area (formed from two distinct centres of population) surrounded by more sparsely populated rural areas.

5.2.3 Cluster Poisson point process

The CPPP datasets are formed through a kind of parent-child process. The parents of the process are generated in the same way for the homogeneous and inhomogeneous processes. That is, they follow PPP models with intensity functions given by equation 5.1 with constant λ_t, and λ_s as a constant function (in the homogeneous case) or defined by equation 5.3 (in the

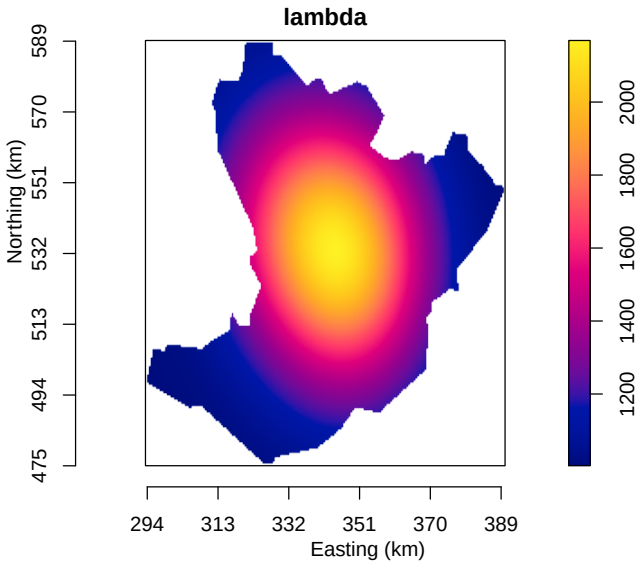


Figure 5.1: Spatially inhomogeneous intensity function λ_s.

inhomogeneous case) The parents can be seen as the causes for an outbreak of disease. After the initial effect of creating the children, the parents themselves are not recorded as events in the final dataset. For instance, the parents could represent interactions between members of the population at risk (say cattle) and wild animals acting as a host for a disease. An example is provided by the relationship between possums and TB spread among cattle. A possum with tuberculosis (TB) may cause the spread of the disease to the surrounding cattle, which are recorded as our events, but the actual possum that caused the spread is unknown and not recorded.

Once a parent is specified in space and time, the distribution of children given the parent is defined as a Poisson point process with intensity function given by

$$\rho(\boldsymbol{x}, t) = \rho_s(\boldsymbol{x})\rho_t(t) \propto \exp\left(-\frac{1}{2}\frac{\| \boldsymbol{x} - \boldsymbol{x}_{parent} \|^2}{(\frac{\tau_1}{2})^2}\right) \times \exp\left(-\frac{1}{\tau_2}|t - t_{parent}|\right) \tag{5.4}$$

where $\boldsymbol{x}_{parent}$ and t_{parent} denote the location and time of occurrence for the parent point. We assume that $t \geq t_{parent}$. It follows that conditional on the parent, each child is spatially distributed according to a bivariate normal distribution, and temporally distributed according to an exponential distribution. When creating a cluster process we can stipulate dispersion (scale) parameters, τ_1 and τ_2 for the spatial and temporal dispersion respectively. We illustrate

this idea in Figure 5.2 where the parents are represented by blue circles, and the children represented by red crosses. The distribution of these children is defined by the intensity function above, so that children typically fall within a circle of radius $2\tau_1$ of a parent indicated by the green arrow. In this example we used $\tau_1 = 2.5$km.

Based on this definition, examples of the spatial temporal process of the children (events) are shown in Figure 5.3. The datasets were stipulated to be created using two parents with an average of seven children per parent. To demonstrate the temporal scale the size of the points signifies the recentness of the event: the larger the point, the more recently the event occurred.

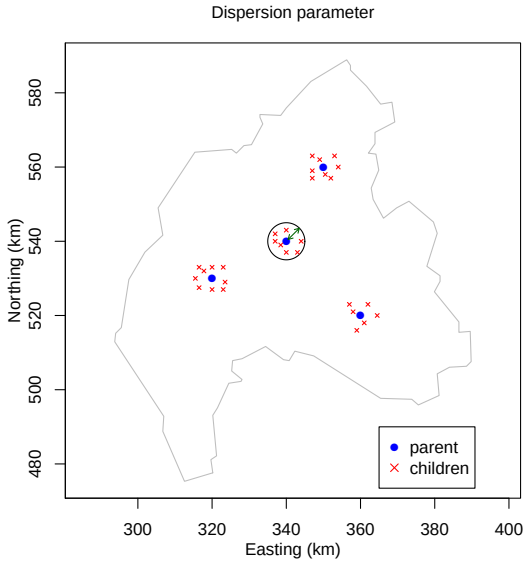


Figure 5.2: Illustration of the procedure for the cluster Poisson point process (CPPP), with parent points represented by blue circles, and the children represented by red crosses. The circle describes the role of the dispersion parameter in determining the concentration of children around each parent.

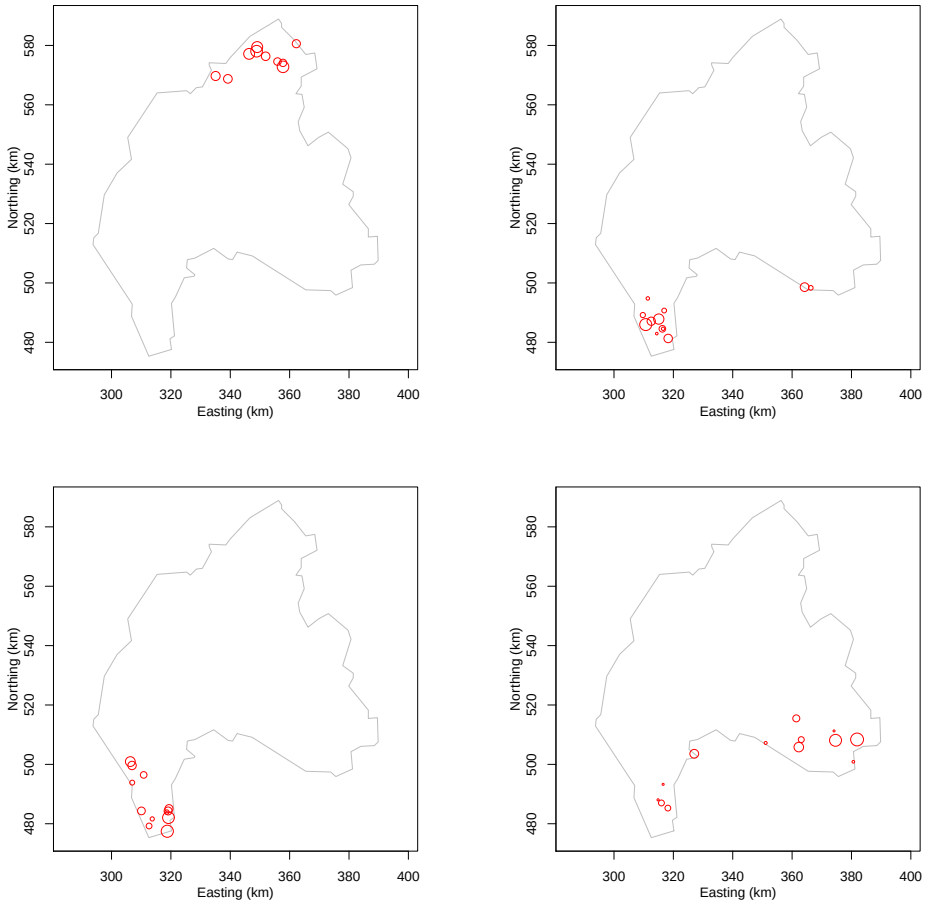


Figure 5.3: Parent/children process with temporal component. Plotting size indicates how recently the case occurred, with large symbols indicating very recent events and small symbols earlier ones.

5.3 Implementation

The datasets were created using the spatial window shown in Figure 5.4. This is the border outline of North Cumbria, England. We used a real geographical window for our analysis, rather then a default square, to reflect the 'typical' properties of study regions in practice.

We implemented the methods described previously with this spatial window to create our four datasets.

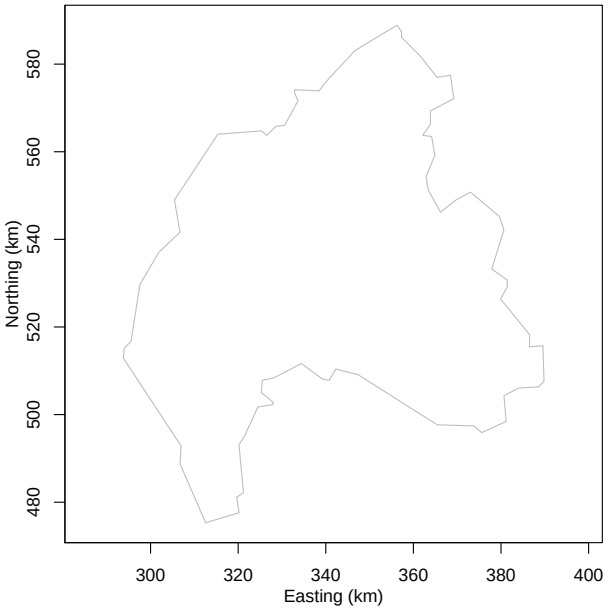


Figure 5.4: Study region for artificial datasets (based on the English county of Northern Cumbria)

5.3.1 Dataset specification

Spatially homogeneous datasets

For the spatially homogeneous Poisson point patterns the underlying intensity λ is constant. Using the stpp package in R (Gabriel et al., 2014), the distribution of points can either be specified as the total number of points over the window or as the expected number of points per unit area. For our datasets we will implement the models specifying the total number of points.

Our PPP dataset was created with the specification of a total number of 100 points and a maximum time window of 50 days, with the final dataset lasting 45 days (100 points were reached within 45 days). In Figure 5.5 we see the progression of the outbreak every 10 days. Red events represent the new cases reported on the day and grey points represent the past events where the size of the plotting point represents the temporal component, the larger the size the more recent the event. The data appears to be spread out over the entire region.

Our CPPP dataset started with five parents and was allowed to progress until a maximum of 100 points was simulated or the temporal component reached a maximum time range of 50 days. As with the homogeneous PPP dataset, λ is constant and the intensity is set by the

maximum number of points when implemented using the stpp package. The spatial dispersion parameter, τ_1 for the children was set to a radius of 2.5km and the temporal dispersion defined by $\tau_2 = 5$ days. We deemed the process infectious and therefore, offsprings' times were always greater than parents' times. The selected dataset reached the maximum number of events (100) within 49 days. In Figure 5.6 we once more look at the spatial-temporal plots at 10 day intervals over the course of the outbreak. Based on these plots we can see that the initial stage of the outbreak ($\leq$10days) is characterised by local spread. By day 20, there has been a long range transmission, followed by a period of local spread. The first grouping appears to still be bubbling away. By day 30 the outbreak appears to be dying out. At day 40, however, we see that the outbreak has not died off and instead observe another long range transmission. By the last day of the recorded outbreak we see significant local spread in a short time frame.

So far we have created two datasets for which it is very easy to spot spatial clustering, and where any clustering that does exist must be due to disease transmission because the underlying intensity λ is constant. We now want to create a more challenging pair of examples with variation in the underlying spatial intensity, so that groupings of points may be attributed to variation in λ or clusters from local transmission. This provides a more challenging and realistic scenario for testing the cluster curve.

Spatially inhomogeneous datasets

For most practical applications it would be entirely unrealistic to assume that the underlying population at risk is uniformly distributed over space. Therefore, we also consider datasets selected from a spatially inhomogeneous population, where variation in intensity over the study area is allowed to occur according to equation 5.3.

Once again we create two datasets, a PPP and a CPPP.

The PPP dataset is specified to have a maximum of 45 cases and a maximum time range of 50 days. The resulting dataset covered all 50 days with 45 cases in total. The resulting spatial temporal plots are shown in Figure 5.7. From these plots we can see that the initial progression of the disease is slow. The cases are concentrated around the centre of the region, with a smaller number in the tails. The presence of a centre grouping for the PPP dataset is of course not a disease cluster but reflects the higher population density.

The CPPP dataset was specified to be formed by 20 parent cases with a mean of 2 children per parent case. Parents were generated through the PPP with intensity function from equation 5.3, and conditional on those, children follow a process with intensity from equation 5.4 with a spatial scale parameter $\tau_1 = 2.5$ and temporal scale parameter $\tau_2 = 5$. Similar to the

PPP dataset, a maximum time region of 50 days is specified. The resulting dataset lasted 49 days and contained 39 events. In Figure 5.8 we illustrate the spatial-temporal distribution of events every 10 days. We see that in the first 10 days no events occurred. By day 20 we see a grouping and a possible long range transmission. Over the next 29 days we see local spread occurring within these two groupings.

In general we see that the PPP dataset covers a lot more of the spatial window while there is a presence of two main clusters (pockets) of disease for the CPPP dataset.

5.3.2 Epidemic curves

We now pretend that we do not know the processes that generated these datasets. Instead, we ask whether it is possible to infer characteristics of the generating process using the epidemic curve.

Homogeneous population

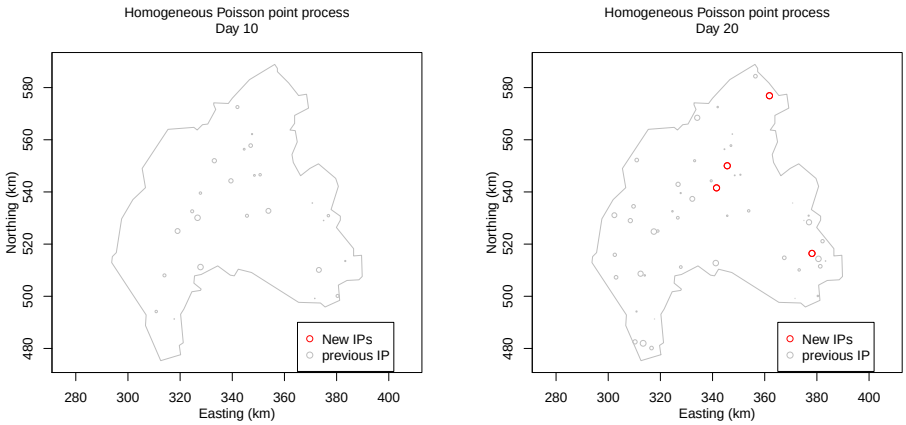
The epidemic curves for the homogeneous datasets are shown in Figure 5.9. These plots show both the daily cases over time as well as the the 7-day moving window. We see from the daily epidemic curve that in the initial stages of the outbreak the two curves; PPP and CPPP, are very similar. In such a situation, an epidemiologist would have very little reason to suspect that the disease processes are, in truth, very different.

Inhomogeneous population

The epidemic curves for the inhomogeneous datasets are shown in Figure 5.10. Similar to the homogeneous case we look at both a temporal daily scale as well as a 7-day moving window. We see that from 10 days into the epidemic for both the daily and the 7-day window there is a similar structure to the epidemic curve. Therefore, from these two curves we are unable to see any difference between the outbreaks and in particular any differences in the progression of the spatial patterns of disease.

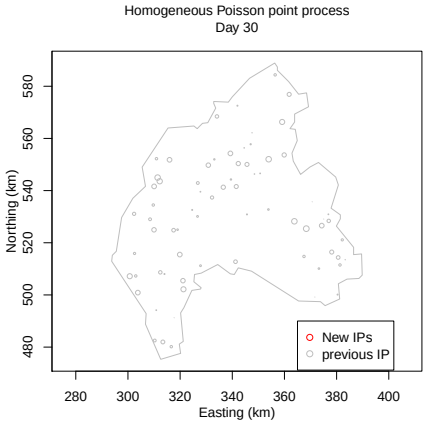
The examples we have used, both with spatially homogeneous and inhomogeneous populations, illustrated the limitations of the epidemic curve (on its own) as a guide to the characteristics of the disease. As the datasets were artificial we know the processes by which they were created. By evaluating the epidemic curves, however, we were unable to distinguish between these processes. This is why we propose the use of a new tool called the cluster

curve alongside the epidemic curves.

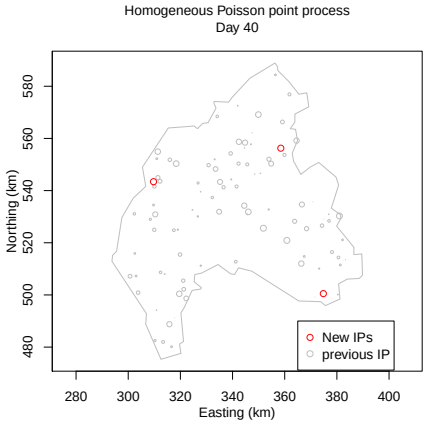


(a) Day 10

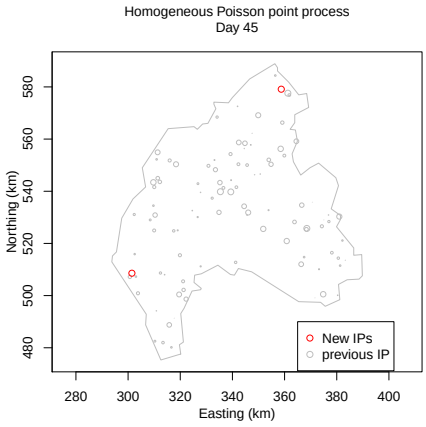
(b) Day 20



(c) Day 30

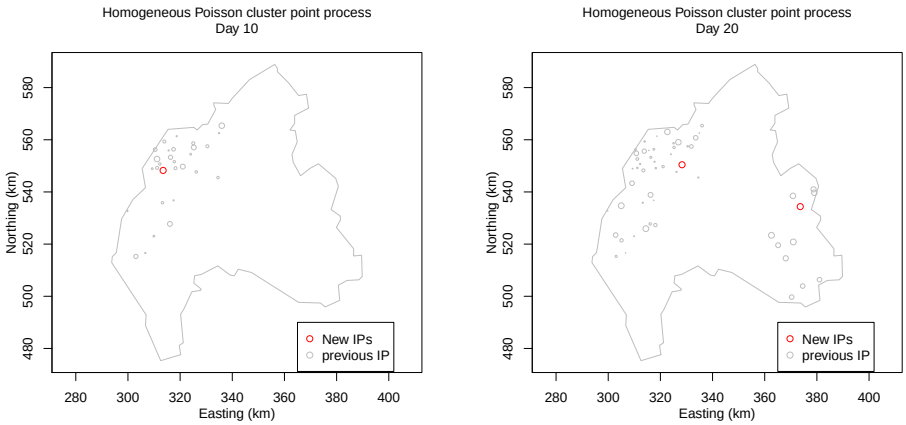


(d) Day 40



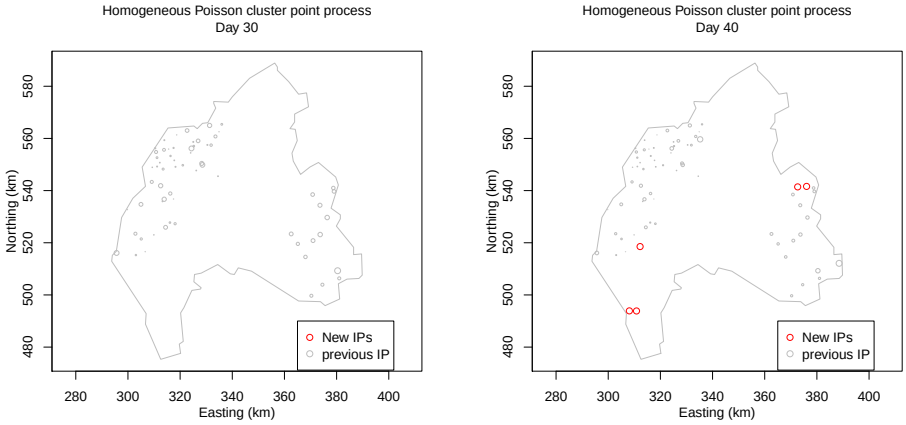
(e) Day 45

Figure 5.5: Spatial-temporal distribution of homogeneous PPP dataset every 10 days, with red representing new cases on the day and grey previous cases. The larger the plotting size the more recent the infection.



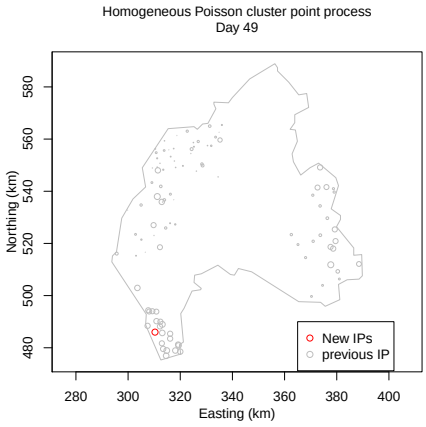
(a) Day 10

(b) Day 20



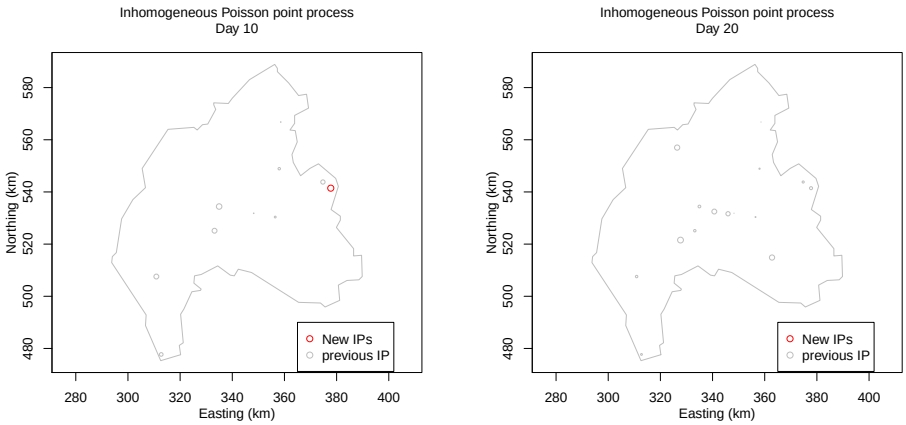
(c) Day 30

(d) Day 40



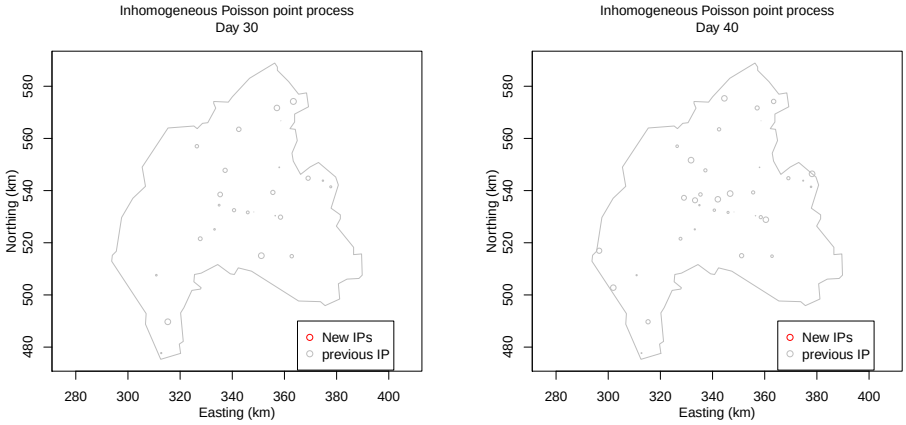
(e) Day 49

Figure 5.6: Spatial-temporal distribution of homogeneous CPPP dataset every 10 days, with red representing new cases on the day and grey previous cases. The larger the plotting size the more recent the infection.



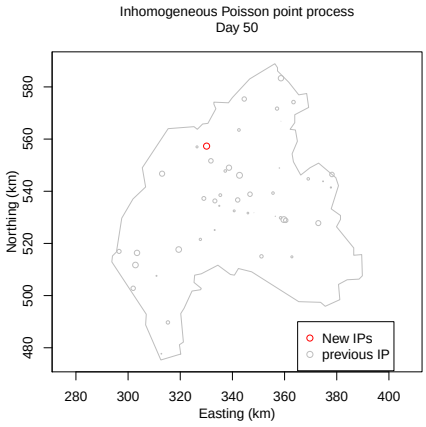
(a) Day 10

(b) Day 20



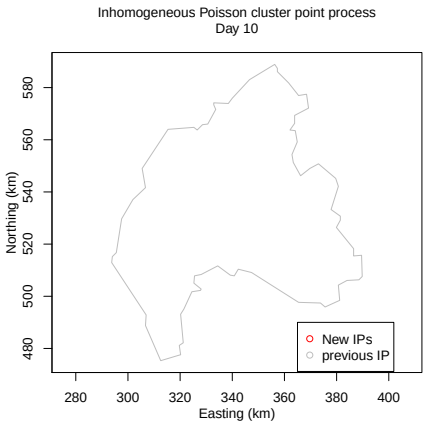
(c) Day 30

(d) Day 40

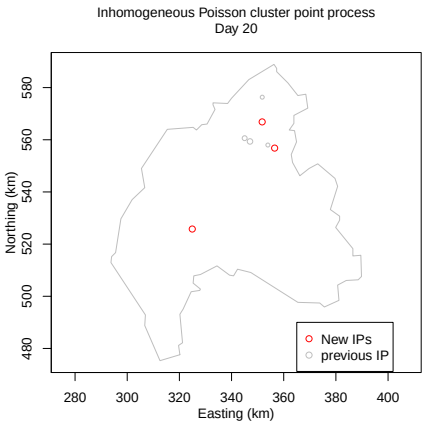


(e) Day 50

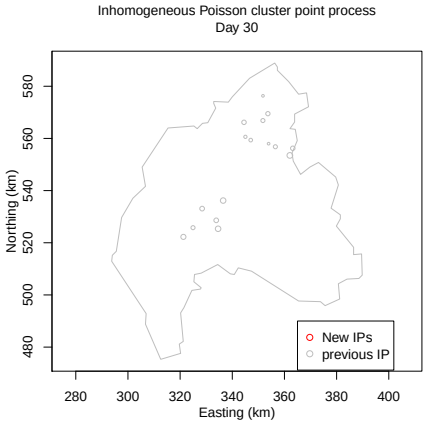
Figure 5.7: Spatial-temporal distribution of inhomogeneous PPP dataset every 10 days, with red representing new cases on the day and grey previous cases. The larger the plotting size the more recent the infection.



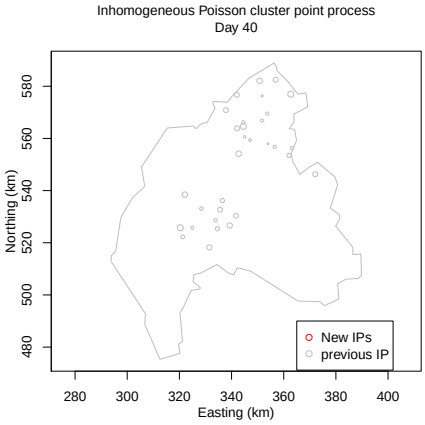
(a) Day 10



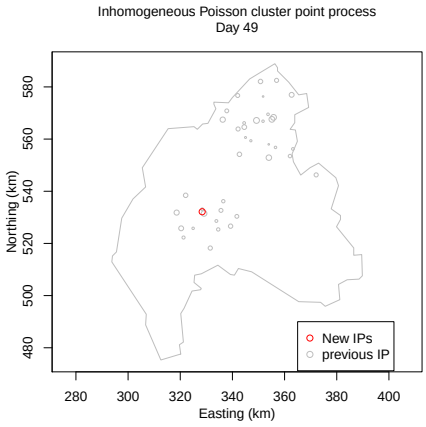
(b) Day 20



(c) Day 30



(d) Day 40



(e) Day 49

Figure 5.8: Spatial-temporal distribution of inhomogeneous CPPP dataset every 10 days, with red representing new cases on the day and grey previous cases. The larger the plotting size the more recent the infection.

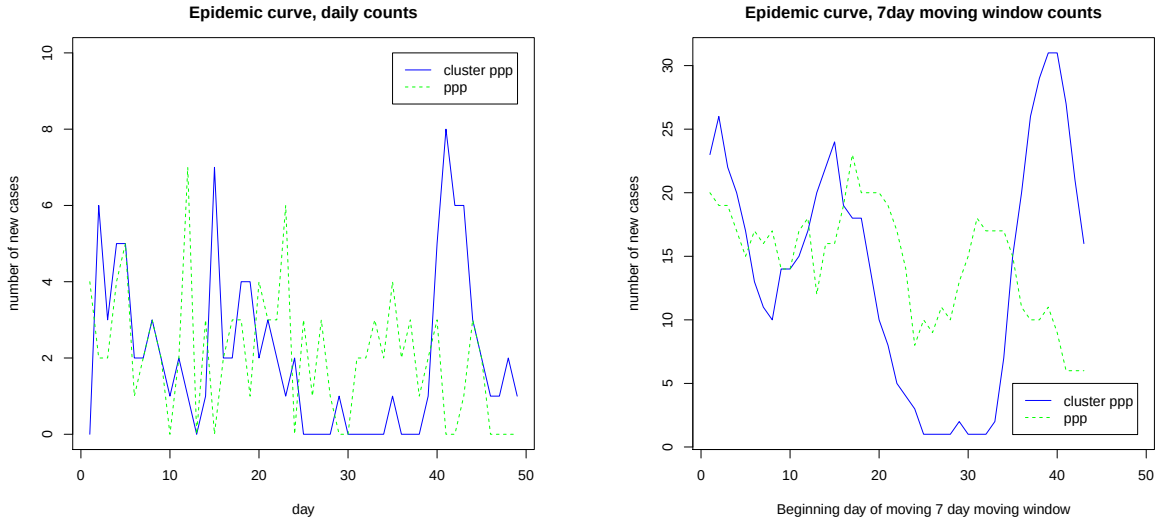


Figure 5.9: Epidemic curves of homogeneous datasets. Blue solid lines represent the cluster Poisson point process and green dashed line a Poisson point process (PPP). Left daily epidemic curve, right 7-day moving window epidemic curve.

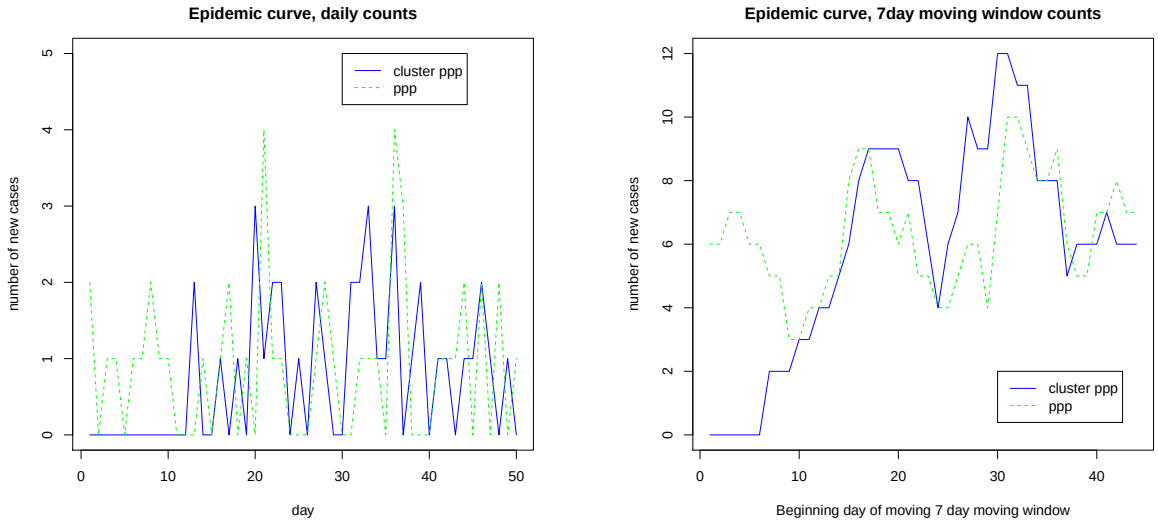


Figure 5.10: Epidemic curve of inhomogeneous datasets. Blue solid lines represent the cluster Poisson point process and green dashed line a Poisson point process (PPP). Left daily epidemic curve, right 7-day moving window epidemic curve.

5.4 Cluster curve

When we look at the epidemic curves we can clearly see temporal trends, but what about the spatial patterns? Though a well trained eye may be able to spot the presence of clustering, its structure can be complex and therefore, difficult to detect and interpret.

If we take our datasets for example, the sets of epidemic curves (PPP and CPPP) are very similar (they were chosen to be that way) but we know, since we created them, that the underlying processes in which they were produced are different, one by a Poisson point process and other by a cluster process based on parents and children. In most circumstances we want to control or eradicate the disease. Methods for disease control will depend critically on its transmission properties, which will be reflected in the patterns of clustering. Therefore, the ability to minimise the impact of the outbreak could be greatly hindered if we fail to consider or investigate the presence of clustering. An epidemiologist would be better able to understand the course of the epidemic if the information from the standard epidemic curve was augmented with a summary of the spatial clustering properties of the process, and how these change with time. This motivates us to define our cluster curve.

The cluster curve makes use of the inhomogeneous K-function (described in detail in section 2.2.3). In essence the K-function measures the degree of spatial clustering by counting the number of observations within a radius (r) while adjusting for the spatial distribution of the population at risk. We are basically looking at how the number of pairs of points over our given radius compares with what is expected given the underlying intensity with no clustering. The cluster curve is created by selecting a radius r, then plotting the difference in the inhomogeneous K-functions based on data falling within a particular time window. For any given value of r, plotting these values for a sequence of time windows covering the entire time period, gives us our cluster curve. We are in effect looking at the changes in spatial clustering through time taking into consideration the groupings that occur because of the spatial variations in the population density.

In preparation for calculation of the cluster curve, we need to subset the data into different time windows, these need not be disjointed. Our usual approach is to use a 7-day moving window, so that the window labelled by time point t contains all the data from day t to day $t + 6$.

Each window must contain a sufficient amount of data points in order to estimate the inhomogeneous K-function, as this function depends upon the population spatial intensity (or density) function λ, which accounts for clustering that would occur because of geographical variations such as urban and rural environments. Such estimation is most straightforward

when we have an exogenous estimate of the population density available, which is the case in all the examples that we consider. If such an independent estimate is not available, then the population density can be estimated directly from the full spatial dataset.

We initially defined our cluster curve function as

$$\Psi(r|t) = \frac{\hat{K}(r|t) - K_{theo}(r|t)}{K_{theo}(r|t)} \tag{5.5}$$

where $\hat{K}(r|t)$ is the observed inhomogeneous K-function statistic at a radius r and $K_{theo}(r|t)$ the theoretical inhomogeneous K-function statistic under the assumption of no clustering at radius r, both based on data for time interval t only. The numerator compares the observed degree of clustering with that expected if cases occur independently. By dividing this difference by the theoretical K-function, we are developing a statistic that does not depend on the overall size of the epidemic. It is equally adept at representing clustering in an epidemic with 20 cases as it is in an epidemic with 2000 cases.

A problem can arise where the cluster curve can take large values if K_{theo} is very small, this problem is quite common with small epidemics. In essence, Ψ becomes very unstable as it stands. We therefore, adapt the original definition to include the addition of a constant (e.g. 1) to the denominator. We now redefine our cluster curve as:

$$\Psi(r|t) = \frac{\hat{K}(r|t) - K_{theo}(r|t)}{(K_{theo}(r|t) + 1)}. \tag{5.6}$$

In practice we will plot $\Psi(r|t)$ against t to produce the cluster curve for radius r. In the analysis of the curve a positive value of the curve represents the presence of clustering. A negative value; a regular pattern and a value around zero; independence.

An illustration of the ideas behind the cluster curve for one time interval is shown in Figure 5.11. Here the purple line represents the difference between the observed and theoretical curves at r = 10km radius (1unit = 10km). From this plot we can also see that there are multiple estimates for the estimated observed inhomogeneous K-function. These are based on the corrections applied; isotropic (iso), translate (trans), border (bord) and border modified (bordm). We proceed using the translate correction method as it is computationally efficient.

5.4.1 Cluster curve smoothing

In the application of the cluster curve we require a minimum of only three events within a window in order to estimate the inhomogeneous K-function (given that an exogenous estimates

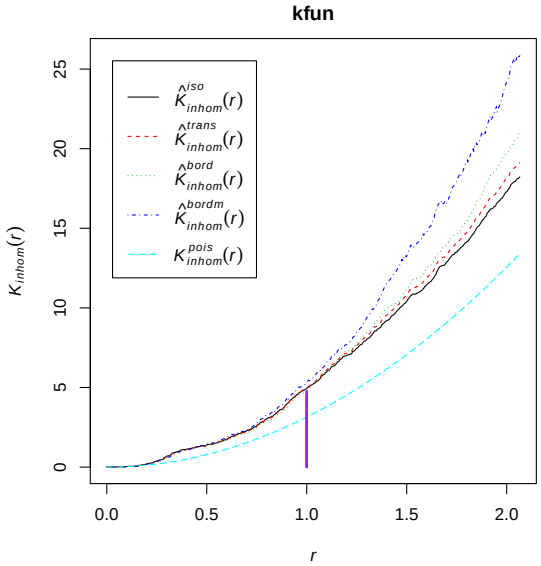


Figure 5.11: Inhomogeneous K-function example, showing the difference between the theoretical and the observed inhomogeneous K-function.

of the population density is available). Low sample sizes in some or all time windows could lead to very noisy cluster curves. Some examples are given in Figure 5.12, where gaps in the curve correspond to time windows with insufficient data to compute estimates of the inhomogeneous K-function. To tackle this issue we provide the option to implement some additional smoothing.

The required level of smoothing depends on the choice of time window length. If a large time window is chosen then the resulting curve is unlikely to be very noisy, possibly rendering smoothing unnecessary. However, the disadvantage of long windows is that they may fail to pick up spatial clusters that last for only a relatively short time.

There are many methods available for smoothing sequences of data. These include kernel smoothing (Wand and Jones, 1995), lowess (Cleveland, 1979), and splines (Green and Silverman, 1994). We chose lowess, because of its reliable performance and ease and speed of implementation. Within the lowess function we used a Gaussian family, therefore, performing the smoothing via fitting least-squares (Jacoby, 2000). The degree of smoothing is controlled by a parameter α that specifies the proportion of the data used to compute each local regression.

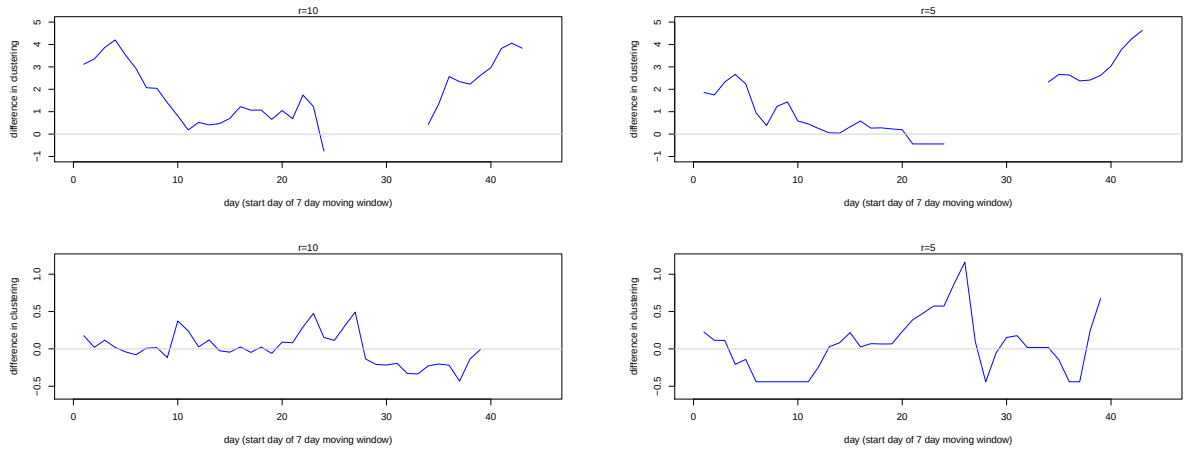


Figure 5.12: Raw cluster curve examples

5.4.2 Implementation in R

Our cluster curve operates by evaluating the difference in clustering for a given radius r. More specifically by fixing r we are examining clustering only up to that distance. We often select r based on prior knowledge of the disease. For example, if it is known that the disease in question rarely causes infection at a range of more than 5km, (as is the case for Foot-and-Mouth disease, for instance) then we might choose to work with the cluster curve with $r = 5$. However, variation in the spread naturally occurs during an outbreak due to factors such as farm locations and geographical features. It may often be the case that the epidemiologist does not have a specific radius in mind, and may want to explore various r values. This motivates development of a software implementation of the cluster curve in which it is straightforward to switch between values of r.

For the ease of switching between radiuses we use a slider. This is an adornment to the plot by which the user can control the choice of r dynamically, using the mouse. An example of the slider and resulting cluster curve is shown in Figure 5.13. This allows users to easily switch between radiuses without having to rerun the R function. As can be seen from this Figure, our slider also allows the degree of lowess smoothing to be adjusted.

Calculation of the inhomogeneous K function is moderately computationally intensive. As a consequence, if this function needs re-evaluating each time a new value of r is selected, the slider will appear very unresponsive. Our solution is to perform an initial evaluation of the K function over a grid of values of r, using the efficient `kinhom` function in the spatstat package in R (Baddeley et al., 2015). The cluster curve and slider are only plotted once this initial step is completed. Any subsequence change of radius is handled through linear interpolation

(implemented using function `approxfun` within R (Becker et al., 1988)). Specifically, this approximates the inhomogeneous K-function at a radius r by using an appropriate convex combination of the values $K(r_1|t)$ and $K(r_2|t)$ computed at the grid points r_1 and r_2 to either side of r. This allows the slider to work more efficiently.

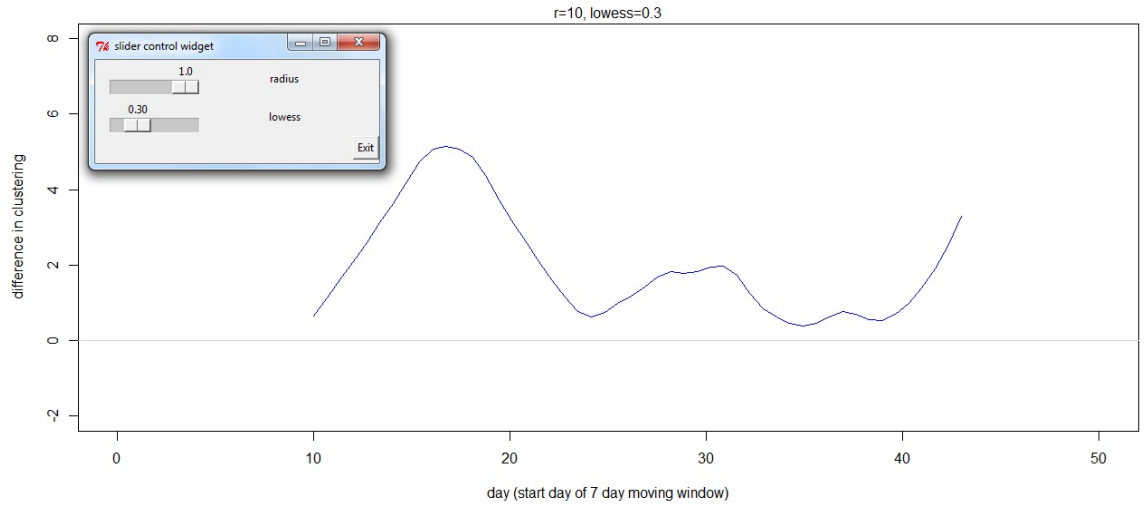


Figure 5.13: Cluster curve slider example.

The application of this tool is shown in the following sections for our spatially homogeneous and inhomogeneous datasets.

5.4.3 Application of the Cluster Curve to simulated data

Homogeneous population

The application of the cluster curve to the homogeneous data is shown in Figure 5.14. We look at radiuses of 2,4,6,8 and 10km with a lowess smoothing where $\alpha = 0.25$. The homogeneous PPP cluster curves are on the left and CPPP on the right.

When we compare the CPPP and PPP cluster curve plots we see that at more time periods for a greater range of radiuses the CPPP dataset appears to have a higher level of grouping amongst events than the PPP dataset. The PPP datasets for almost all radiuses at all time windows shows little difference from what is expected based on a homogeneous Poisson process. For the CPPP datasets we see that the greatest degree of clustering difference occurs at the larger radiuses. We also see that the heightened degree of clustering occurs in peaks and troughs throughout the entire time period. This is generally characterised by periods of

local spread followed by a long range transmission which then causes a new cluster of local spread.

In the interpretation of these curves negative areas almost certainly do not indicate 'anti-clustering', rather, they are an artefact of the methodology. Specifically, the K_{theo} curve when computed from data that is in truth clustered will give results averaged over clusters and areas without data, so that in the latter areas it will indicate that there are greater distances between points that is anticipated by a PPP model. It follows that the dips in the right hand plots are of little interest.

As mentioned previously, when we look at the epidemic curves the two curves are similar, especially between day 5 and day 25 with a few days delay between peaks, but when we look at the clustering structure of the two datasets they are different. This information on the spatial structure was unable to be easily detected via the epidemic curves.

Inhomogeneous population

The application of the function to the inhomogeneous data is shown in Figure 5.15. We look at radiuses of 2,4,6,8 and 10km with the homogeneous PPP on the left and CPPP on the right. This was implemented using true population intensity $\lambda(\boldsymbol{x}, t)$ as an input into the inhomogeneous K-Function.

Once again we see that the CPPP dataset has a higher degree of clustering when compared to the PPP dataset. For the CPPP dataset we have an initial large amount of clustering, a moderate degree in the middle of the epidemic and and increasing degree in the latter stages. This increase in the latter stages suggest that the end of the outbreak was mainly characterised by local spread. For the PPP dataset we see slightly more clustering indicated than we did for the homogeneous PPP dataset, but still with little deviations from what is expected. The small deviations in the PPP plots could be the presence of noise in the cluster curve, we will consider methods of addressing this in chapter 6.

When we look at the epidemic curves from day 10 onwards we see very little difference between the temporal patterns for the two datasets. However, from the spatial plots and the cluster curve we once again can see that the spatial distribution of the two cases are different, with the CPPP dataset having significantly more clustering.

A basic examination of the spatial distribution may suggest more clustering in the inhomogeneous case than the homogeneous, our cluster curve adjusts for this through its use of the inhomogeneous (as opposed to homogeneous) K-function.

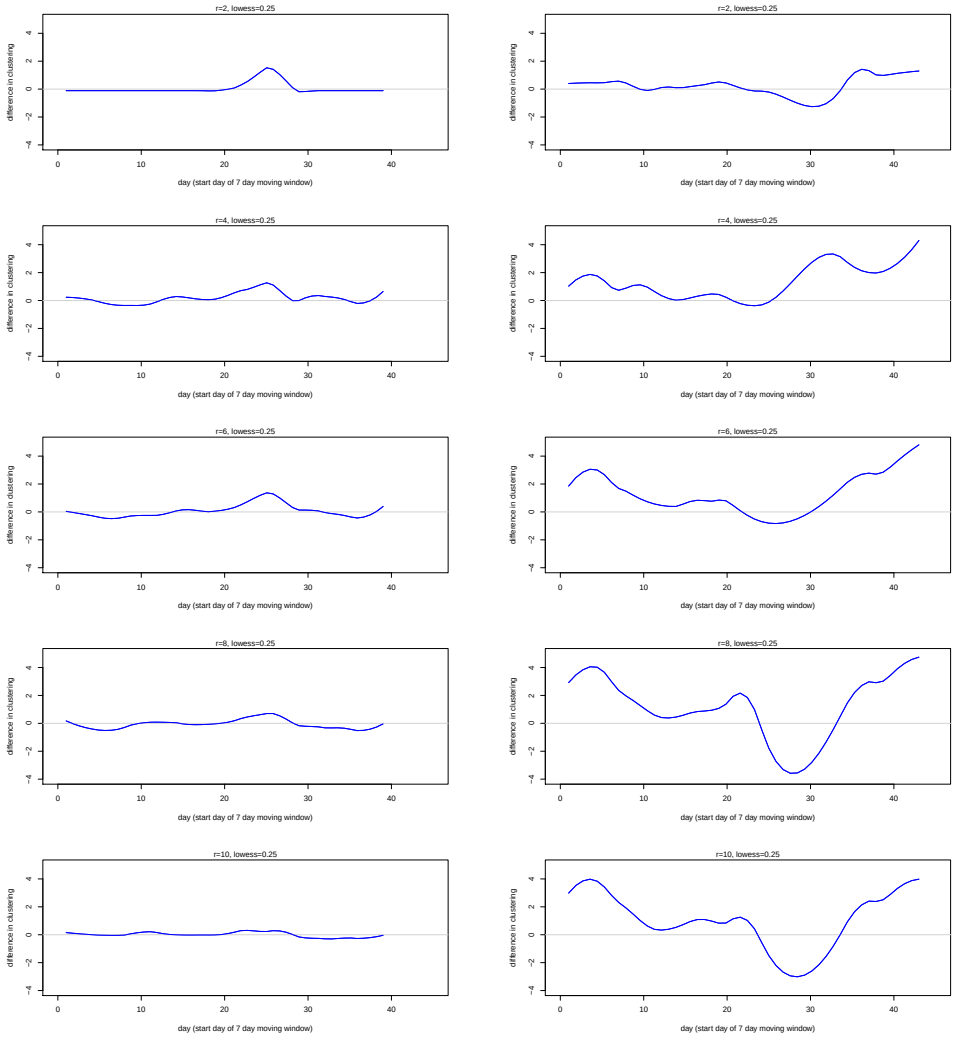


Figure 5.14: Cluster curve for spatially homogeneous datasets at a radius or 2,4,6,8 and 10km. On the left is the Poisson point process (PPP) and on the right the cluster Poisson point process (CPPP). Plots are all on the same axes scale for comparability

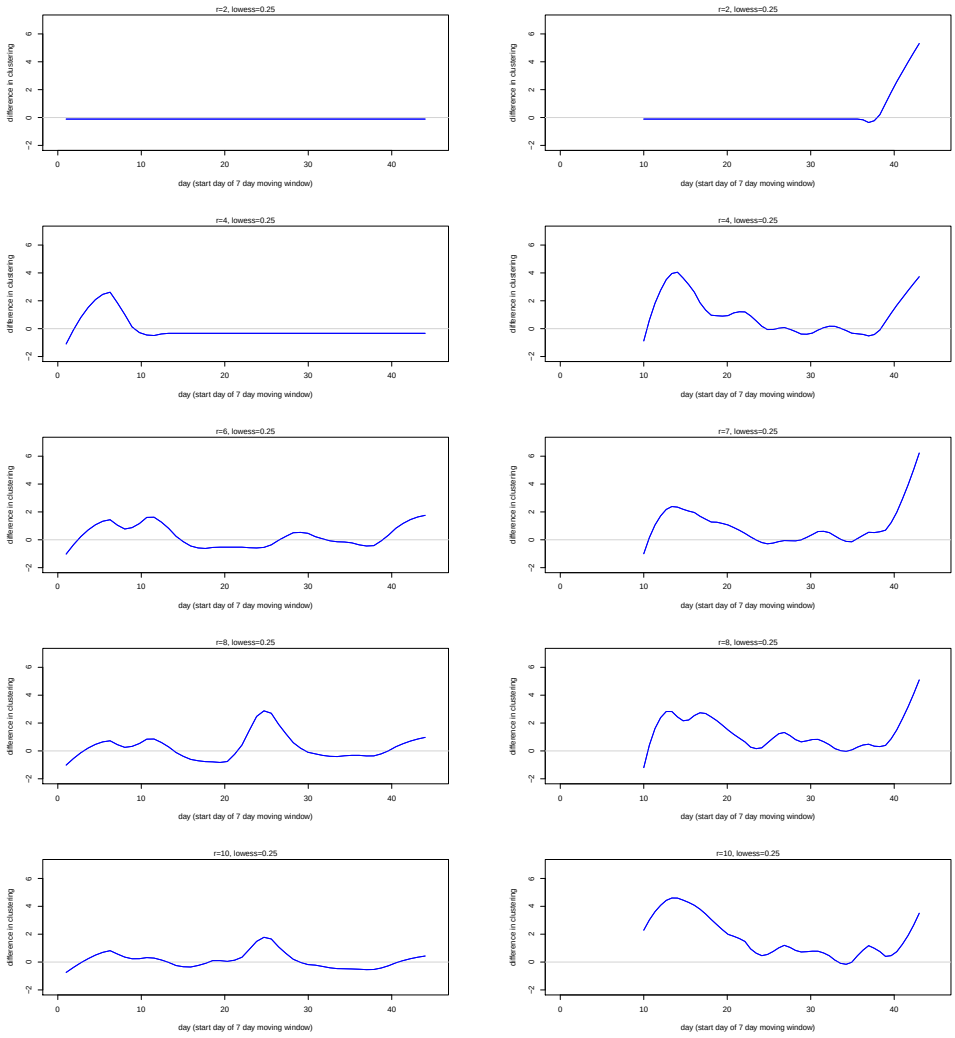


Figure 5.15: Cluster curve for spatially inhomogeneous datasets at a radius or 2,4,6,8 and 10km. On the left is the Poisson point process (PPP) and on the right the cluster Poisson point process (CPPP).

5.5 Integrated cluster curve

The ability to vary r is both a strength and a weakness of the cluster curve. It is a strength in that it allows the epidemiologist to examine clustering over ranges s/he thinks are important for the disease in question, but a weakness in that one might end up having to look at many plots, where the goal is to produce a simple tool for initial exploration of spatio-temporal epidemic data. In these circumstances it may be more interesting to look at the degree of clustering difference over a given range of radiuses. For this we propose the integrated cluster curve.

Our integrated cluster curve tool is a variation of our cluster curve tool in which we look at the difference in the spatial clustering through time over a given range of radiuses $[0, r_{\max}]$ rather than for a single radius. More specially our preliminary definition is:

$$\Psi_{int} = \int_0^{r_{\max}} (K_{obs}(r|t) - K_{theo}(r|t))\,\mathrm{dr} \tag{5.7}$$

where for every time interval we evaluate the degree of clustering by taking the area between the two curves over a specified range of radiuses. This is illustrated in 5.16, where we evaluate the integrated cluster curve over the range $[0, r_{\max}]$ for $r_{\max} = 10$km (or 1 unit). Equation 5.7 is evaluated using numerical integration.

The initial definition (5.7) of the integrated cluster curve is very dependent on the spatial scale of the data coordinates and the size of the epidemic outbreak. We therefore, scale by dividing the integrated difference by something. There are many methods in which this can be carried out, with no one that is clearly superior. Here we suggest three possible methods:

$$\Psi_{int} = \frac{\int_0^{r_{\max}} (K_{obs}(r|t) - K_{theo}(r|t))dr}{\int_0^{r_{\max}} K_{theo}(r|t)dr} \tag{5.8}$$

$$\Psi_{int} = \frac{\int_0^{r_{\max}} (K_{obs}(r|t) - K_{theo}(r|t))dr}{\int_0^{r_{\max}} \frac{1}{2}(K_{obs}(r|t) + K_{theo}(r|t))dr} \tag{5.9}$$

$$\Psi_{int} = \frac{\int_0^{r_{\max}} (K_{obs}(r|t) - K_{theo}(r|t))dr}{\int_0^{r_{\max}} (1 + K_{theo}(r|t))dr}. \tag{5.10}$$

Equation 5.8 takes the area under the theoretical curve as the denominator, equation 5.9 takes the average, and equation 5.10 has the addition of one, similar to the method applied to the cluster curve for fixed radius. Which method is chosen does not have a great impact on the conclusions drawn as we are mainly focused on the shape of the curve not the scale on the y-axis. Nonetheless, the scale can be of interest when comparing cluster curves (e.g. for the same disease over different regions). We proceed using method 5.8. Similar to the cluster

curve, we plot the standardised area for every time window which creates our integrated cluster curve.

The range of integration is generally chosen based on prior knowledge of the disease, for example the radiuses covered by protection and surveillance zoning. In particular we generally set the maximum radius, $r_{\max}$ to a maximum buffer zone that is thought to be appropriate to contain local spread. For example with FMD the FAO suggests a surveillance zone to cover a minimum of a 10km radius, but this will vary depending on factors such as geographical locations (Geering and Lubroth, 2002).

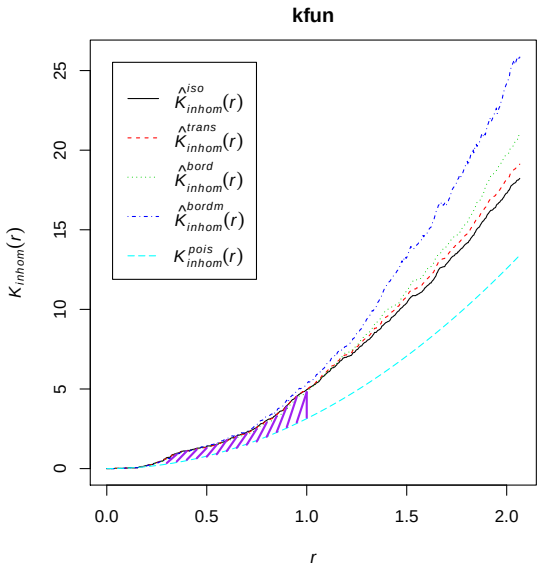


Figure 5.16: Inhomogeneous K-function example, showing the area difference between the theoretical and the observed inhomogeneous K-function curves for a range of 0 to 1 units (0-10km).

5.5.1 Application of the Integrated Cluster Curve to simulated data

Homogeneous population

The application of the integrated cluster curve to the homogeneous data is shown in Figure 5.17. We evaluate over the radiuses less than 5km (top) and less than 10km (bottom) for our PPP (left) and CPPP (right) datasets.

If we evaluate the difference over the area for both 5 and 10 km, we can see that there is no clustering for the PPP datasets.

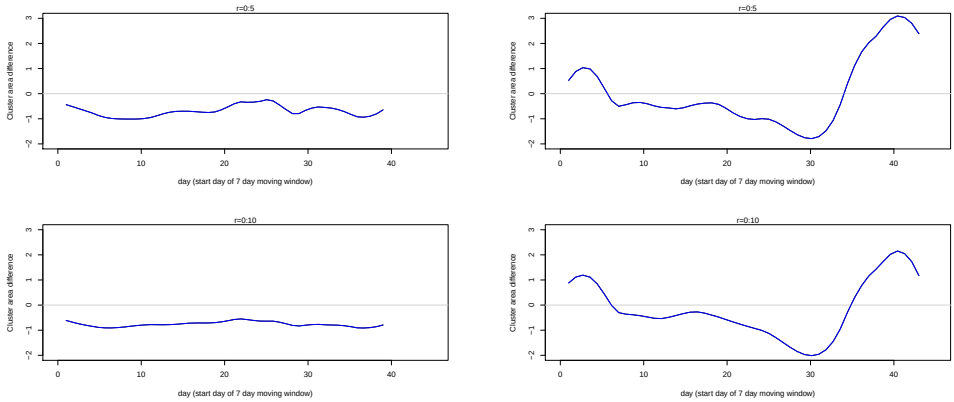


Figure 5.17: Integrated cluster curve for spatially homogeneous datasets. On the left we see the PPP datasets and on the right CPPP. The top plots show the evaluation over a radius of less than 5km and the bottom plots over the radius of less than 10km.

For the CPPP we see a higher degree of clustering, greater for the 5 km radius then the 10km, mainly at the beginning and end of the outbreak. We also see a clustering structure similar to that which is seen in the cluster curve evaluations. The similar structure but to the lesser extent at 10km suggest that the real clustering is largely within 5km, so the larger radius just waters the effect down.

Inhomogeneous population

The application of the integrated cluster curve to the homogeneous data is shown in Figure 5.18. As with the homogeneous examples we evaluate the datasets over the radiuses less than 5km (top) and less than 10km (bottom) for our PPP (left) and CPPP (right) datasets.

For the inhomogeneous datasets we see a greater degree of difference between the theoretical and observed curves in comparison to the homogeneous cases. Once again we see very little presence of clustering for PPP datasets at both radius range and for the CPPP dataset we see a comparable shape to the individual radius curves, with a large degree of clustering at the beginning and the end of the outbreak.

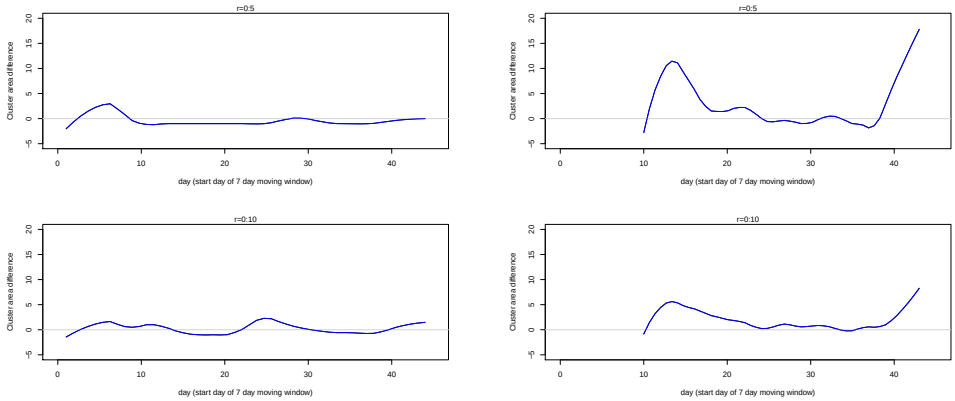


Figure 5.18: Integrated cluster curve for spatially inhomogeneous datasets. On the left we see the PPP datasets and on the right CPPP. The top plots show the evaluation over a radius of less than 5km and the bottom plots over the radius of less than 10km.

5.6 Real world application

5.6.1 England 2001 FMD outbreak

A full description of the epidemic and the data is provided in section 3.3.

We focus our attention purely on the cases that occurred, ignoring the pre-emptively culled farms. The spatial distribution of the confirmed cases in the 2001 epidemic in Northern England is shown in Figure 5.19. Along with all treated farms, information was available on all farms with susceptible animals within Northern England and the population on each of these farms. From this we were able to create an estimated farm density to be used as our spatial intensity function $\lambda_s(\boldsymbol{x})$. We did this by applying kernel smoothing over the spatial location of the farms. The bandwidth selected for the English farm population was 8km, with the resulting farm density shown in Figure 5.20.

Figure 5.21 shows the daily epidemic curve as well as the 7-day moving window. We can see a steep increase in the number of cases that peak at around 50 days. This then follows a steady decrease with a long drop off and a few small outbreaks.

The application of the cluster curve to the English data is shown in Figure 5.22. We look at a radius range of 2km to 10km in 2km increments with a lowess smoother where $\alpha = 0.3$. We see the greatest values of clustering occurring at a radius of 2 km and that later in the epidemic when the number of new cases is less, the degree of clustering is at its greatest. The curve also indicates clustering at all times (always above zero). Clustering at short ranges

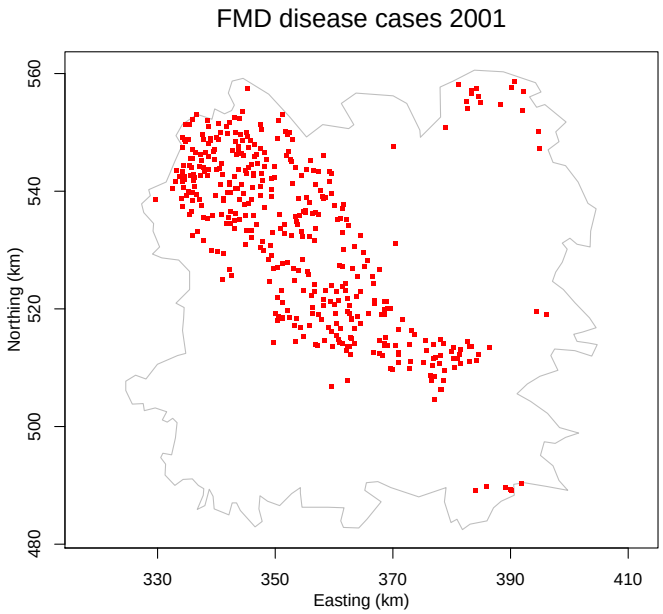


Figure 5.19: Spatial distribution of England 2001 FMD cases.

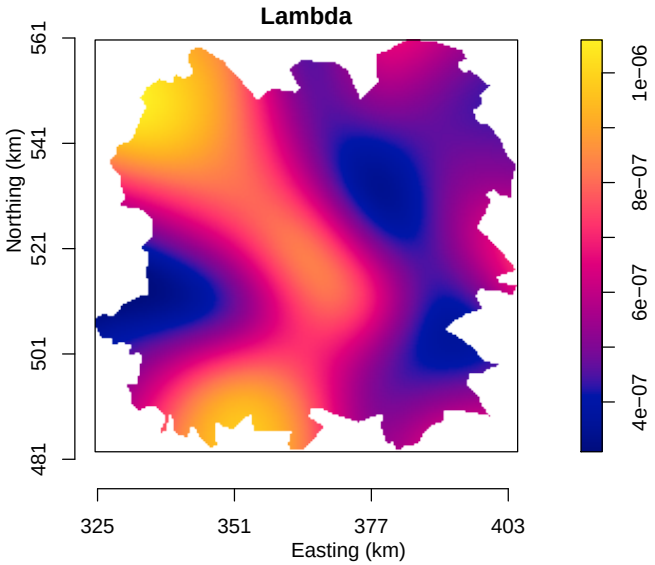


Figure 5.20: Farm density in Northern England

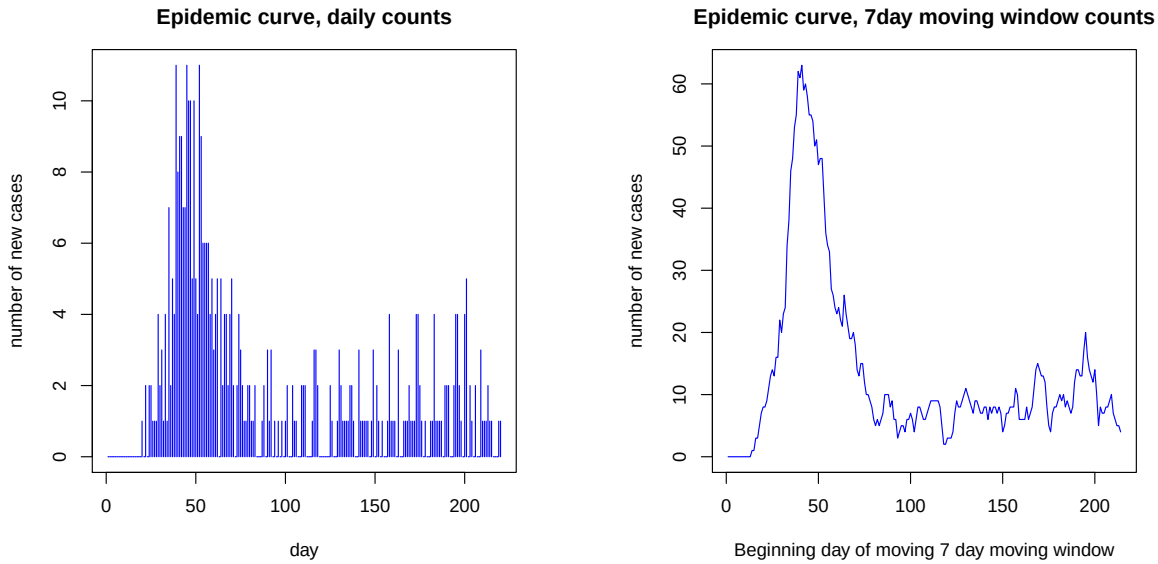


Figure 5.21: Epidemic curve of Northern England 2001. Left: daily epidemic curve, right: 7-day moving window epidemic curve.

fits in with what is known about FMD transmission. We know from previous analysis that 50% of new infected premises (IPs) were located within 3km, and 80% within 10km, after the non-movement ban. Also all farms within 1.5km of an IP was pre-emptively culled for control processes. This could suggest why we have a steep decrease in the degree of clustering from 4km onwards: the clustering is mainly occurring within 1.5km (picked up by a radius of 2km) before culling and this effect is watered down when we look at larger radiuses. We must be cautious about over-interpreting the big bump around 100 days, since that is when the data are sparse and may in part be noise.

The application of the integrated cluster curve to the English data is shown in Figure 5.23. We look at a radius range of 0:5km and 0:10km. These where chosen as previous analyses of FMD outbreaks, suggested a protection zone of 3-5km and a surveillance zone of 10km (Shimshony, 1988; Bergevoet and van Asseldonk, 2014). For both radius ranges we see a similar structure: a higher peak shortly after 100 days and a later, wider shorter peak after 150 days. We see the greatest degree of clustering in the later stages of the outbreak when the number of new cases has reduced. This suggests that the cases occurring were clustered and that the latter stages of the outbreak and spread of the disease was due to local transmission. However, caution must be taken as this could just be due to noise, reflecting the small number of cases.

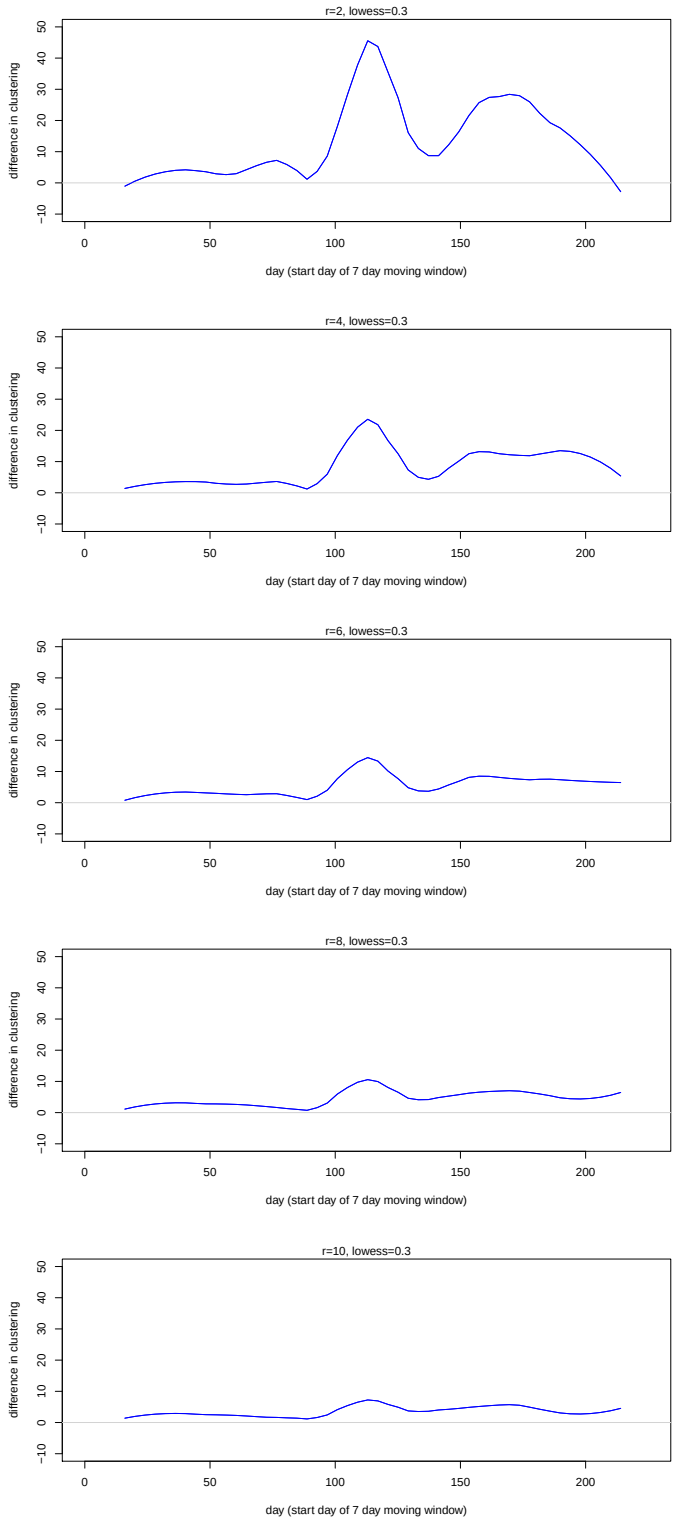


Figure 5.22: Cluster curve for FMD outbreak in England 2001 at 2,4,6,8 and 10km radius

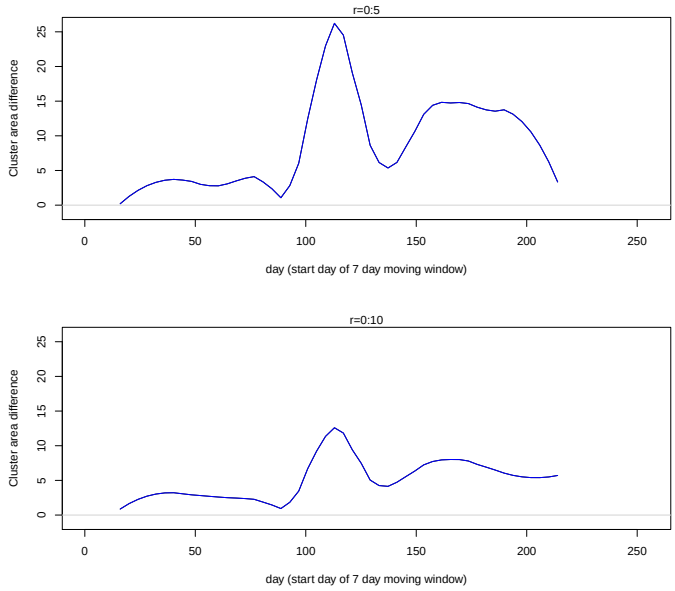


Figure 5.23: Integrated clustering curve for FMD outbreak in England 2001 0:5km and 0:10km radius

5.6.2 Japan 2010 FMD outbreak

A full description of the 2010 outbreak of FMD in Japan is provided in section 3.4.

We focus our attention on all confirmed cases of FMD in Miyazaki, ignoring the ring vaccinated farms. The spatial distribution of these events are shown in Figure 5.24. Similar to the English outbreak, we have information on all susceptible farms within our study region. We used these locations to obtain a kernel estimate of the spatial intensity function $\lambda(\boldsymbol{x})$. This was implemented using a bandwidth of $h_x = 1.3$km (smoothing in the x direction) and $h_y = 2.3$km (smoothing in the y direction), chosen using Scott's rule of thumb bandwidth selector (Scott, 1992). The resulting intensity function is shown in Figure 5.25.

We can evaluate the temporal pattern of the outbreak by looking at the epidemic curves shown in Figure 5.26. On the left we have a daily temporal scale and on the right the 7-day moving window. We can see that the initial spread of the disease was slow, but once several events occurred, the disease spread rapidly.

The application of cluster curve is shown in Figure 5.27 at radiuses of 2, 4, 6, 8,and 10 km. From these plots we can clearly see that the disease events were highly clustered at small radiuses, suggesting a large amount of local spread, with minimal long range transmission. In evaluating the Japanese epidemic there are three key dates. On day 20 a method of culling

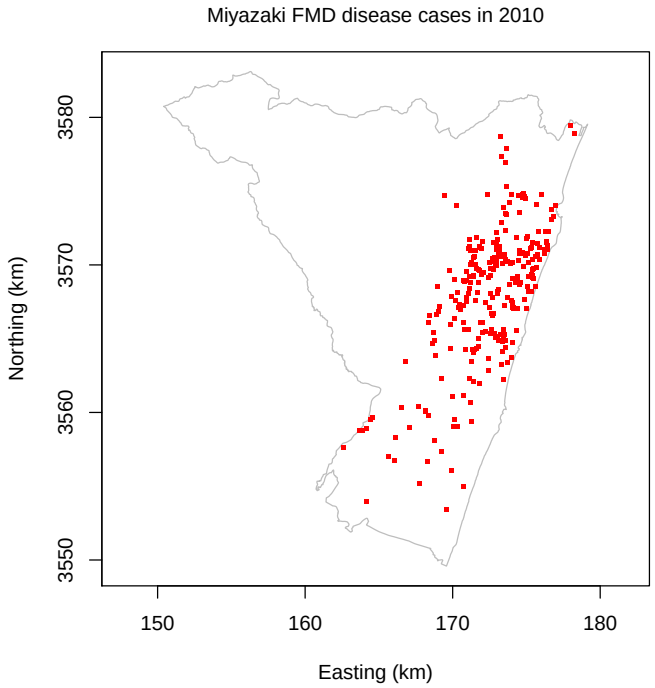


Figure 5.24: Spatial distribution of Japan 2010 FMD outbreak

and disinfection of infected premises was implemented, on day 48 a state of emergency was declared and on day 52 a method of 10km buffer ring vaccination was implemented. The outbreak had an intervention time lag of 11 days between notification to the completion of culling and disinfection (Nishiura and Omori, 2010). If we look at the spatial clustering at 2km we see that by the time the vaccination interventions were implemented the local spread seems to be under control.

The application of the integrated cluster curve is shown in Figure 5.28. We see that there is a high degree of overall clustering, which is particularly noticeable in the version of the curve with the 5km maximum. The initial stages of the outbreak appear to characterised by a clustering pattern, with the later stages of the epidemic appearing to be characterised by sporadic cases, rather than significant local outbreaks.

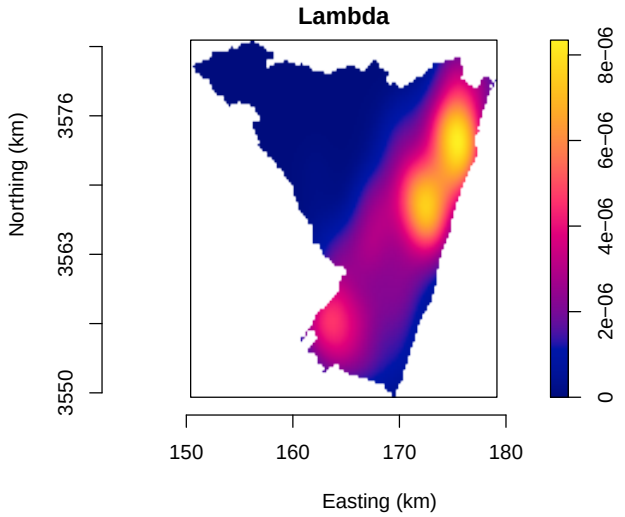


Figure 5.25: Estimated farm density in Miyazaki, Japan

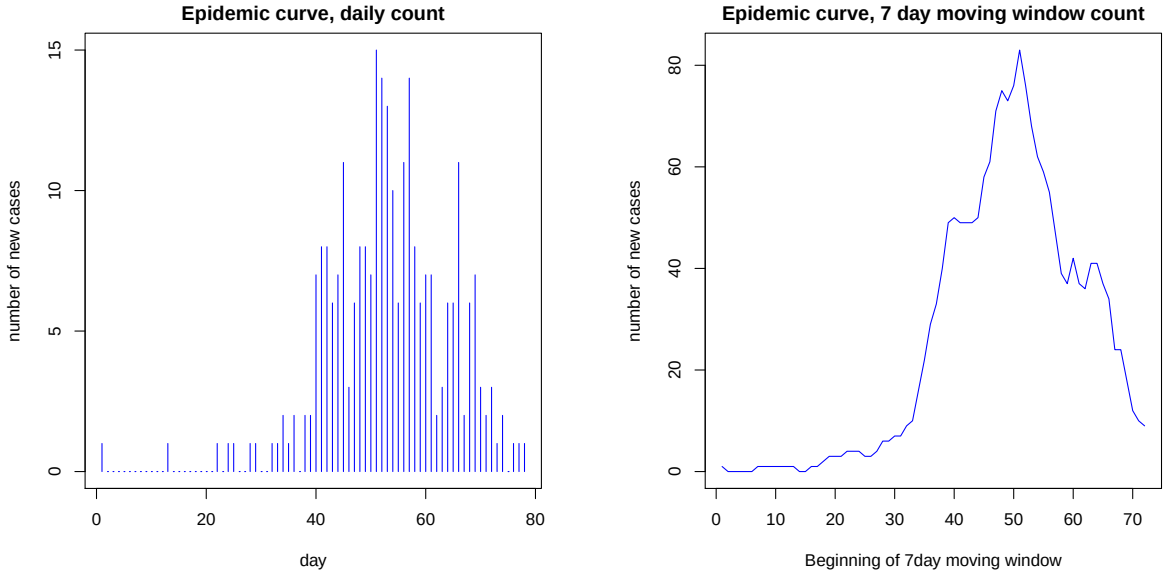


Figure 5.26: Epidemic curve of Miyazaki, Japan 2010. Left: daily epidemic curve, right: 7-day moving window epidemic curve.

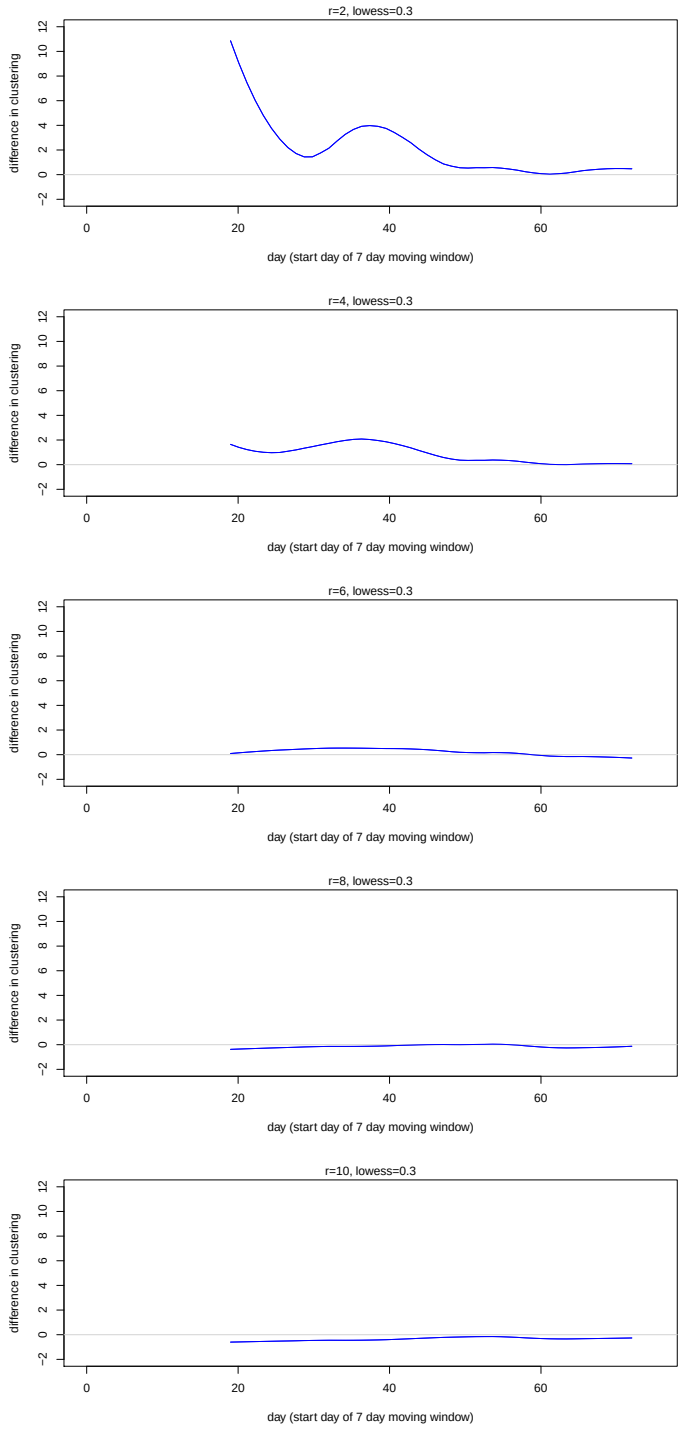


Figure 5.27: Clustering curve for FMD outbreak in Miyazaki, Japan 2010 at 2,4,6,8 and 10km radius

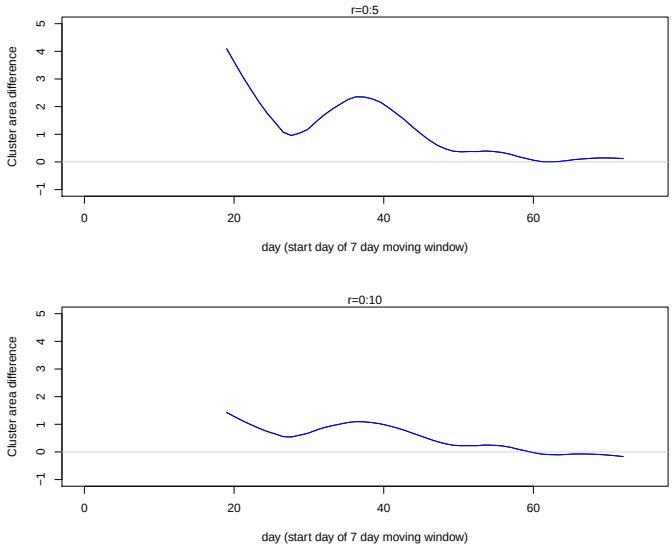


Figure 5.28: Integrated clustering curve for FMD outbreak in Miyazaki, Japan 2010 0:5km and 0:10km radius

5.7 Discussion

When analysing epidemiological data it is common practice to use epidemic curves to examine the temporal patterns amongst the data. They can provide great insight into likely mechanisms of disease spread and to the well trained eye can hint at the presence of clustering. However, clustering of diseases can be subtle and complex, making it difficult to draw inferences about clustering and spatial transmission in many situations.

The presence of clustering commonly occurs with disease as they generally are infectious or point source in nature. With infectious diseases the failure to consider and/or investigate the presence of clustering could hinder methods to control and eradicate the disease.

Our tools, the cluster curve and the integrated cluster curve, describe clustering in point process patterns through second order analysis; more specifically the use of the inhomogeneous K-function. Our tools evaluate the presence of spatial clustering through time by calculating the difference in the theoretical and the observed levels of clustering within the data in subintervals in time. The cluster curve is not aiming to detect clusters of disease in space and time, instead it is a simple means of describing how the pattern of spatial clustering varies with time.

To illustrate the use of our tools we created two sets of pairs of datasets, the first pair based on a spatially homogeneous underlying population and the second on a spatially inhomogeneous population. For each of these we created two datasets, one a Poisson point process and the other a cluster Poisson point process. Within each underlying population for the two datasets we selected the datasets based on similar epidemic curves. For both populations the epidemic curves offered little insight into the spatial distribution of the data. From the application of our cluster curve and integrated cluster curve we can see that the spatial structure of the datasets are very different, with the presence of clustering correctly identified for our clustered Poisson point processes only.

In our application to the real world cases we evaluate the the changes in spatial clustering through time of FMD, an infectious disease. These cases suggest the presence of spatial clustering through time. Definitive conclusions can be difficult because of the possibility of quite big bumps occurring largely by chance in time periods with sparse data. In the following chapter we will look into extension of the cluster curve including the presence of significance indicators.

In both the real world situations, intervention strategies were applied to the control the outbreak. This raises the question; how do intervention methods affect the clustering properties of an outbreak? We will explore this in chapter 7.

Chapter 6

Extensions to the cluster curve

6.1 Introduction

In the previous chapter we discussed the motivation and the methodology of our tool, 'the cluster curve', to be used alongside epidemic curves to investigate the change in spatial clustering through time. However, to improve the utility of the cluster curve for the epidemiological research community there are two major things that we need to address: indicators of statistical significance to aid interpretation of the curve, and ease of use for those uncomfortable with a command line package like R.

One of the limitations we faced with our original cluster curve is the possibility for over-interpretation when data is sparse, therefore, making it difficult to distinguish between true clustering and noise. Here we will investigate the use of pointwise envelopes to address whether the difference between the theoretical and observed curves is statistically significant, indicating that the data is more clustered then expected.

Secondly in this chapter we will consider the functionality of our tool. It is one thing to create a method to investigate clustering through time, but we want our tool to be usable to a wider research community (with no coding required by the user). To that end we will describe our use of the *shiny* R package to make our tool a functional URL webpage. We do this by providing two examples, one with our simulated datasets and the other where users can upload their own datasets.

6.2 Highlighting significance in the cluster curve

Our cluster curve is a measure of spatial clustering through time. In some situations it will be difficult to tell whether bumps in the cluster curve indicate important clustering from local disease transmission, or are primarily noisy artefacts. For example, in Figure 6.1 it appears that there is the presence of clustering between 20-30 days but this is based only on 4 data points at its peak so may well reflect just random variation. Therefore, in many circumstances it will be helpful to provide a method to assess statistical significance. We do this through the application of envelopes for the inhomogeneous K-function. These are estimates of the K-function from simulated datasets based on the theoretical under the assumption of no clustering. See Baddeley et al. (2014) for a general discussion of the use of simulation envelopes for testing for point process models.

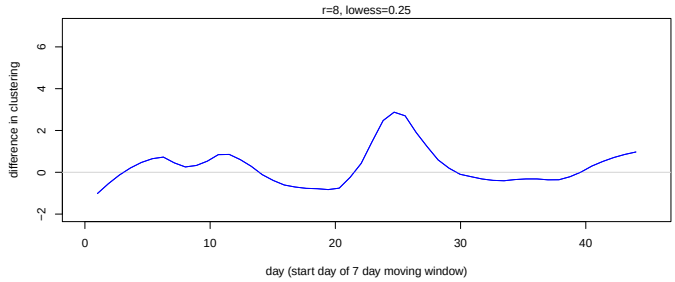


Figure 6.1: Cluster curve inhomogeneous PPP 'bump' example

6.2.1 Methodology

To create our significance bands we use pointwise envelopes. For any given value of r, these work by using suitable quantiles to define lower and upper confidence limits for the K-function under the assumption of no clustering. Values outside this range are deemed statistically significant. In more detail, for each time window we start by simulating M datasets from an inhomogeneous Poisson process with intensity function defined by the population (spatial) density, $\lambda(\boldsymbol{x})$ scaled to produce patterns with the same mean number of points over the region as are in the observed pattern over the window. For each of these simulated datasets we compute the inhomogeneous K-function. See Figure 6.2 for an example.

For each radius r over a grid of values we then compute suitable quantiles from the set of K-function values $\{K^{[i]}(r|t)\colon i = 1, \ldots, M\}$, where $K^{[i]}$ is the function obtained from the ith simulated dataset. For instance, for a 95% significance level we would compute 2.5% and 97.5% quantiles from the set of $K(r|t)$ values (see Figure 6.3). These quantiles then define

the limits of 'normal' behaviour under the assumption of non-clustering. Let $K_{obs}(r|t)$ denote the inhomogeneous K-function estimated from the observed data in time window t. Then if the value $K_{obs}(r|t)$ breaches the limits, this indicates statistically significant clustering (if beyond the upper limit) or statistically significant repulsion (if below the lower limit) at time t. In general we will be interested in the first of these.

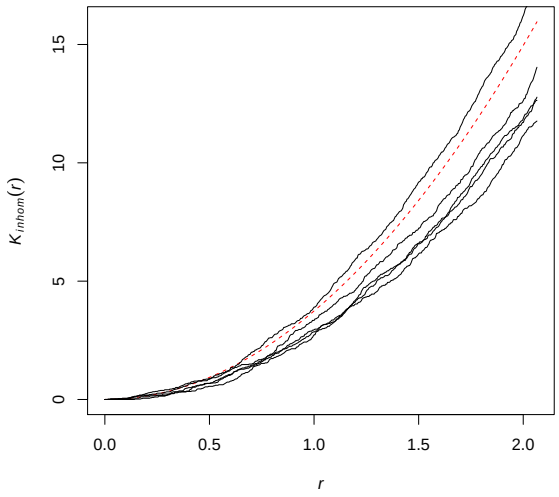


Figure 6.2: Inhomogeneous K-function example: dashed red line represents a theoretical inhomogeneous K-function based on the data, and the black solid lines K-functions derived from 5 simulated datasets (under the assumption of no clustering)

Implementation in R

We implemented our significance indicators using the function *envelope* in the R package spatstat (**?**). We specified the number of simulations to 99 and the desired significance level to 5% for a breach of the corresponding upper envelope quantile. The significant clustering indicators were coded to appear as red dots for each time window, as illustrated in the next section.

6.2.2 Application of cluster curve with significance marks to simulated data

In Figure 6.4 we demonstrate the use of the significance indicators for the spatially homogeneous CPPP dataset. Here we can see the same curves as seen in 5.14 but this time have the

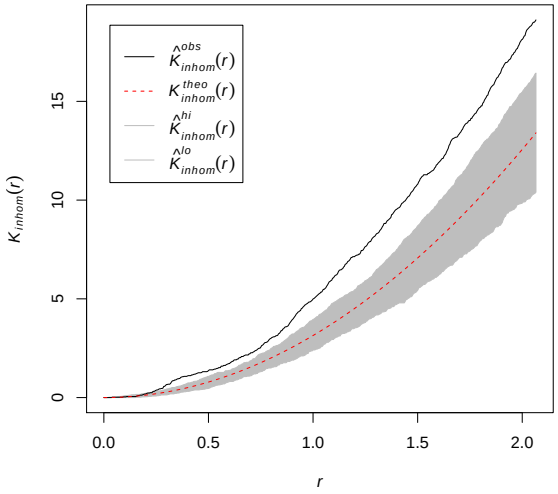


Figure 6.3: Inhomogeneous K-function example for the translate correction

presence of red dots to indicate those time windows for which the clustering is significant. We can see that the majority of positive differences are significant, particularly at the beginning and end of the outbreak. This suggests that when the data appears to be clustered, this clustering is significantly different to what is expected for a simple (un-clustered) Poisson process. The largely negative values in the latter around days 25-30 occur when the there are no new cases recorded (epidemic curves shown in Figure 5.9). One apparent anomaly is the presence of an indicator marker at $t = 24$ for the $r = 2$ cluster curve, even though the curve is marginally negative at that point. This is an artefact of the lowess smoothing procedure. The unsmoothed value of the curve is positive at that time point.

In Figure 6.5 we demonstrate the curve for the spatially inhomogeneous CPPP dataset. We can see that the initial and the latter stages appear to be significantly clustered, with this more prominent in the larger radiuses. The peak number of cases coincides with periods of reduced levels of clustering (epidemic curve Figure 5.10). Again, the red dot at an apparently negative value of the cluster curve at time $t = 37$ for the $r = 2$ cluster curve is an anomaly caused by the lowess smoother.

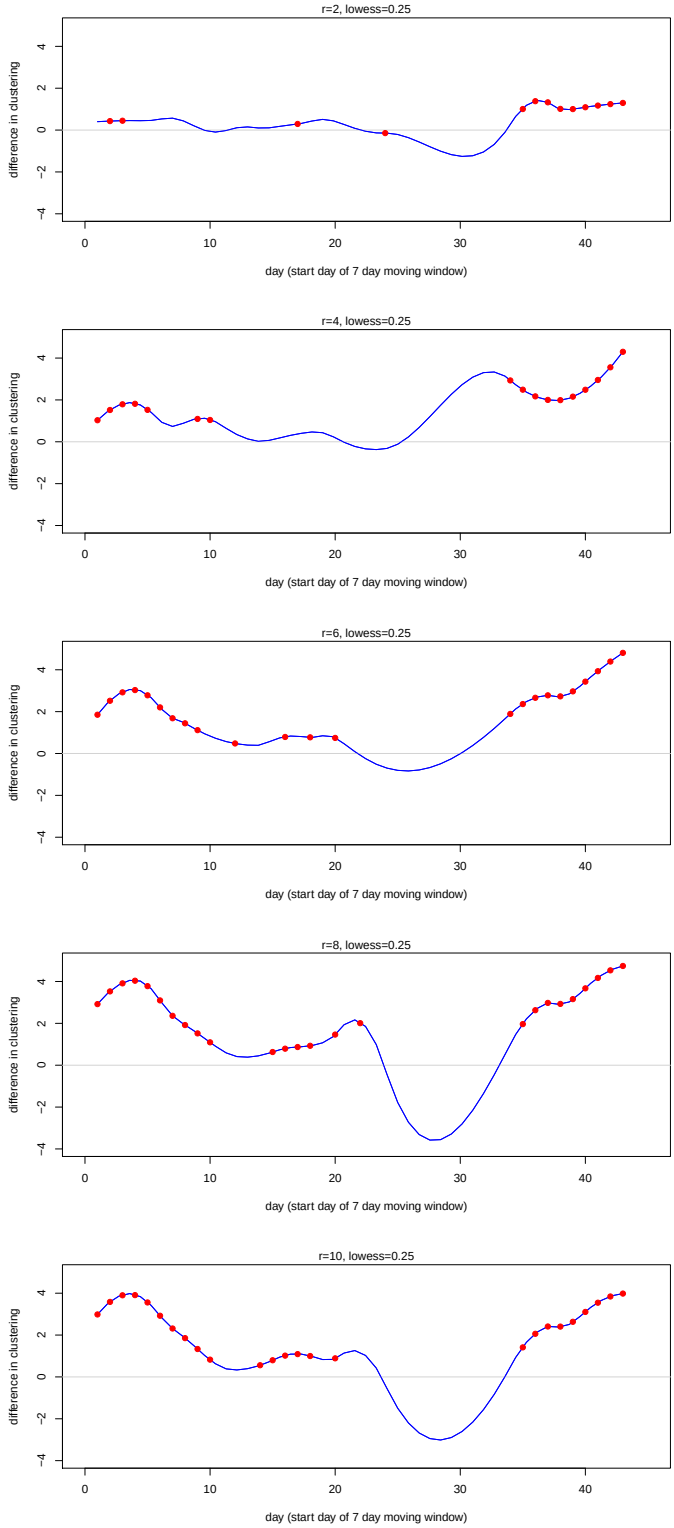


Figure 6.4: Cluster curve with significance dots for spatially homogeneous CPPP dataset at a radius or 2, 4, 6, 8 and 10km.

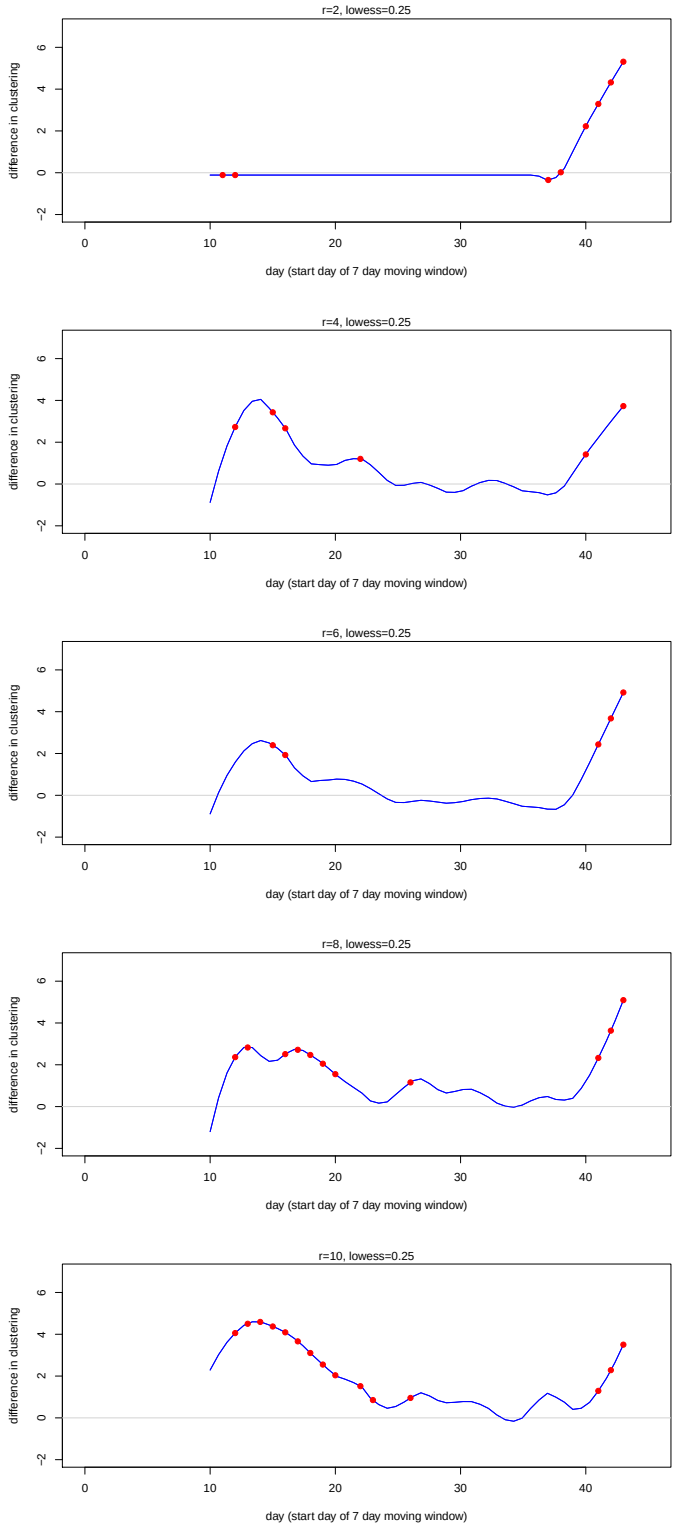


Figure 6.5: Cluster curve with significance dots for spatially inhomogeneous CPPP dataset at a radius or 2, 4, 6, 8 and 10km.

6.2.3 Real World Application

England FMD 2001

The application of the enhanced cluster curve to the 2001 FMD outbreak in England is shown in Figure 6.6 and at a 2km radius paired with the epidemic curve in Figure 6.7. We can see that the majority of the FMD outbreak in Northern England was characterised by a significant degree of clustering. What is interesting to see is that even at some time points where the cluster curve takes quite small (positive values), the degree of clustering is significant. These significance at the small degrees of difference match with the peak number of cases of the epidemic curve. Overall the outbreak appears to be characterised by local spread. In chapter 5 we were concerned with the over-interpreting of the large bump after 100 days, as the data was sparse, here we can see that the cases in these time windows appear to be significantly clustered.

Japan FMD 2010

We look again at the FMD outbreak in Japan and apply our cluster curve with significance dots. The results are displayed in Figure 6.8 with the radius of most significance paired with the epidemic curve in Figure 6.9. We see that most of the clustering occurred at the lower radius and only one time point on the cluster curve appears to be statistically significant. With an isolated significant point like this, we should bear in mind that it may be a false positive result. Whatever the case, there is clearly far less clustering for the Japan outbreak than the English.

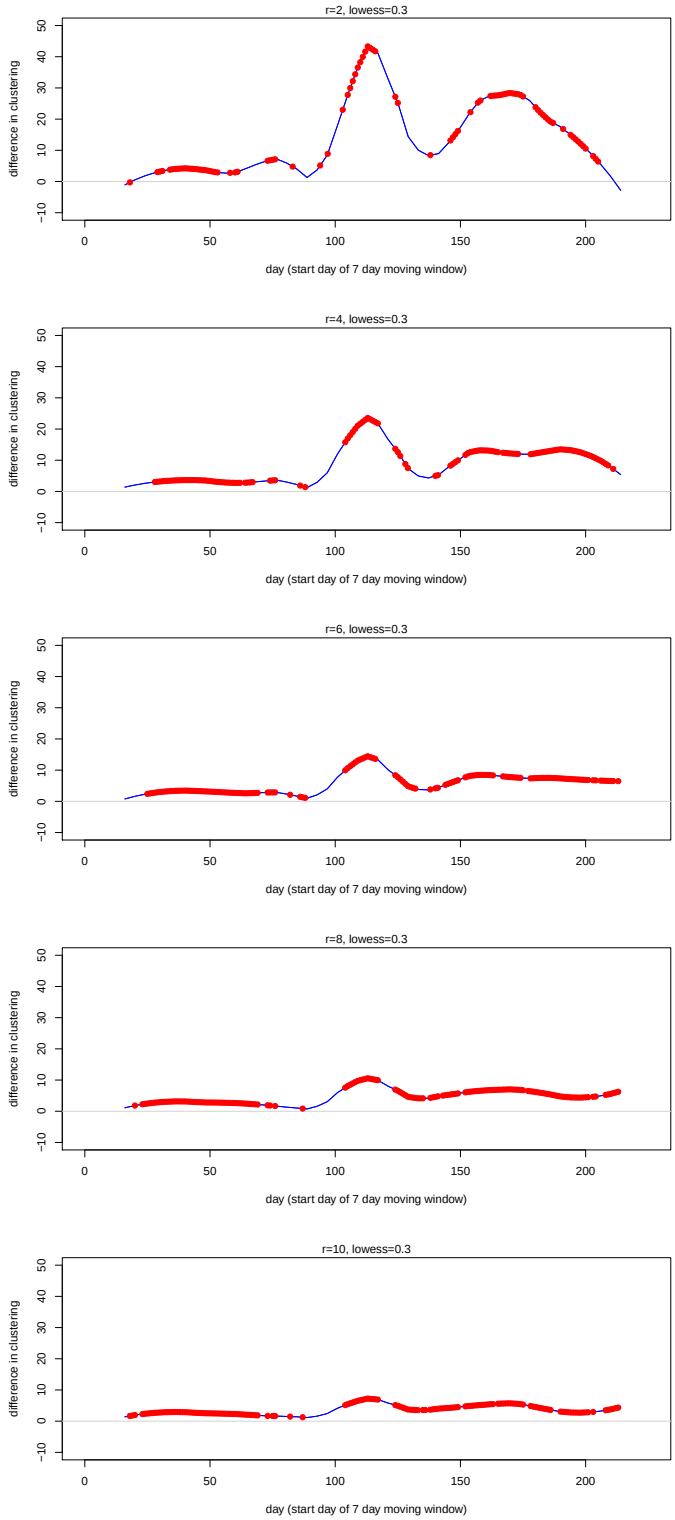


Figure 6.6: Cluster curve with significance dots for FMD outbreak in England 2001 at 2, 4, 6, 8 and 10km

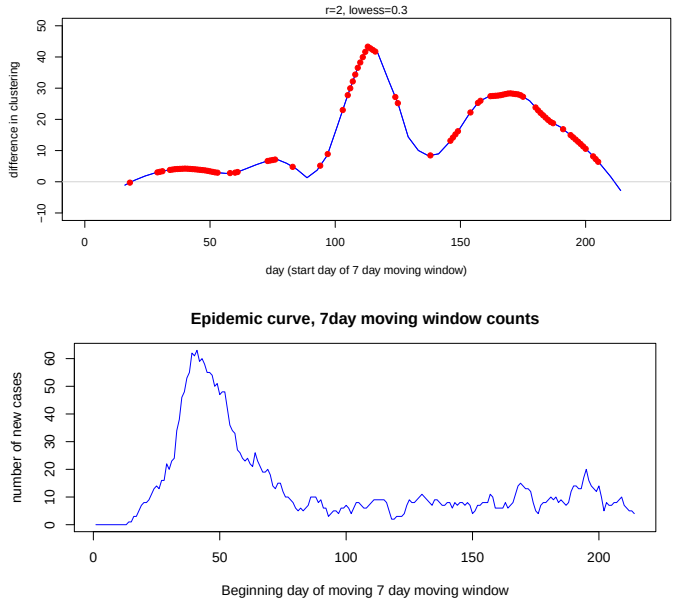


Figure 6.7: Epidemic curve and cluster curve for 2km radius of the English FMD outbreak

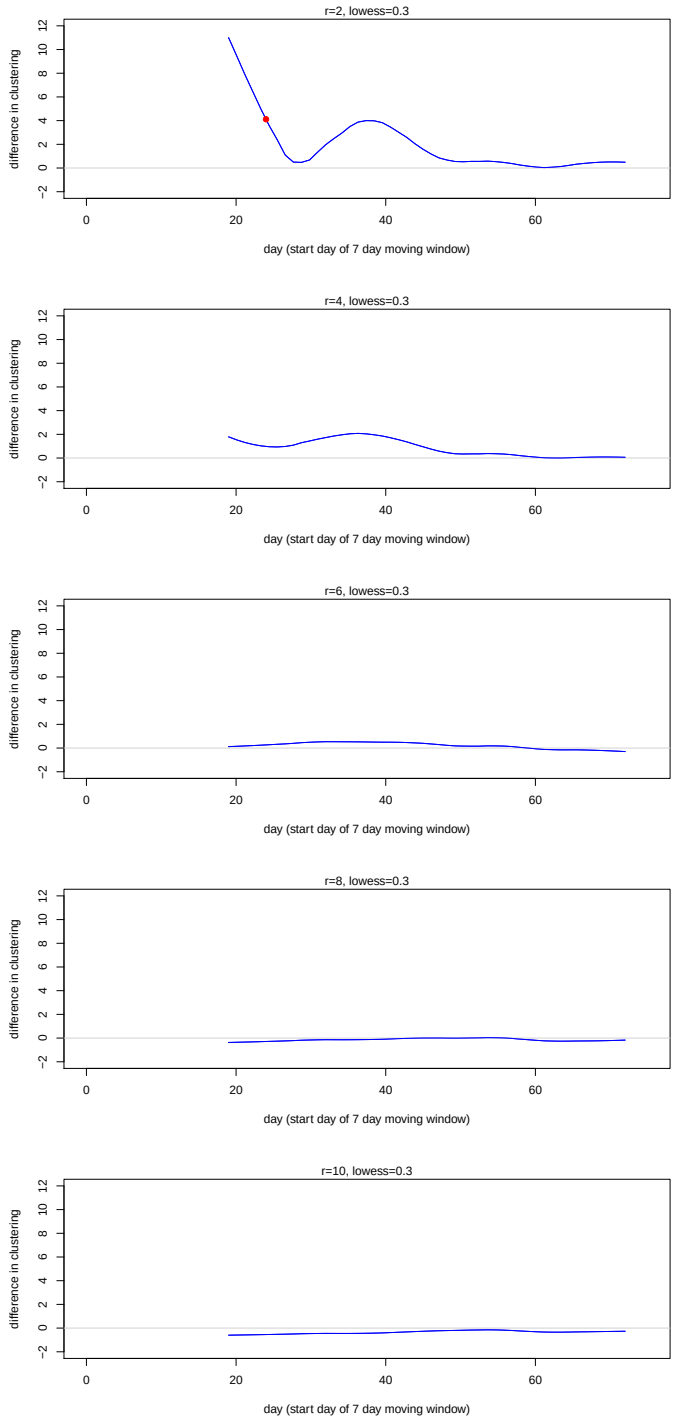


Figure 6.8: Cluster curve with significance dots for FMD outbreak in Miyazaki, Japan 2010 at 2,4,6,8 and 10km

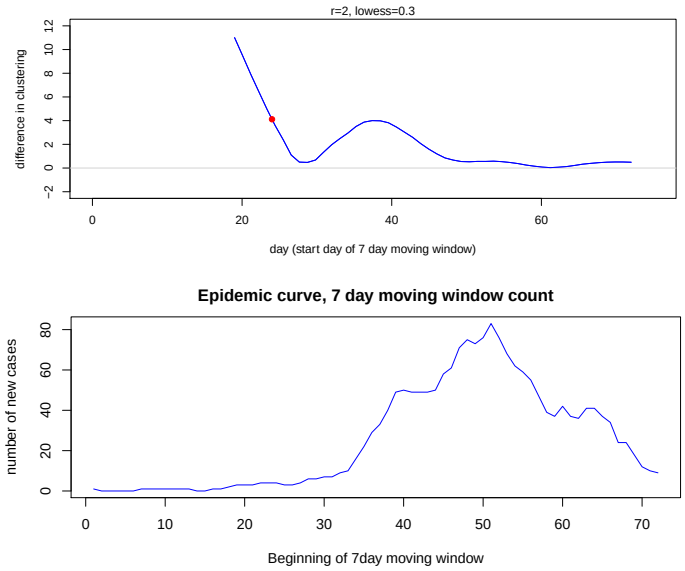


Figure 6.9: Epidemic curve and cluster curve for 2km radius of the Japanese FMD outbreak

6.3 Shiny application of the cluster curve

Shiny is an RStudio project (Rstudio, 2012) which allows the creation of a web application for R analyses. The *Shiny* package allows the combination of the computation power of R with the interactivity of the web, with no web development skills required to write them (Chang, 2015).

6.3.1 Implementation in R-studio

The *Shiny* package allows the creator to build two interfaces, one that is an interactive interface for the user and one that takes the input specified by the user and updates the R code that creates the output. The interaction of the user can be as small as choosing the plotting colour through to the user being able to upload their own data.

The *Shiny* package runs on three main files: *ui.R*, *server.R* and *helper.R*. The *ui.R* file supplies information on the app user interface design; what is present in the side and main panels; and titles, descriptions and so on. The *server.R* tells the app what each slider/button/input does, where the information comes from and how to create the display plot. The *helper.R* files provides all the background information including function code and required packages. There are several methods to host the web application. These include the use of Github (GitHub, 2015) which makes available the ZIP file that a user can then run in R, and shinyapps.io (by Rstudio, 2015) which is RStudio's hosting service for Shiny apps. We use the latter to publish our prototype cluster curve apps.

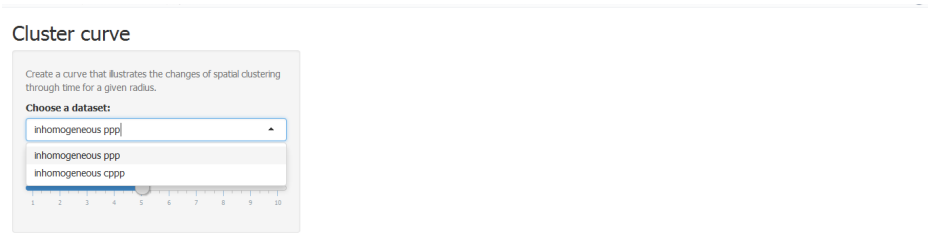
6.3.2 Cluster curve online prototypes

Here we showcase two online prototypes for the cluster curve tool. The first is the application of the cluster curve to the inhomogeneous datasets. The prototype allows for the selection of which dataset is to be used and then which radius to look at. The second prototype allows the user to upload their own data for application of the cluster curve.

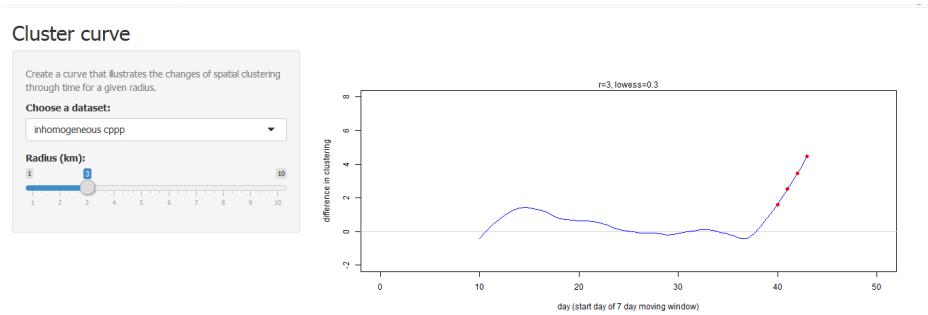
Inhomogeneous data example

Our first online prototype illustrates the application on the cluster curve to the two simulated spatially inhomogeneous datasets. The steps of this application are shown in Figure 6.10. The first step (a) is the selection of which dataset is to be used. This is chosen from a drop down

menu. The next step (b) is the selection of the radius; the default is set to 5km but this can be easily changed by adjusting the the slider. In this example, of step 'b', we have changed the slider to a value of 3km. Thus creating the cluster curve plot to the right of subplot(b).



(a) Data selection



(b) Cluster Curve

Figure 6.10: Cluster curve example

Upload own data

A visual representation of the steps to apply the cluster curve prototype to one's own data is shown in Figure 6.11. The process is as follows:

1. Upload the boundary file (subplot (a)). After clicking on 'choose file' an open box will appear. As a shape file contains multiple layers, within this open file, select all layers (.dbf, .prj, .shp, .shx). The current prototype is only set up for metric coordinate

systems and currently does not handle latitude and longitude.

2. Selecting the unit that the shapefiles are in (subplot(b)) presents two possibilities; m or km. This allows the program to make corrections as the selected radius is in km.

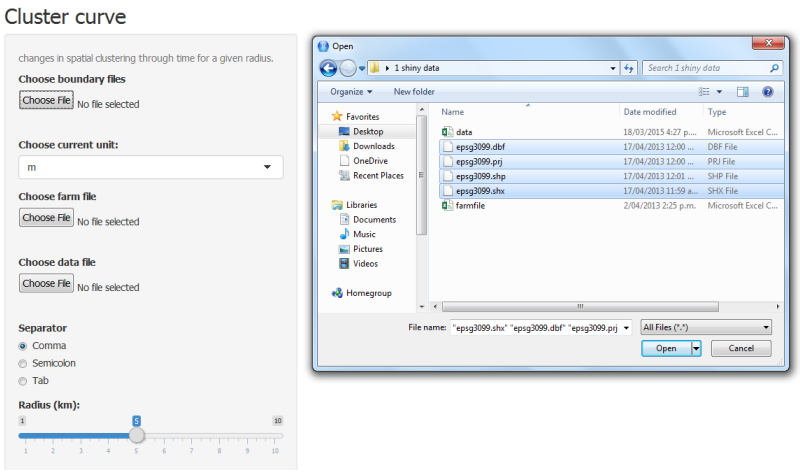
3. Upload the farm file. This can be in the format of CSV or text.

4. Upload the data file. Similar to the farm file this can be in the format of CSV or text. For the farm and data files the day of the outbreak needs to have the header "day" and the x and y coordinates "xcoord" and "ycoord" respectively.

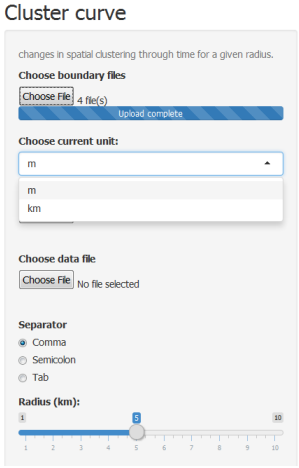
5. Specify the separators.

6. Once all datafiles are uploaded the program will run the function in the background, then produce the cluster curve plot (subplot(c)).

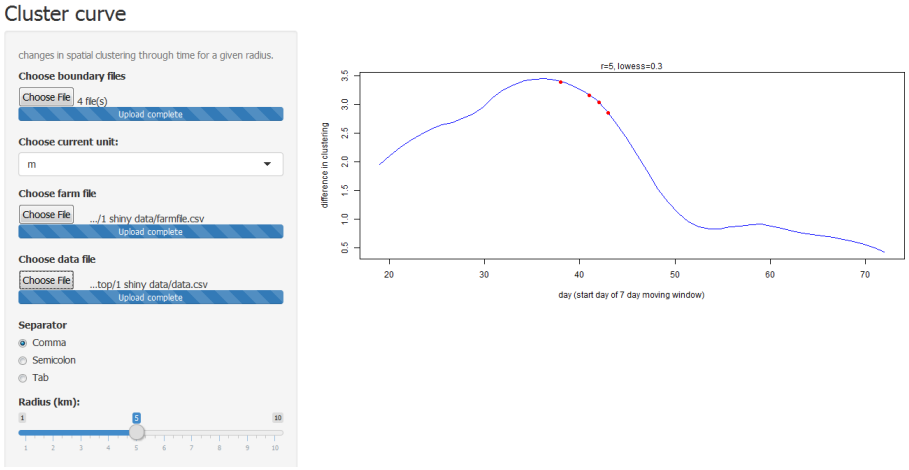
7. The radius at which we wish to evaluate the cluster curve can then be changed by adjusting the slider.



(a) Shapefile upload



(b) Unit selection



(c) Cluster Curve

Figure 6.11: Cluster curve prototype

6.4 Discussion

In chapter 5 we noted that one of the main limitations in our cluster curve tool was the inability to definitively differentiate between true clustering and noise due to small sample sizes within a given time window. In this chapter we looked at the inclusion of significance dots to try to reduce the influence of noise. The significance indicators were created using simulated envelopes. If the observed K-functions is greater than the upper limit value for the K-function simulated under the assumption of non-clustering, then the degree of clustering in that time frame for that radius is indicated as significantly clustered.

We illustrated the significance indicators on the simulated cluster point process data sets, where it proved effective at identifying the presence of clusters. We then applied the extended cluster curve to the real world outbreaks of FMD. Previously our application to the real world cases of FMD suggested the presence of spatial clustering through time. However, we found it difficult to make definitive conclusions because of sparse data. In this chapter we included the addition of significance indicators. From this we saw that the England outbreak appeared to be significantly clustered throughout the outbreak even when the data is sparse. In contrast we see that the Japanese outbreak of FMD does not seem to be significantly clustered. This suggests that the Japan outbreak was very effectively handled by veterinary authorities.

Lastly we discussed our tool that was developed for the use alongside the epidemic curve to identify changes in spatial clustering through time. In theory we want our tool to be used by a wide range of researchers, including those without a strong coding background. We therefore want to make our tool user friendly. To enable this we have created a prototype of the cluster curve and uploaded it to an online hosting platform. Here researchers can look at an example of the implementation of the cluster curve through our simulated inhomogeneous dataset examples, and secondly upload their own data and shape files to carry out the cluster curve analysis.

Chapter 7

Application to FMD intervention strategies

7.1 Introduction

In the real world, outbreaks are not left to run their natural course but are controlled in some form. For example, the 2001 outbreak of FMD in England was eradicated through the use of depopulation and pre-emptive culling, and the 2010 outbreak in Miyazaki was controlled using a method of depopulation and vaccination to eliminate FMD.

The method or combination of methods applied vary for diseases, countries and whether the interventions are designed to control or eradicate. Possible intervention methods include: slaughter, contact reduction, chemical use (disinfection, pesticides), vaccination, environment and/or management controls (husbandry practices, education) and no controls (Christensen, 2001). The methods chosen to control a disease can have dramatic economic consequences.

To examine the impact of intervention strategies on the changing patterns of spatial clustering through time we used InterspreadPlus (Stevenson et al., 2013) to create models of FMD outbreaks. InterspreadPlus is a software designed to provide a framework for modelling the spread of infectious disease. The intervention methods investigated included:

- No control (NC) where the virus was allowed to run its natural course,
- Depopulation (D) where FMD positive farms are culled,
- Vaccination (V) where a 5km buffer ring vaccination program is applied,

- Depopulation and vaccination (DV) where FMD positive farms are culled and a vaccination program applied,
- Depopulation and pre-emptive culling (DED) where FMD positive farms are culled and all farms within 5km of the infected farm are culled.

We looked at the effects of these various strategies using simulated outbreaks in Border counties, Great Britain (see section 3.5.2 for specification of which counties) and Miyazaki, Japan.

7.2 Simulated datasets

7.2.1 Scenarios

In this section we provide a more detailed account of the control strategies considered.

Depopulation

When implementing the depopulation intervention strategy only the farms that are confirmed to have FMD are treated. For this treatment all susceptible animals on a confirmed FMD positive farm are culled. No surrounding farms are immediately affected, unless later confirmed with FMD.

Vaccination

For the vaccination control method a 5km radius around an infected farm is instigated. The border of this radius is where vaccination of all susceptible animals starts. Vaccination is then applied in an inwards motion creating a buffer ring, which is used to try to prevent further local spread out of the 5km radius. An example of the buffer ring is shown in Figure 7.1.

Depopulation and vaccination

When we combine the depopulation and vaccination methods we proceed by culling the confirmed farm of all susceptible animals and then vaccination is applied, starting at a 5km radius and working towards the depopulated farm. This method combines an initial removal of the infection source with a method to limit spread by creating an immune buffer ring.

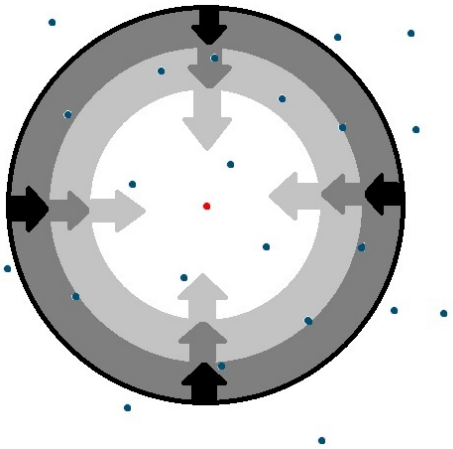


Figure 7.1: Vaccination buffer ring example.

Depopulation and pre-emptive culling

This method of control has the greatest impact on the animal population. For this method we once again use a 5km radius of intervention, but this time we work out from the centre of the reported case. The initial farm is depopulated of all susceptible species. Then all surrounding farms are depopulated working outwards until all farms within the 5km radius have been depopulated. The pre-emptive culling regime is shown in Figure 7.2.

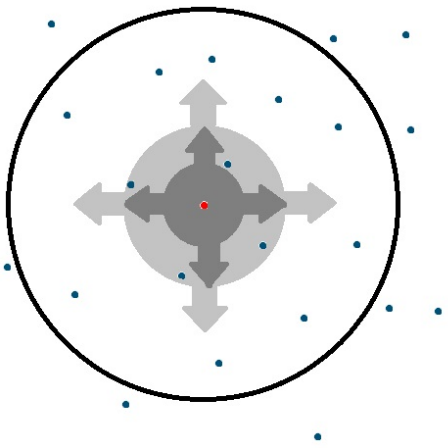


Figure 7.2: Pre-emptive culling example.

No control

In order to see the natural progression of the disease a dataset was simulated with the same epidemic history as if no interventions were applied. For this simulation we monitored the outbreak until the maximum time length of 100 days.

7.2.2 Data generation

The datasets were created using InterspreadPlus. We describe the process of creating an Interspread simulation in section 3.5.

Two sets of simulated outbreaks with the different intervention strategies applied were reviewed. One was based on the animal populations of Border counties, Great Britain and the other Miyazaki, Japan. The basic structure of these outbreak datasets are explained in section 3.5.2 for Border counties, Great Britain and for Miyazaki, Japan. The intervention strategies for the different places were specified identically.

Each simulation generated a complete outbreak of FMD with a maximum of 100 days monitored.

7.3 Border counties, Great Britain

The simulated Border counties datasets (NC, D, V, DV, DED) are explained in detail in section 3.5.2. We considered diseased farms to be our cases, and look at clustering based on these cases rather than including farms that have been preemptively culled, for example. Identifying the farms affected by intervention strategies as cases of disease, would mask the natural clustering of the outbreak, as these strategies would produce artificial clusters of premises surrounding diseased farms.

In the application of the cluster curve we once again estimate a farm density based on the location of all the farms. We apply kernel smoothing with a bandwidth of 8km to provide the inhomogeneous K-function with the underlying spatial intensity $\lambda(\boldsymbol{x})$ for the population. The resulting intensity function is shown in Figure 7.3.

In Figure 7.4 we see the spatial distribution of the outbreaks under the five intervention strategies. All the unaffected farms are shown in grey, infected in red, vaccinated in blue and pre-emptively depopulated in green. We can see that most of the outbreak seems to result in pockets of disease. In the NC (a) plot we can see that the the pockets of disease are larger,

with no unaffected farms within the centre of a pocket. With the intervention methods V, DV, and DED we see that the pockets of disease are surrounded by the intervention methods, stopping the spread of the disease by reducing contact.

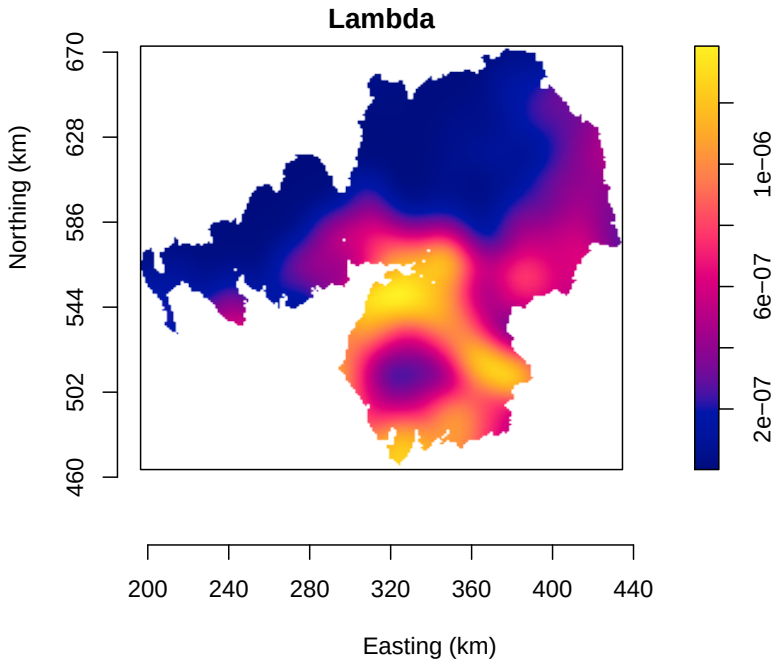


Figure 7.3: Estimated Border counties farm density, obtained using kernel estimation with a bandwidth of 8km.

7.3.1 Spatial-temporal plots

In Figure 7.5 we illustrate the first 60 days of the outbreak and its spatial temporal distribution. We can see that by day 20 the pockets of disease locations were determined and local spread followed. For the depopulation intervention strategy (Fig 7.6) we see that the last day a case was reported was day 31. After day 10 there were only a few long distance transmission, with the rest of the outbreak being characterised by local spread. After day 20 the outbreak appears to die out. The vaccinated outbreak (Fig 7.7) lasted slightly less time then the depopulation case, with the last two events occurring on day 26. What is interesting to see is that on the day 10 plot there are two groupings of cases around the coordinates (390,550). However, as the outbreak progresses we see that these two groups merge into one.

This is not surprising as we would expect a high probability of the farms in between getting the disease as their are multiple sources of infection.

With the combination of depopulation and vaccination (Fig 7.8) we once again see a drop in the length of the outbreak, with day 22 being the last recorded event. When depopulation and pre-emptive culling was applied (Fig 7.9) we see the last events occurring on day 24. For each of these intervention strategies we see that the spatial temporal distribution for day 10 is the same. This is because the initial stage of the outbreaks is determined by the epidemic history, which is kept constant.

7.3.2 Epidemic curves

Figure 7.10 shows the daily and the 7-day moving window epidemic curves for the simulated Border counties outbreak with each of the intervention strategies. We can see that the first 10 days of the outbreak are the same for all the different strategies, because each outbreak is started with the same epidemic history. In the analysis of the epidemic curve we see that there are no major differences between the control strategies, excluding no intervention.

We can see from the curves that if the outbreak simulation was allowed to extend past the 100 day maximum that the no control outbreak would have continued to spread. The outbreak has not fully run its course at 100 days.

7.3.3 Cluster curve

To evaluate the spatial clustering through time for the intervention strategies we apply the cluster curve. First we evaluate all intervention methods at a radius of 2km on a comparable scale, Figure (7.11). Here we can see that the overall level of clustering in the no control method (a) lasted substantially longer then the intervention strategies. We also see that there are slow changes in the spatial clustering through time, while when intervention strategies are applied we see more dramatic changes. In comparison for the three intervention methods that affect more than the FMD positive farm we see that the initial pattern is 'M' shaped. The depopulation method has the initial two peaks similar to the other three, however more peaks are present before the outbreak appears to die down. From the epidemic curves alone we were unable to gain this insight into the spatial structure.

In the consideration of significance at a radius of 2km we see that in the last stages of the outbreak for depopulation, vaccination and depopulations and pre-emptive culling, the events appear to be significantly clustered. This corresponds to the fact that the final stages

of these outbreaks were confined to a single local cluster, from which no further long range transmission occurred, indicating the success of these control strategies.

Due to the great difference in the size of the largest and smallest outbreak, we will proceed using the individually scaled plots for each intervention method. The individually scaled plots for each intervention method will occur for both the x and y axes. The x (i.e. time) axis is truncated for the instances in which the outbreak is eliminated in a shorter period of time. The y axis is scaled to the intervention method's degree of clustering. It is important to keep in mind the different axis scales when making comparisons between intervention methods.

For our simulated outbreak with no control methods (Fig 7.12) we can see that at smaller radiuses we have a significant degree of spatial clustering through time. At all radiuses investigated the cluster curve is positive indicating that the data is clustered. However, only at small radiuses is this clearly statistically significant. This suggests that the spread of disease is largely driven by transmission over very short distances, typically 1-2km. If we look at the structure at the smaller radiuses we see that the degree of clustering occurs in waves with even the troughs suggesting the presence of significant groupings suggesting clustering. This reflects the emergence of multiple clusters through time, as new isolated cases (generated perhaps through long range transmission) swiftly infect surrounding farms.

In the FMD simulated outbreak where confirmed cases led to depopulation of those farms (Fig 7.13) we see once again at the smaller radius a significance in the degree of clustering. At the larger radiuses there is an increase in the degree of clustering in the latter stages. This coincides with the slight increase in the number of events, as seen in the epidemic curves. The depopulation method works to try to reduce local spread by removing the infection source, but not entirely eradicating it as infection transmission can occur before clinical signs.

When we apply the vaccination method to control the disease we apply the treatment initially in a 5km radius ring, and then work inwards. This provides a buffer ring to help stop further transmission. The cluster curve application (fig 7.14) shows that at the smaller radiuses there is the presence of significant clustering in bursts throughout the outbreak. This could have occurred because local spread within the buffer ring.

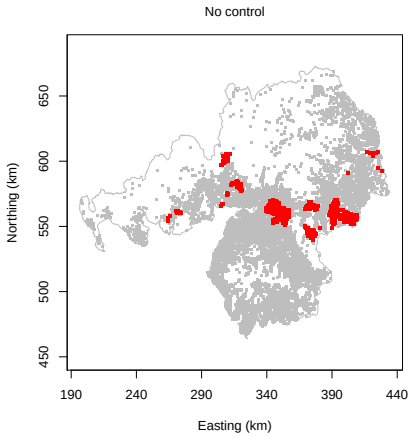
With the combination of vaccination and depopulation (Fig 7.15), for radiuses from 2:5km we see an almost identical structure to that of vaccinations alone. The main difference occurs at a radius of 1km, where the degree of clustering is reduced but with a greater presence of peaks and variations. Once again this could be due the buffer rings restricting the range of pockets of local spread.

The application of depopulating infected and suspect premises is shown in Figure 7.16. Once

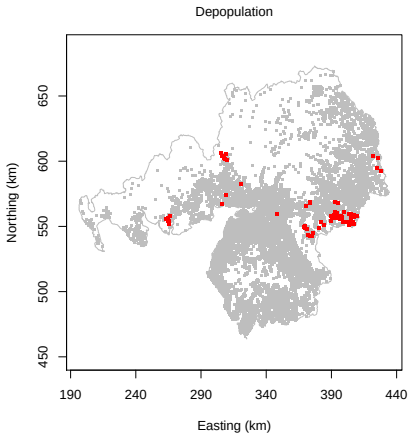
again we see a similar structure to the other methods that involved treatment to more then the farm, this similarity being more prominent at the larger radiuses. The main difference between the curves for the intervention methods that involved vaccination and this method is the presence of a steep increase at end of the outbreak. This steep peak coincides with the last increase in cases at the end of the epidemic, as shown by the epidemic curves. This suggests the presence of the identification of a new case, followed by the start of local spread which is rapidly controlled by the intervention method.

7.3.4 Integrated Cluster Curve

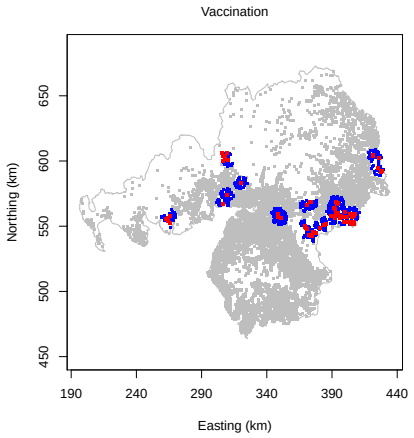
Our integrated cluster curve for the intervention strategies is shown in Figure 7.17 evaluating the function over a radius range of 0-5km. When the disease is allowed to follow its natural progression we see that the data appears to be constantly clustered throughout the studied time interval. With all intervention methods we see the same initial structure, this is expected due to the specified epidemic history. For depopulation we see that the latter part of the outbreak has a large increase in the degree of clustering until the control method manages to stop the outbreak. In the latter stages of the vaccination method we see a decrease then a small bump at the very end, suggesting that at the very last cases of the outbreak where clustered. This could be due to the cases being trapped within the buffer ring. When we combined depopulation and vaccination we see a very similar structure to vaccination alone, however the hump at the last stage of the outbreak is no longer present. For DED we once again see that as the outbreak died out the last cases were within close proximity to one another.



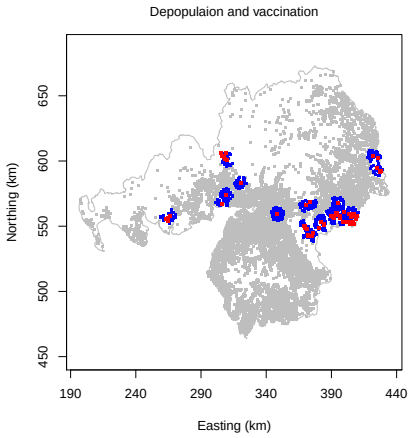
(a) No intervention



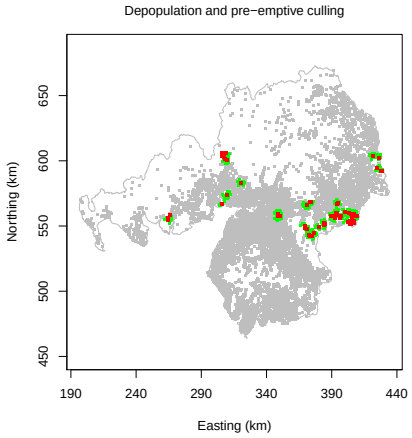
(b) Depopulation



(c) Vaccination

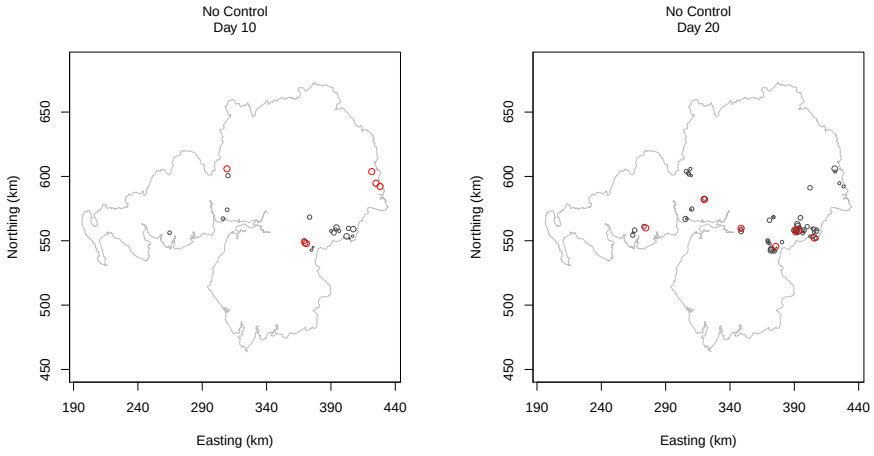


(d) Depopulation and vaccination



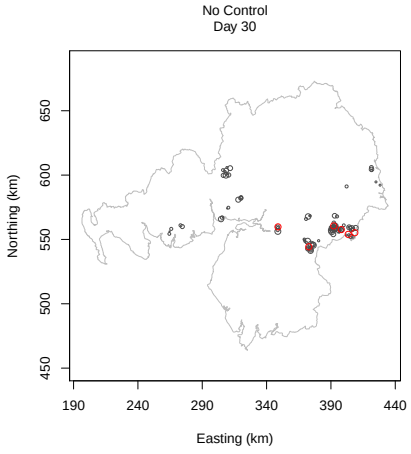
(e) Depopulation and pre-emptive culling

Figure 7.4: Image plots of simulated FMD outbreaks in Border counties: (a) No intervention (b) depopulation (c) vaccination (d) depopulation and vaccination (e) depopulation and pre-emptive culling. Grey represents farms with no FMD and no intervention, blue where no FMD but vaccination applied, green where no FMD but depopulation occurred and red where FMD is present.

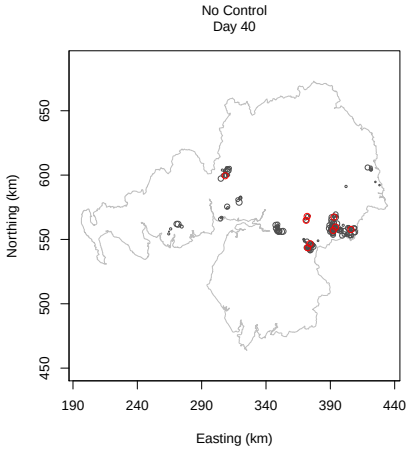


(a) Day10

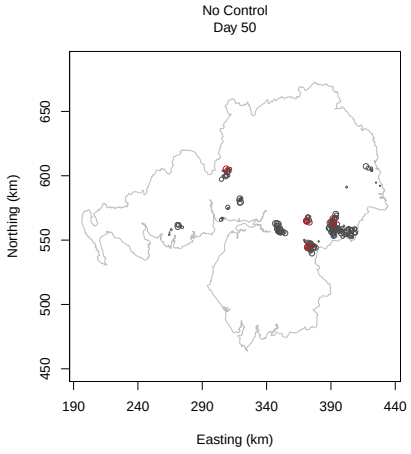
(b) Day20



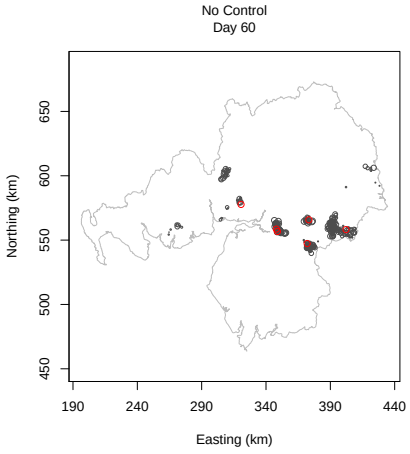
(c) Day30



(d) Day40

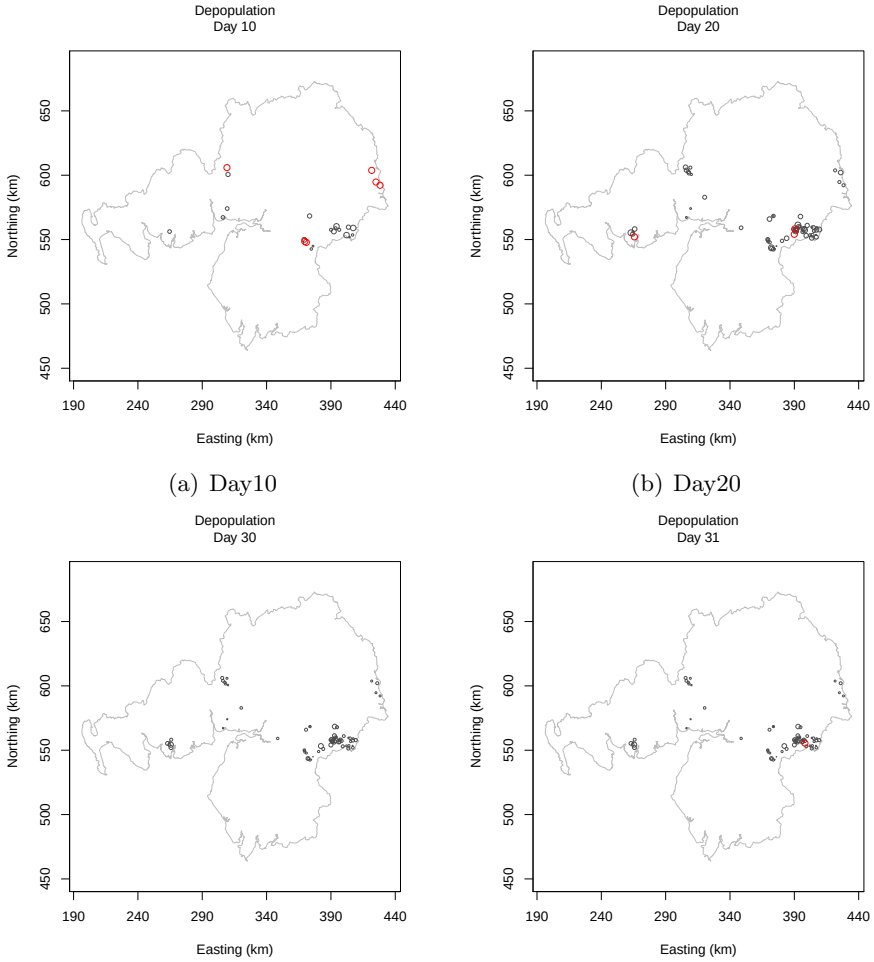


(e) Day50



(f) Day60

Figure 7.5: Spatial-temporal 10 day window plots of simulated No control FMD outbreak in Border counties: (a) Day 10 (b) Day 20 (c) Day 30 (d) Day 40 (e) Day 50 (f) Day 60. Red represents the farms diagnosed with the presence of FMD has been diagnosed on the day, Grey represents the previously diagnosed farms. The plotting size represents the temporal scale, with the larger the point the more recently it was diagnosed.



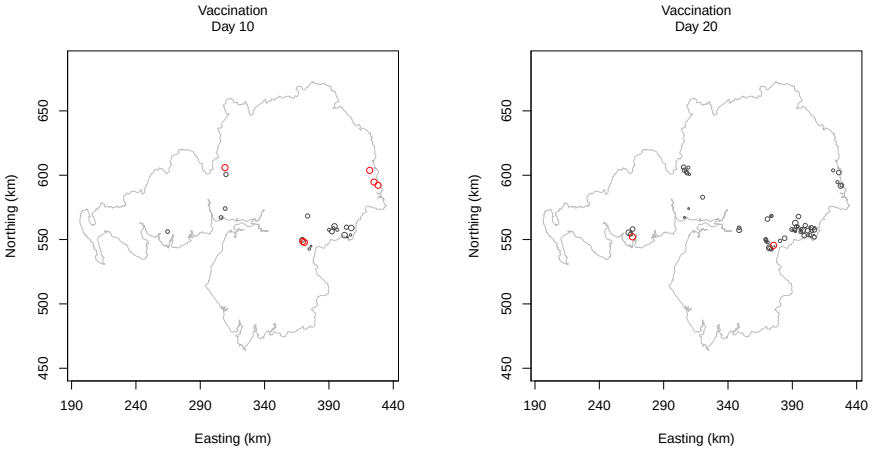
(a) Day10

(b) Day20

(c) Day30

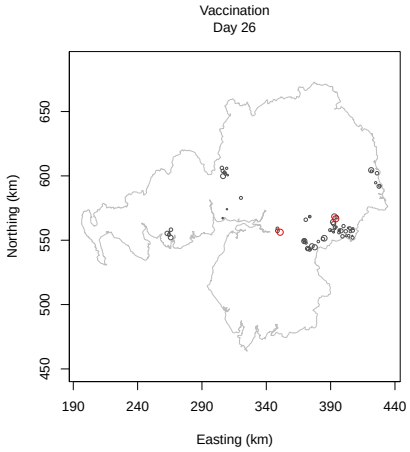
(d) Day31

Figure 7.6: Spatial-temporal 10 day window plots of simulated depopulated FMD outbreak in Border counties: (a) Day 10 (b) Day 20 (c) Day 30 (d) Day 31. Red represents the farms diagnosed with the presence of FMD has been diagnosed on the day, Grey represents the previously diagnosed farms. The plotting size represents the temporal scale, with the larger the point the more recently it was diagnosed.



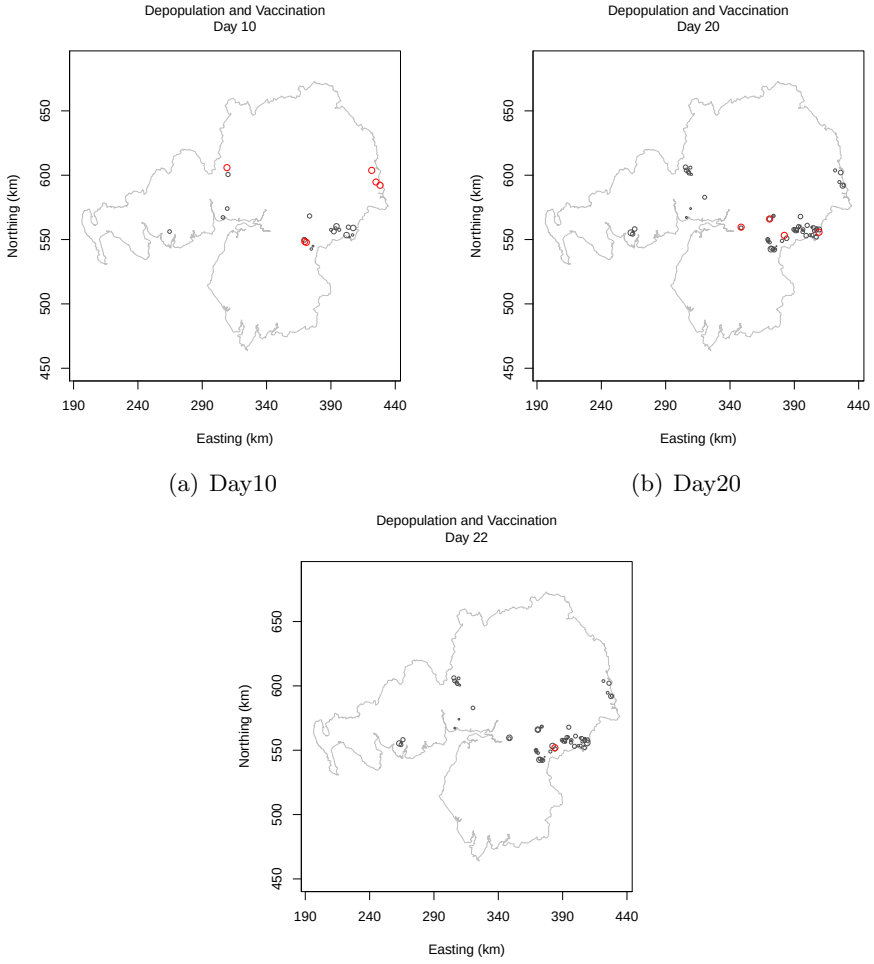
(a) Day10

(b) Day20



(c) Day26

Figure 7.7: Spatial-temporal 10 day window plots of simulated vaccinated FMD outbreak in Border counties: (a) Day 10 (b) Day 20 (c) Day 26. Red represents the farms diagnosed with the presence of FMD has been diagnosed on the day, Grey represents the previously diagnosed farms. The plotting size represents the temporal scale, with the larger the point the more recently it was diagnosed.

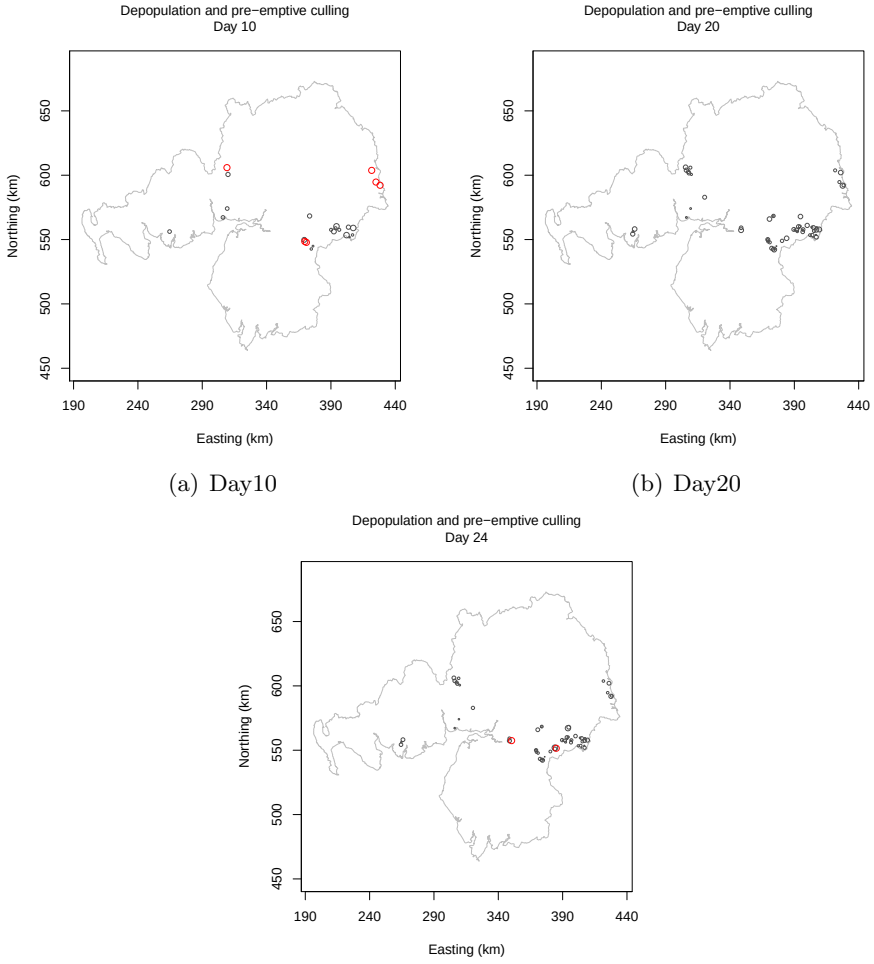


(a) Day10

(b) Day20

(c) Day22

Figure 7.8: Spatial-temporal 10 day window plots of simulated depopulated and vaccinated FMD outbreak in Border counties: (a) Day 10 (b) Day 20 (c) Day 22. Red represents the farms diagnosed with the presence of FMD has been diagnosed on the day, Grey represents the previously diagnosed farms. The plotting size represents the temporal scale, with the larger the point the more recently it was diagnosed.



(a) Day10

(b) Day20

(c) Day24

Figure 7.9: Spatial-temporal 10 day window plots of simulated depopulated and pre-emptive culled FMD outbreak in Border counties: (a) Day 10 (b) Day 20 (c) Day 24. Red represents the farms diagnosed with the presence of FMD has been diagnosed on the day, Grey represents the previously diagnosed farms. The plotting size represents the temporal scale, with the larger the point the more recently it was diagnosed.

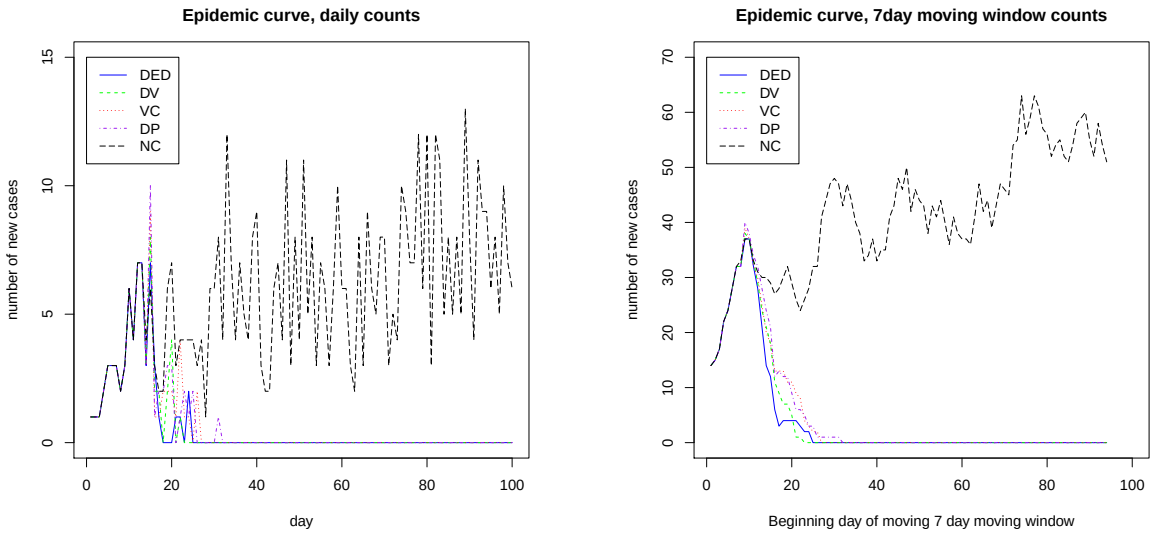
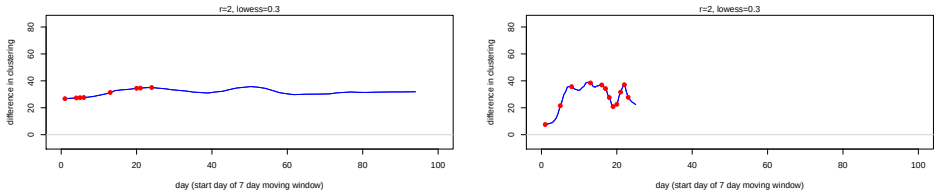
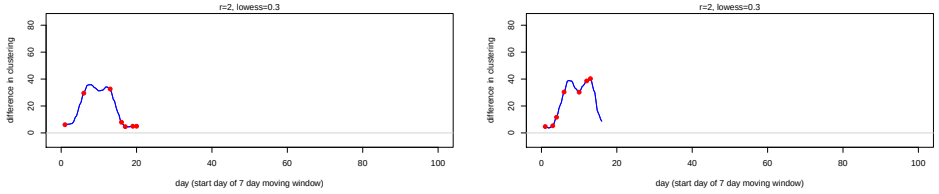


Figure 7.10: Epidemic curve of Border counties intervention datasets. Left daily epidemic curve, right 7-day moving window epidemic curve.



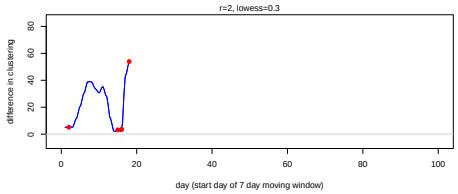
(a) No control

(b) Depopulation



(c) Vaccination

(d) Depopulation and Vaccination



(e) Depopulation and pre-emptive depopulation

Figure 7.11: Cluster curve for all intervention strategies at a radius of 2km on a comparable scale.

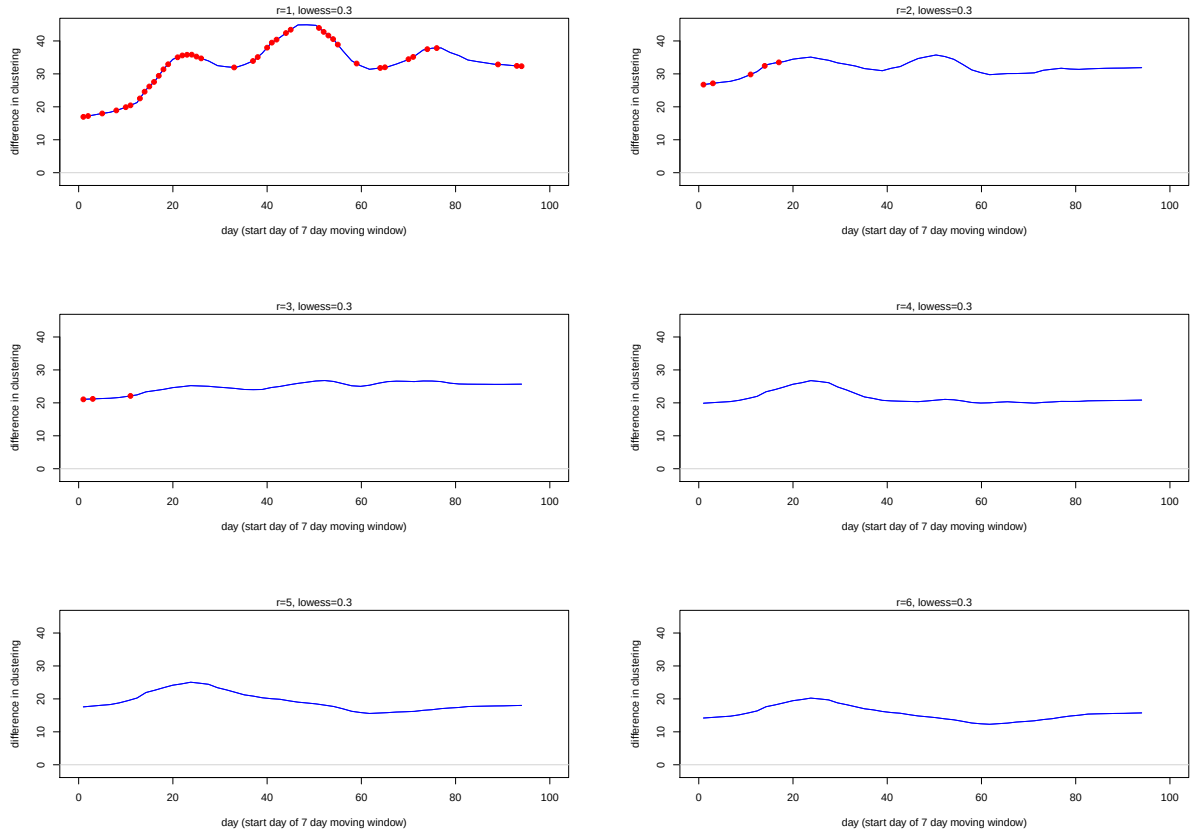


Figure 7.12: Cluster curve for the Border counties FMD simulated outbreak with no control methods

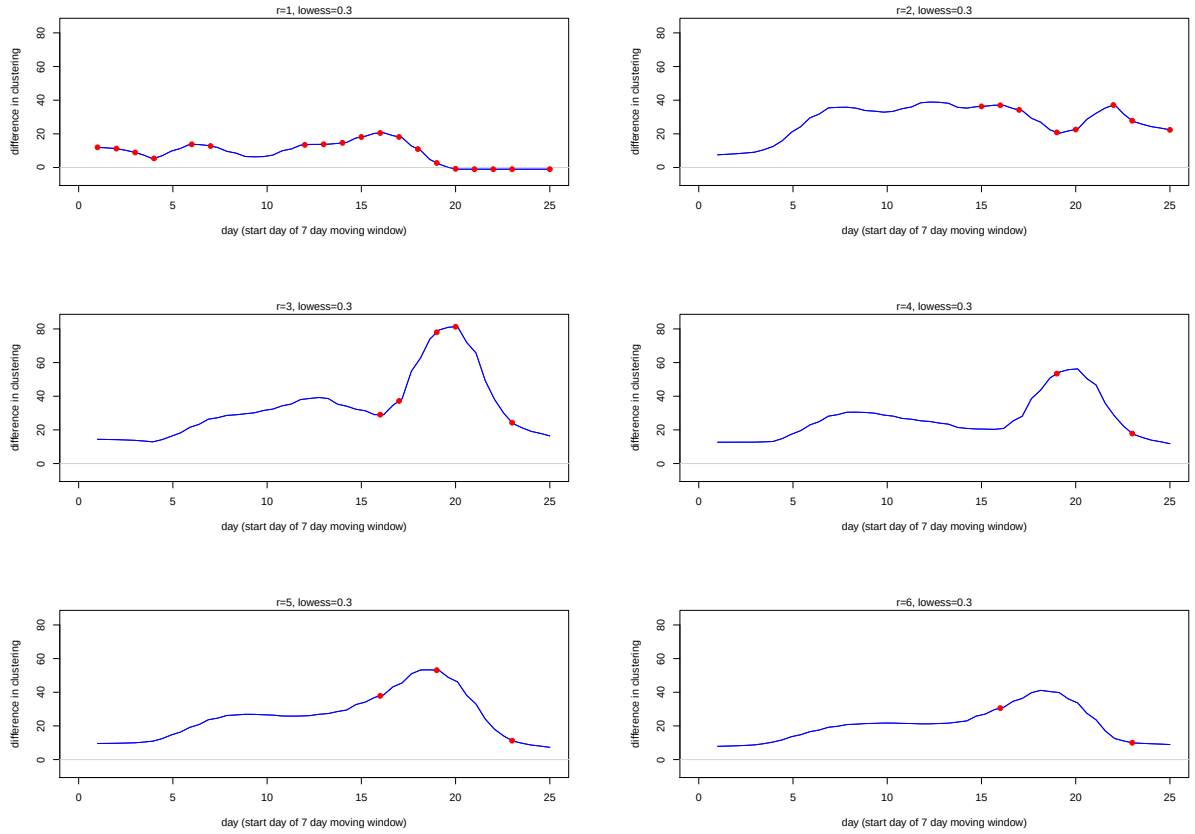


Figure 7.13: Cluster curve for the Border counties FMD simulated outbreak with depopulation control methods

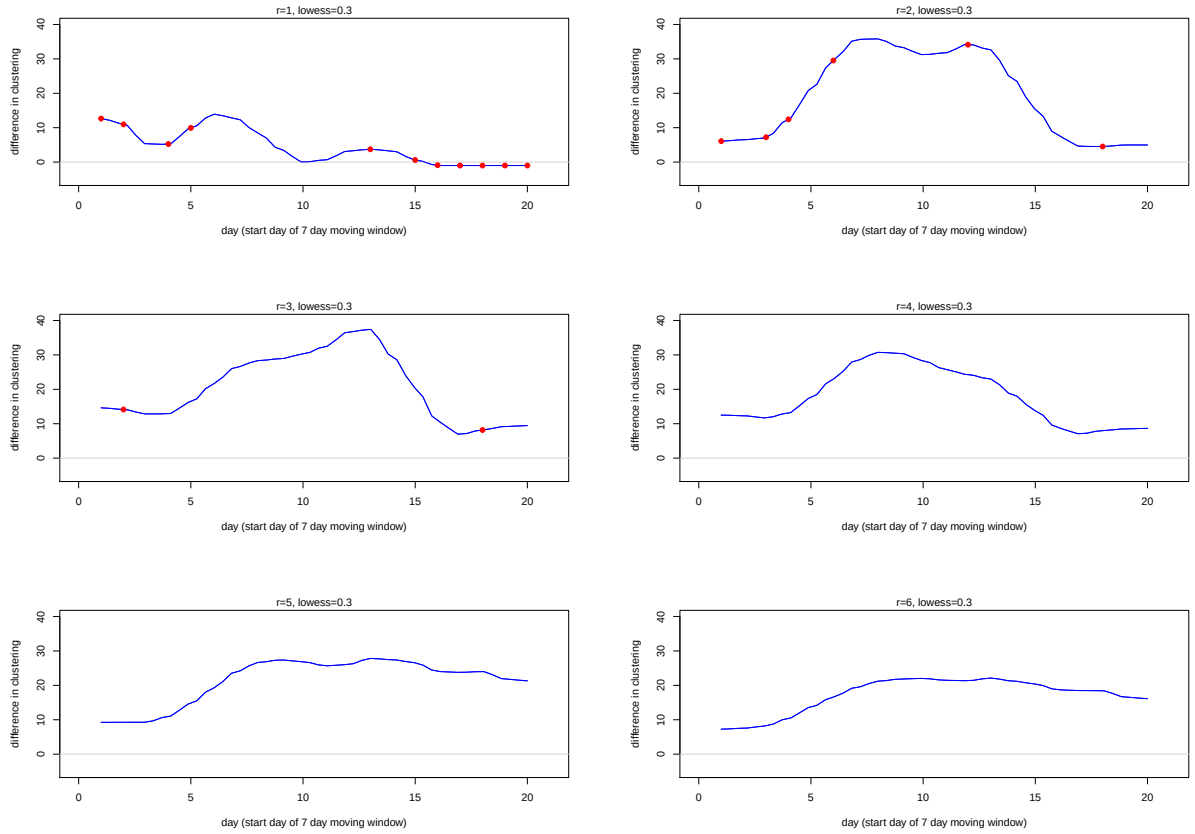


Figure 7.14: Cluster curve for the Border counties FMD simulated outbreak with vaccination control methods

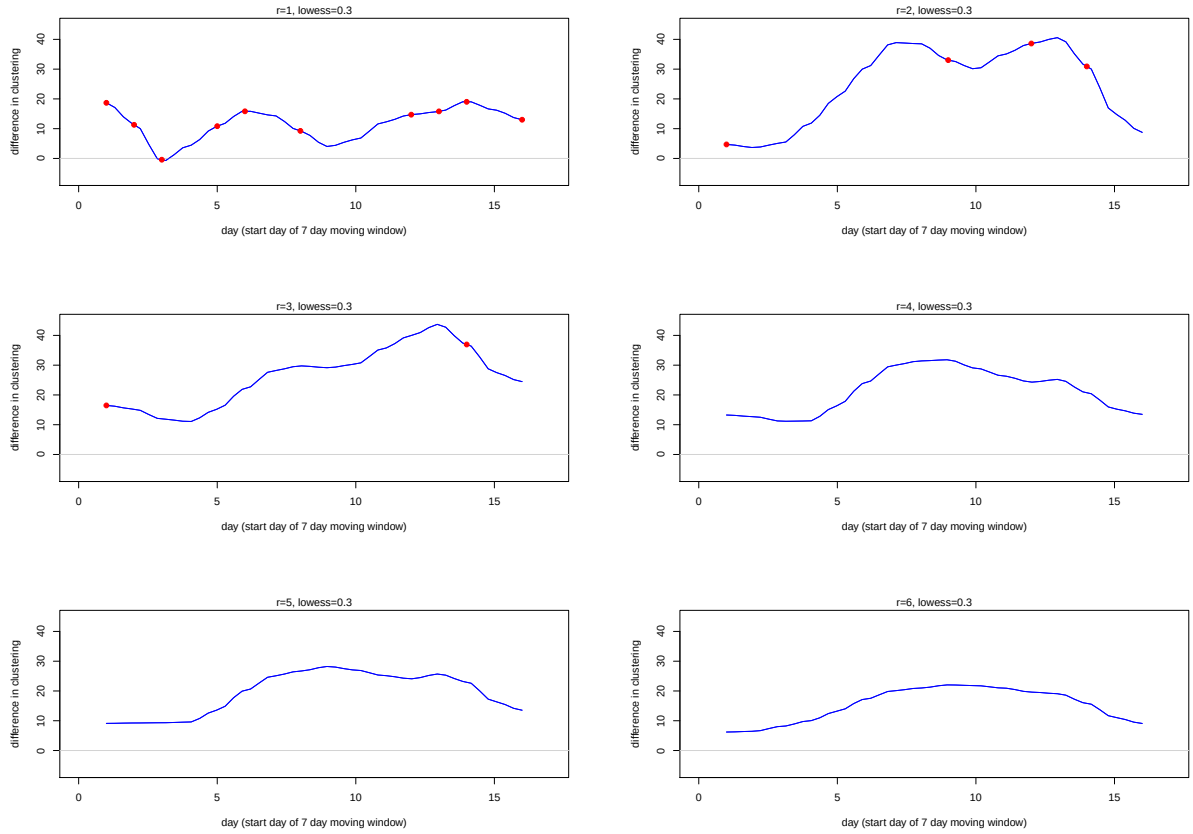


Figure 7.15: Cluster curve for the Border counties FMD simulated outbreak with depopulation and vaccination control methods

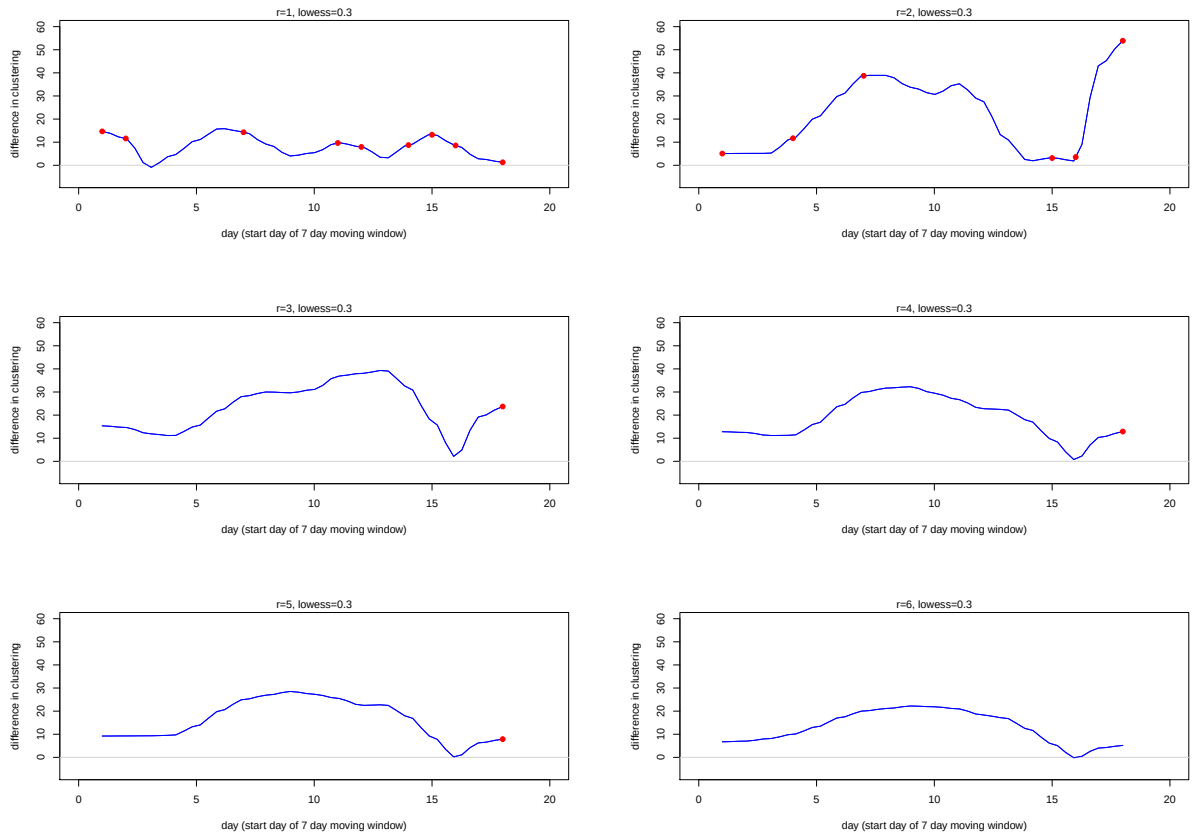
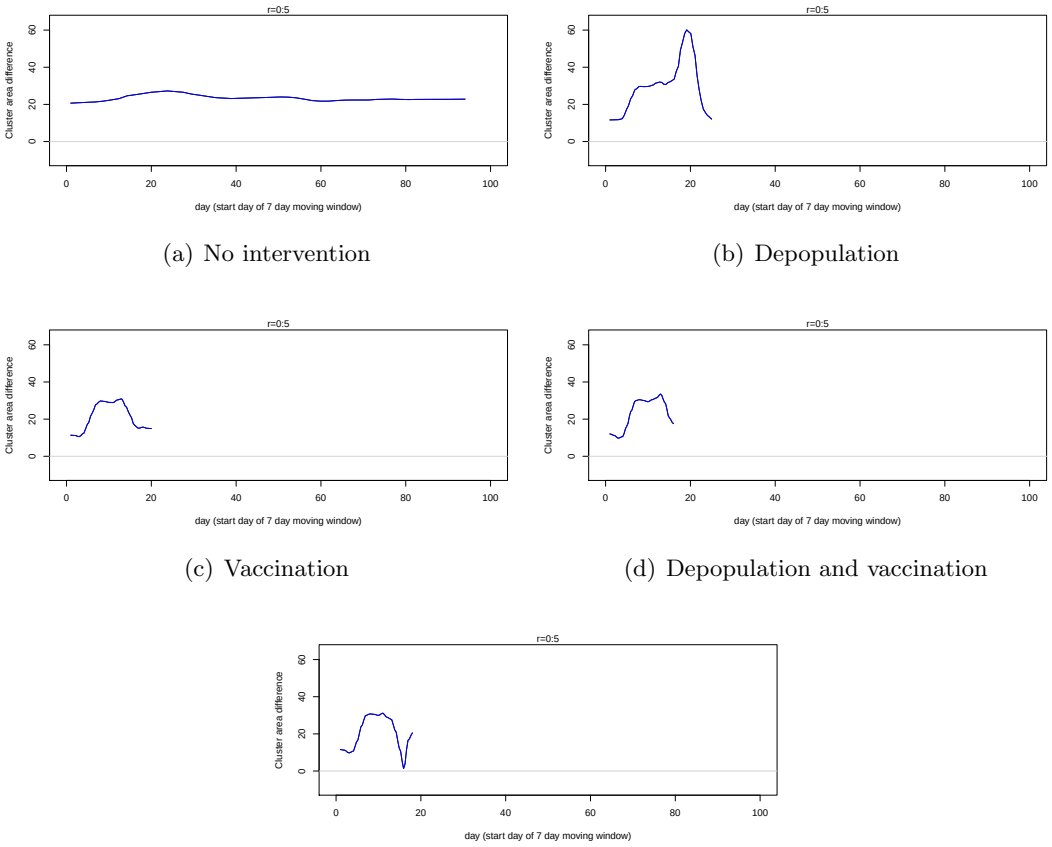


Figure 7.16: Cluster curve for the Border counties FMD simulated outbreak with depopulation and pre-emptive culling control methods



(a) No intervention

(b) Depopulation

(c) Vaccination

(d) Depopulation and vaccination

(e) Depopulation and pre-emptive culling

Figure 7.17: Integrated cluster curve for simulated FMD outbreaks in Border counties over a radius of 0:5km : (a) No intervention (b) depopulation (c) vaccination (d) depopulation and vaccination (e) depopulation and pre-emptive culling.

7.4 Miyazaki, Japan

The simulated Miyazaki datasets is explained in detail in section 3.5.2. Once again we focus our attention on just the diseased farms (as opposed to those subject to some intervention, like pre-emptive culling) when looking at the clustering structure through time.

In the application of the cluster curve we once again produce a farm density estimate plot based on the location of all the farms. This provides the model with the underlying intensity for the population. location to obtain a kernel estimate of the spatial intensity function $\lambda(x)$. This was implemented using a bandwidth of $h_x = 1.3km$ (smoothing in the x direction) and $h_y = 2.3km$ (smoothing in the y direction), chosen using Scott's (Scott, 1992) rule of thumb bandwidth selector. The resulting intensity function is shown in Figure 7.18.

The spatial distribution of each outbreak is shown in Figure 7.19. We can see that unlike the Border counties example, these outbreaks seem to involve most of the farms within the areas. For the methods where more than the infected farm received treatments there appears to be more treated farms than confirmed cases.

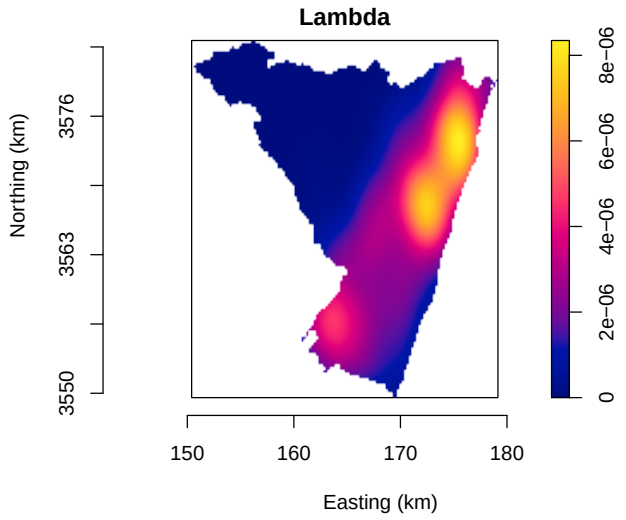


Figure 7.18: Estimated farm density in Miyazaki, Japan

7.4.1 Spatial-temporal plots

Similar to the Border counties outbreaks we see that for all intervention strategies the spatial-temporal plots for the first 10 days are the same due to the specification of the epidemic history. When we allow the outbreak to naturally progress we see the spatial temporal for the first 60 days in Figure 7.20. By day 20 we have a large grouping in the upper right of the window with a few odd new cases in the lower area. By day 30 the cases spread south. For the day 50 and 60 plots even though we do not have any red cases, there is still active infection as the plotting size of some of the past events indicate that they are recent.

Similar to the no control (NC) case, for depopulation (Fig 7.21) we look at the first 60 days of the outbreak. If we compare these two outbreaks, even though the temporal length for the two scenarios is very similar there are substantially less cases in the depopulation example. The vaccination control outbreak (Fig 7.22) lasted 37 days and the combination of depopulation and vaccination (Fig 7.23) lasted 36. As we progress through the intervention strategies (NC, D, V, DV, DED) the number of cases decreases. The lessening of the cases generally occurs in the southern parts of the window. With the depopulation and pre-emptive culling example (Fig 7.24) we observe that the outbreak lasts 19 days and appears to cover less of the study window. However if we look at Figure 7.19 (e) we see that the outbreak does in fact affect the majority of farms. These farms were just pre-emptively culled before they could contract the disease in an attempt to control the outbreak. For all of the plots the location of large groupings around the upper right of the study window is not surprising given the underlying population density (Fig 7.18).

7.4.2 Epidemic curves

Figure 7.25 shows the daily and 7-day moving window epidemic curves. We see that all the curves have similar initial stages of the outbreak. The epidemic history determines the first 7 days of each intervention method but we can observe from the plot that the initial stages are very similar for all curves past this specified 7 days. All the curves have a similar shape with DED dropping off first, followed by DV, VC,D and finally NC. The fact that the no control strategy dataset decreases to near 0 before the maximum time frame, unlike the Border counties case, indicates that the disease ran out of susceptible farms. This is re-enforced by the spatial distribution plots.

7.4.3 Cluster curve

When comparing the application of the cluster curve on a common scale (Fig 7.26) we see that the late stages of the outbreak for NC is substantially larger then the others. Fixing all comparisons to this scale means we loose the ability to see subtle changes for the other methods, hence we continue with each method on their own scale. We scale in both the x and y axes. The x (time) axis is truncated for the instances in which the outbreak is killed off in a shorter period of time. The y axis is scaled to the intervention methods degree of clustering. It is important to keep in mind the different axis scales when making comparisons between intervention methods.

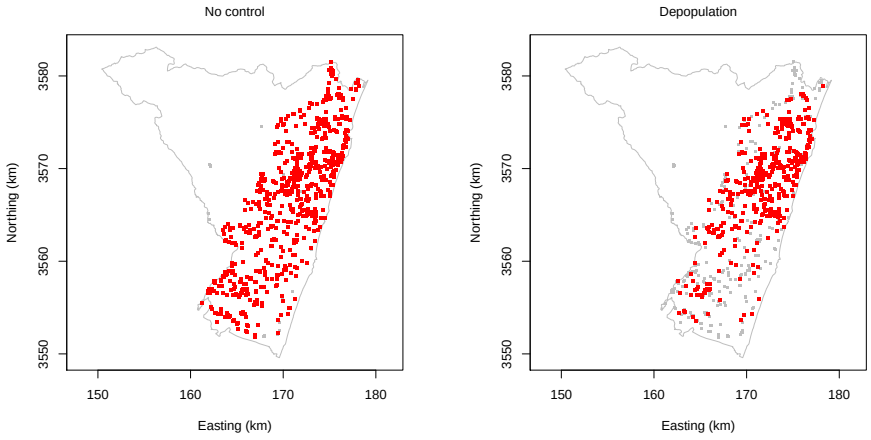
The large bump present at the end of the NC cluster curve application (Fig 7.27) is not deemed significant and occurs when the temporal windows occupy very few points. Therefore this may just be a noisy artefact. The random appearance of significant dots could be the occurrence of the location of points within close proximity within only that 7-day window or simply a false positive. The depopulation example (Fig 7.28) exhibits a degree of increased clustering through time for radiuses greater then 2km. This is once again where the cases are sparse. The increase at the end of the outbreak for depopulation suggests that this method was not completely effective in controlling the local spread resulting in a secondary outbreak of local clustering. With V (Fig 7.29), DV (Fig 7.30), and DED (Fig 7.31) strategies we see very little change in spatial clustering through time from a radius of 2km onwards. When we compare the cluster curves at a radius of 1km we see that D, V and DV have a similar structure while DED appears to have a similar spread to that which is expected after the initial distribution specified by the epidemic curve. This suggests that the DED strategy was very effective at shutting down the local spread of the outbreak.

Overall with the Japanese simulated intervention epidemics we do not see a large amount of significance, even at the larger degrees of clustering difference. However, this does not mean that there is not the presence of clustering. Rather, it may reflect the difficulty in attaining statistical significance in outbreaks with rather few new cases in each time window.

7.4.4 Integrated Cluster Curve

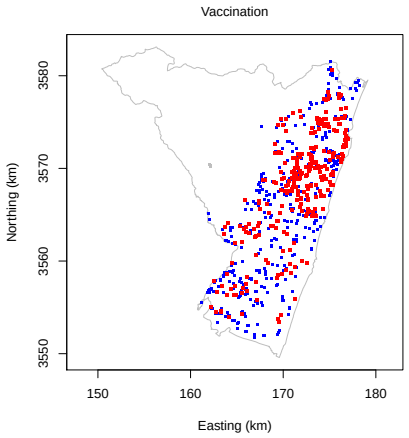
When we look at the degree of clustering through time over a range of 0:5km (Fig 7.32) for all intervention methods we can see the hint that the initial phase of the outbreak is grouped. For NC and D, where none or only the infected farm is treated, we see a similar clustering structure where the cases in the later time windows of the outbreak appear to be significantly clustered, with the depopulation scenario just to a lesser degree. The presence of the increased

clustering in the latter stages of the outbreak could be artefacts resulting from the sparsity of data, or could suggest that the last events are the result of local spread. With the three methods that involve the surrounding farms we once again see a similar structure. Over the range of radiuses 0:5km, the degree of clustering is minimal and seen to decrease with time.

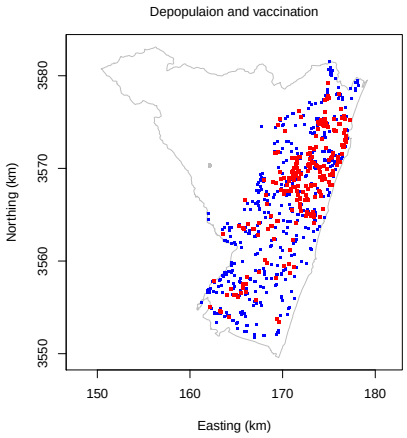


(a) No intervention

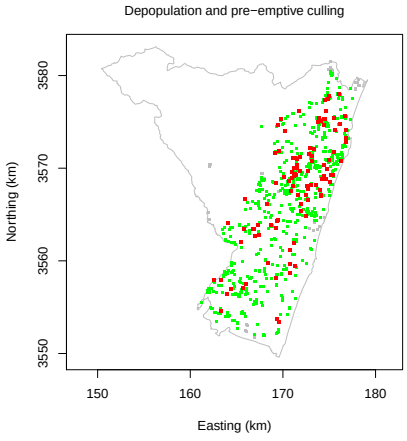
(b) Depopulation



(c) Vaccination

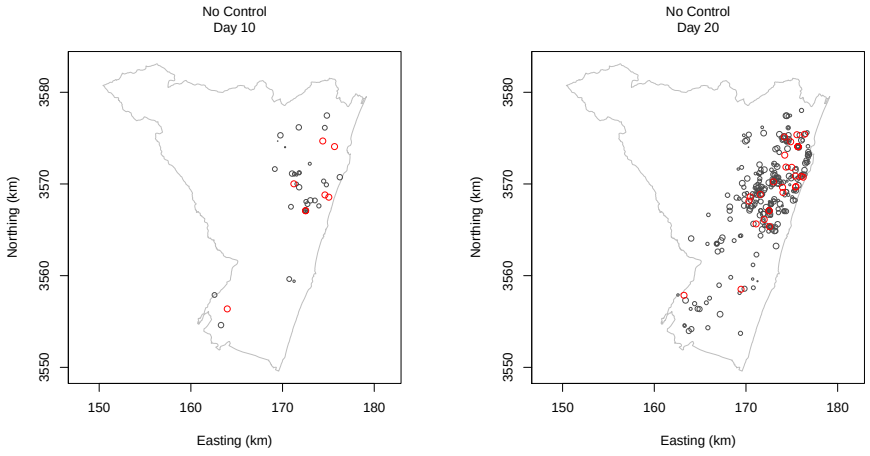


(d) Depopulation and vaccination



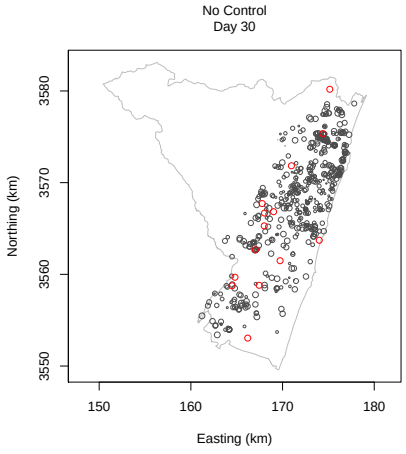
(e) Depopulation and pre-emptive culling

Figure 7.19: Image plots of simulated FMD outbreaks in Miyazaki (Grey all farms, Red FMD positive farms) : (a) No intervention (b) depopulation (c) vaccination (d) depopulation and vaccination (e) depopulation and pre-emptive culling. Grey represents farms with no FMD and no intervention, blue where no FMD but vaccination applied, green where no FMD but depopulation occurred and red where FMD is present.

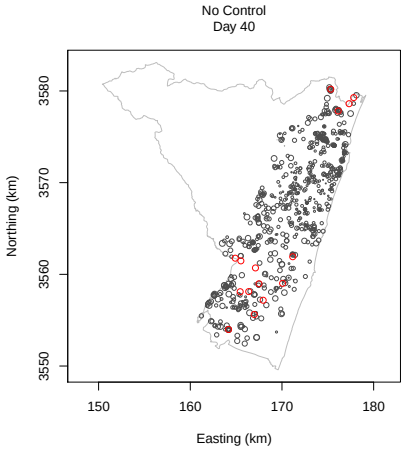


(a) Day10

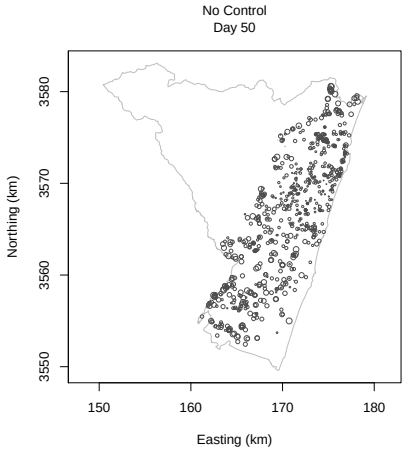
(b) Day20



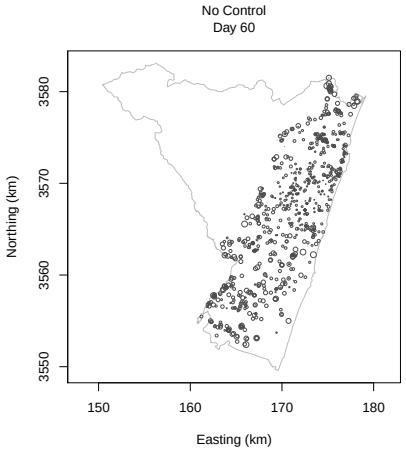
(c) Day30



(d) Day40

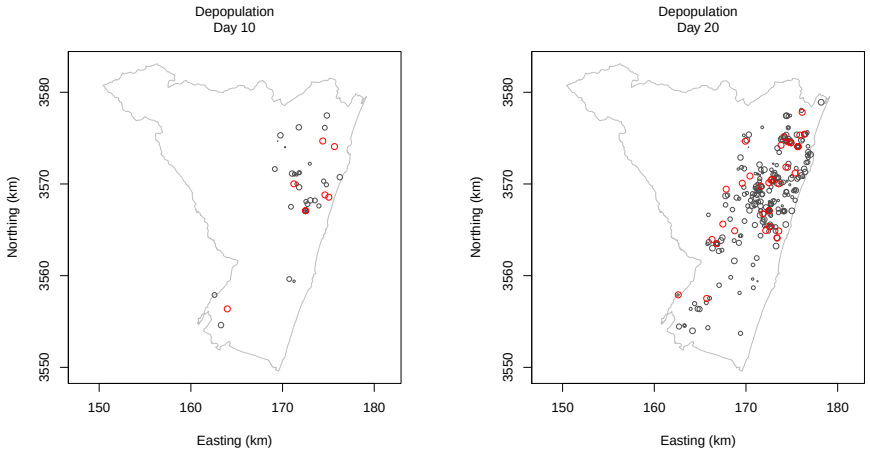


(e) Day50



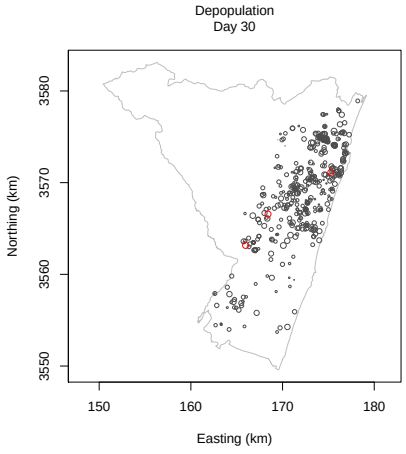
(f) Day60

Figure 7.20: Spatial-temporal 10 day window plots of simulated No control FMD outbreak in Miyazaki : (a) Day 10 (b) Day 20 (c) Day 30 (d) Day 40 (e) Day 50 (f) Day 60. Red represents the FMD positive farms recorded on the day, grey represents the previously diagnosed farms. The plotting size representing the temporal scale, with the larger the point the more recently it was diagnosed.

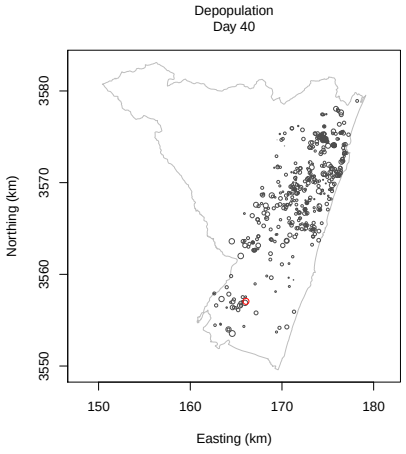


(a) Day10

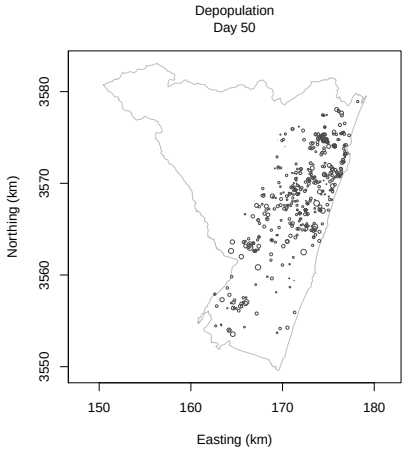
(b) Day20



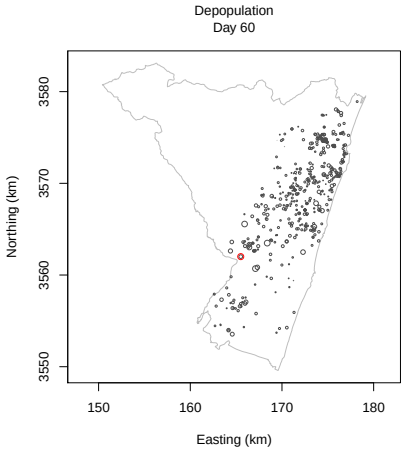
(c) Day30



(d) Day40

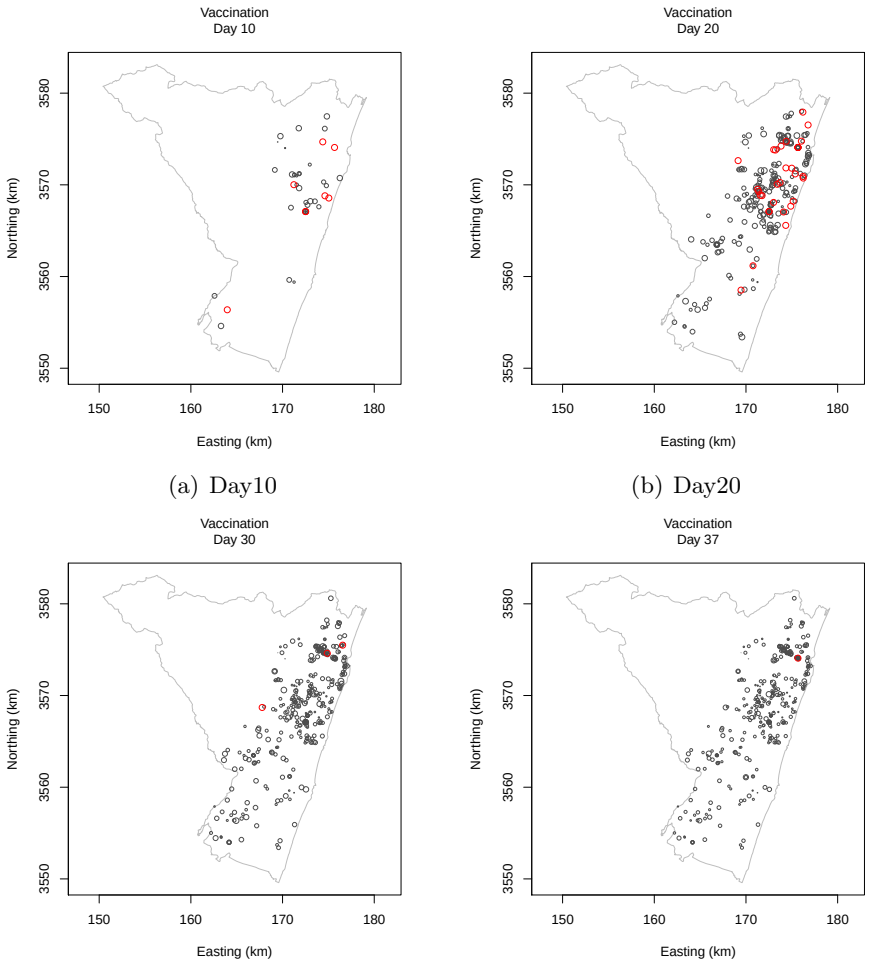


(e) Day50



(f) Day60

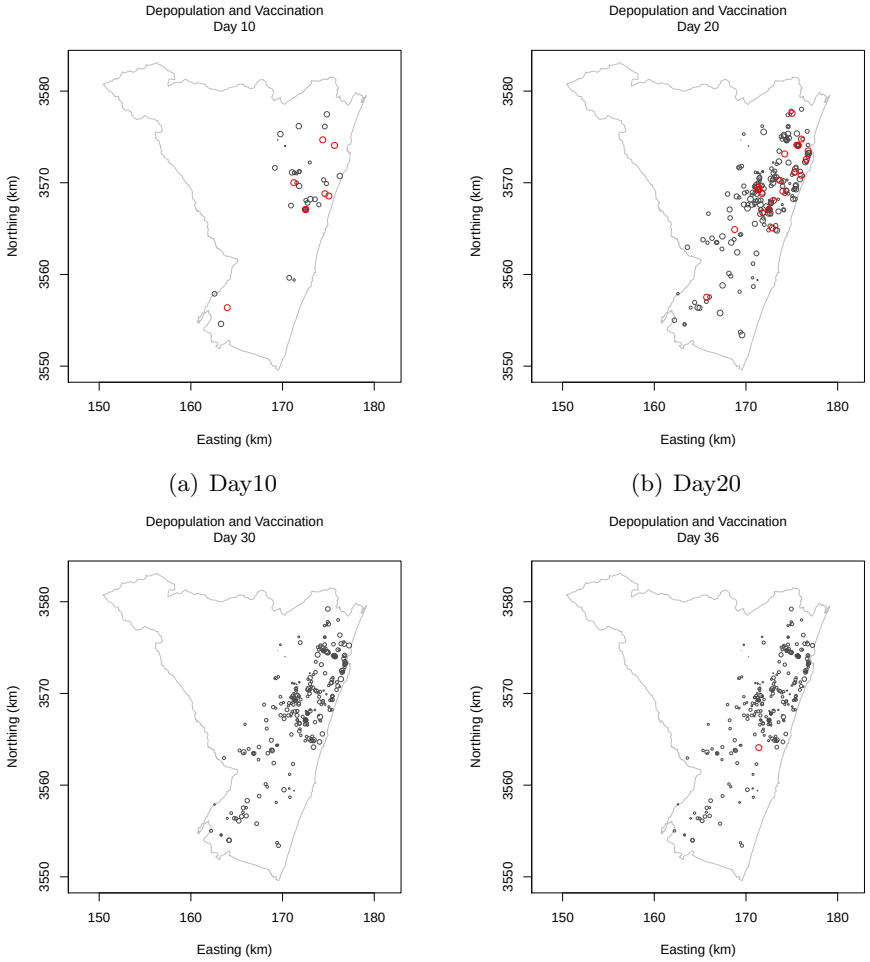
Figure 7.21: Spatial-temporal 10 day window plots of simulated depopulated FMD outbreak in Miyazaki : (a) Day 10 (b) Day 20 (c) Day 30 (d) Day 40 (e) Day 50 (f) Day 60. Red represents the FMD positive farms recorded on the day, grey represents the previously diagnosed farms. The plotting size representing the temporal scale, with the larger the point the more recently it was diagnosed.



(a) Day10 (b) Day20

(c) Day30 (d) Day37

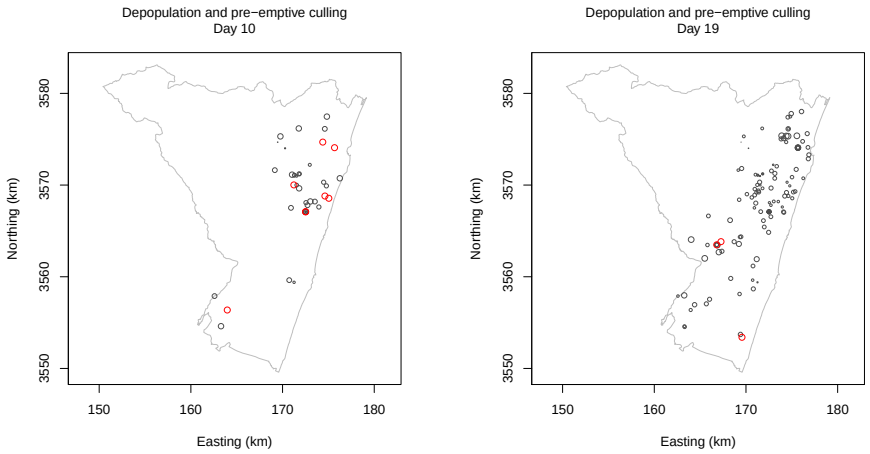
Figure 7.22: Spatial-temporal 10 day window plots of simulated vaccinated FMD outbreak in Miyazaki : (a) Day 10 (b) Day 20 (c) Day 30 (d) Day 37. Red represents the FMD positive farms recorded on the day, grey represents the previously diagnosed farms. The plotting size representing the temporal scale, with the larger the point the more recently it was diagnosed.



(a) Day10 (b) Day20

(c) Day30 (d) Day36

Figure 7.23: Spatial-temporal 10 day window plots of simulated depopulated and vaccinated FMD outbreak in Miyazaki : (a) Day 10 (b) Day 20 (c) Day 30 (e) Day 36. Red represents the FMD positive farms recorded on the day, grey represents the previously diagnosed farms. The plotting size representing the temporal scale, with the larger the point the more recently it was diagnosed.



(a) Day10 (b) Day19

Figure 7.24: Spatial-temporal 10 day window plots of simulated depopulated and pre-emptive culled FMD outbreak in Miyazaki : (a) Day 10 (b) Day 19. Red represents the FMD positive farms recorded on the day, grey represents the previously diagnosed farms. The plotting size representing the temporal scale, with the larger the point the more recently it was diagnosed.

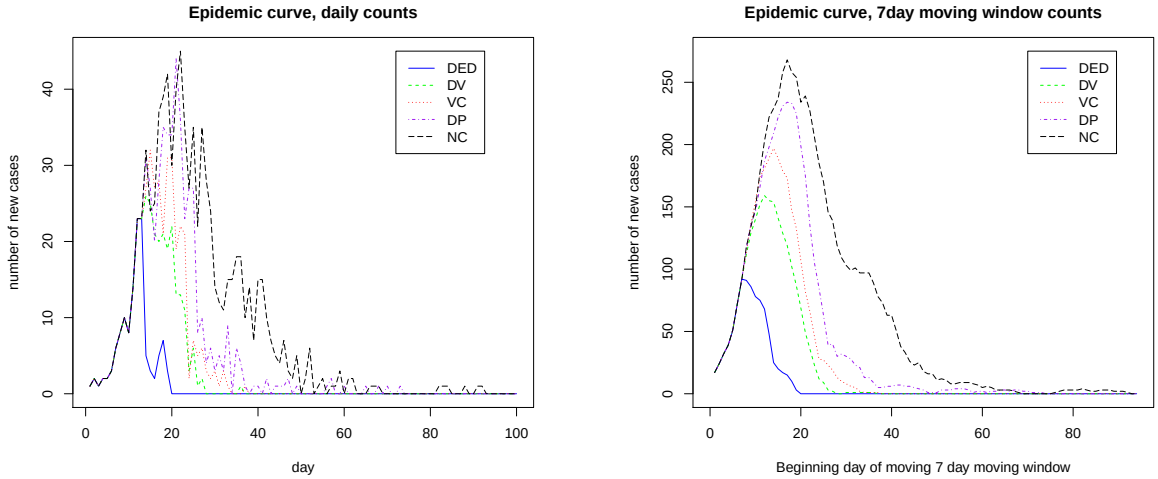
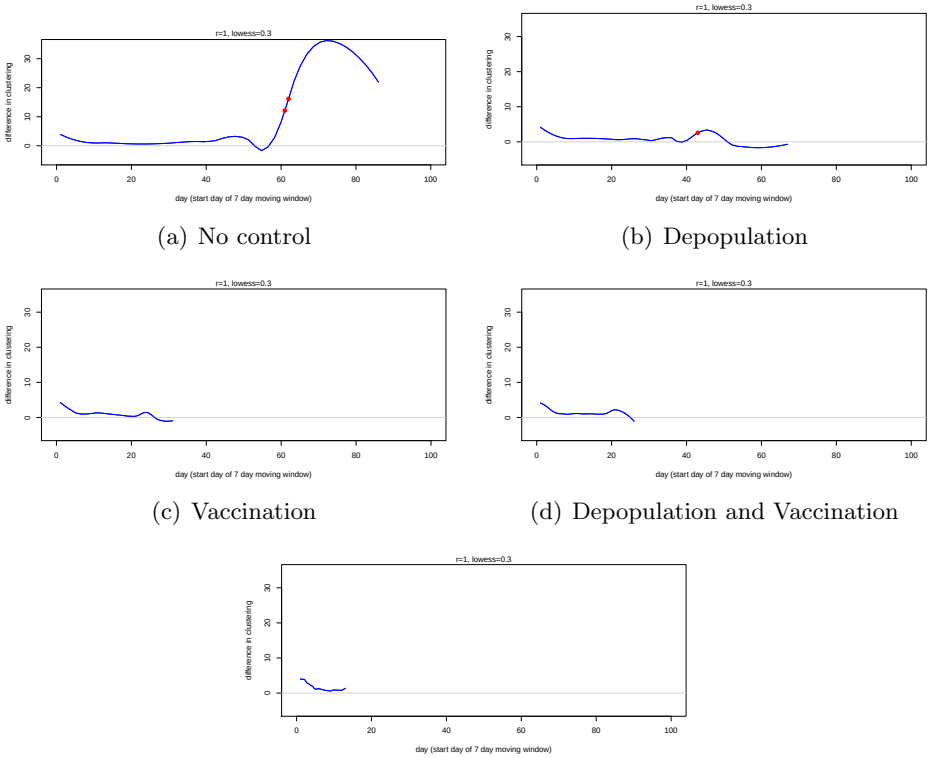


Figure 7.25: Epidemic curve of Miyazaki intervention datasets. Left daily epidemic curve, right 7-day moving window epidemic curve.



(a) No control

(b) Depopulation

(c) Vaccination

(d) Depopulation and Vaccination

(e) Depopulation and pre-emptive depopulation

Figure 7.26: Cluster curve for all intervention strategies at a radius of 1km for Miyazaki FMD outbreaks on a comparable scale. redoing

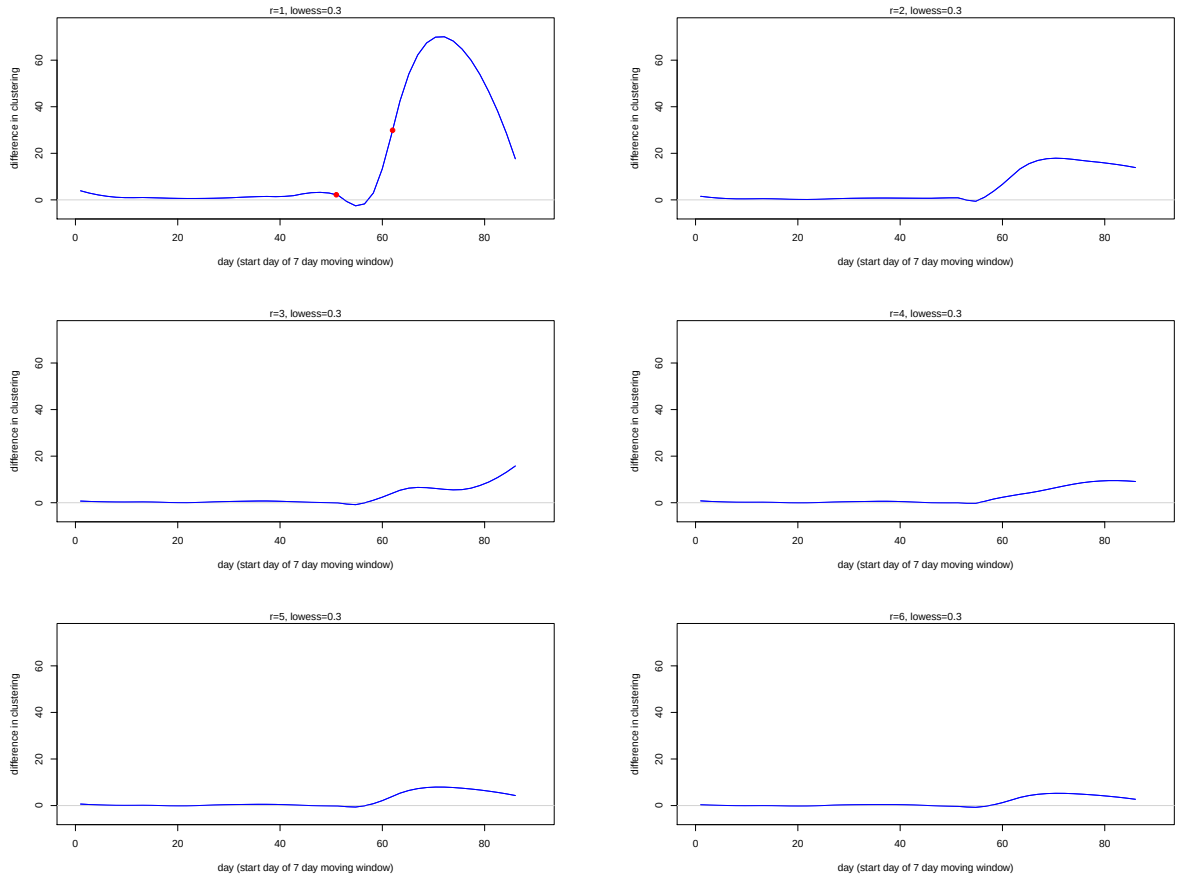


Figure 7.27: Cluster curve for the Miyazaki FMD simulated outbreak with no control methods

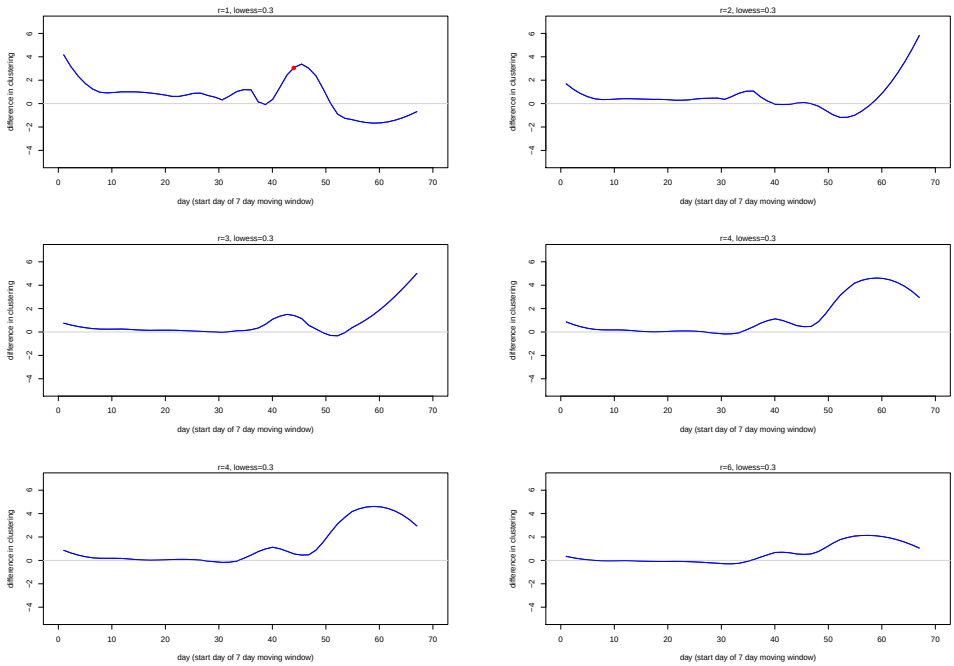


Figure 7.28: Cluster curve for the Miyazaki FMD simulated outbreak with depopulation control methods.

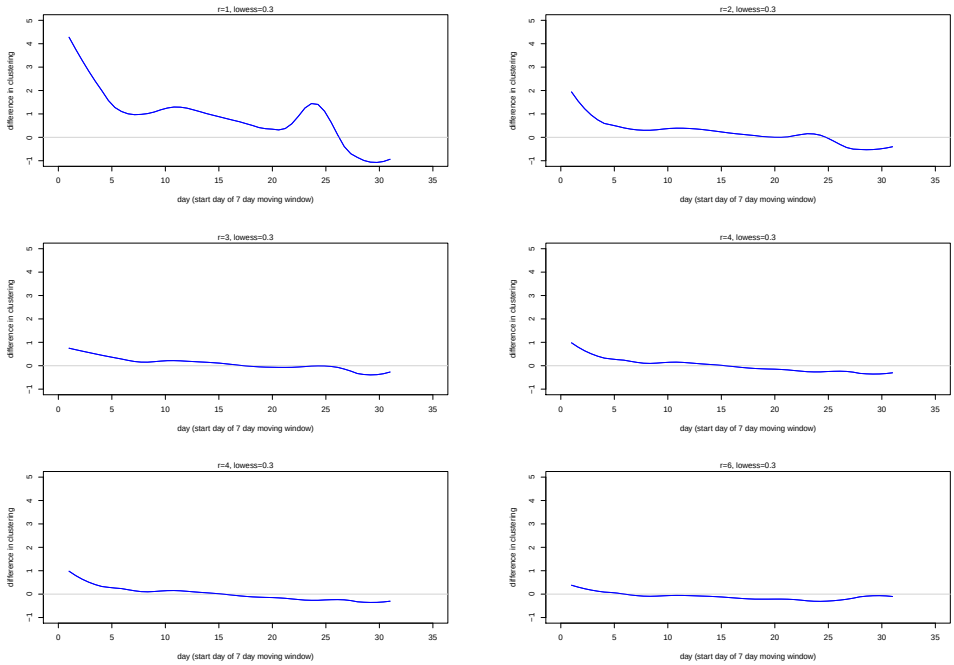


Figure 7.29: Cluster curve for the Miyazaki FMD simulated outbreak with vaccination control methods.

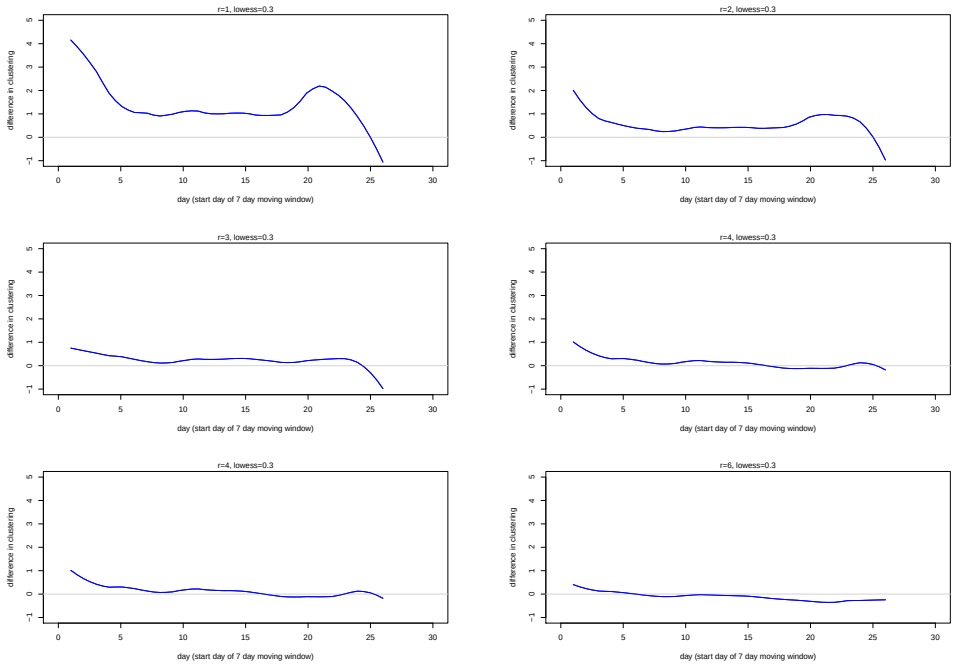


Figure 7.30: Cluster curve for the Miyazaki FMD simulated outbreak with depopulation and vaccination control methods.

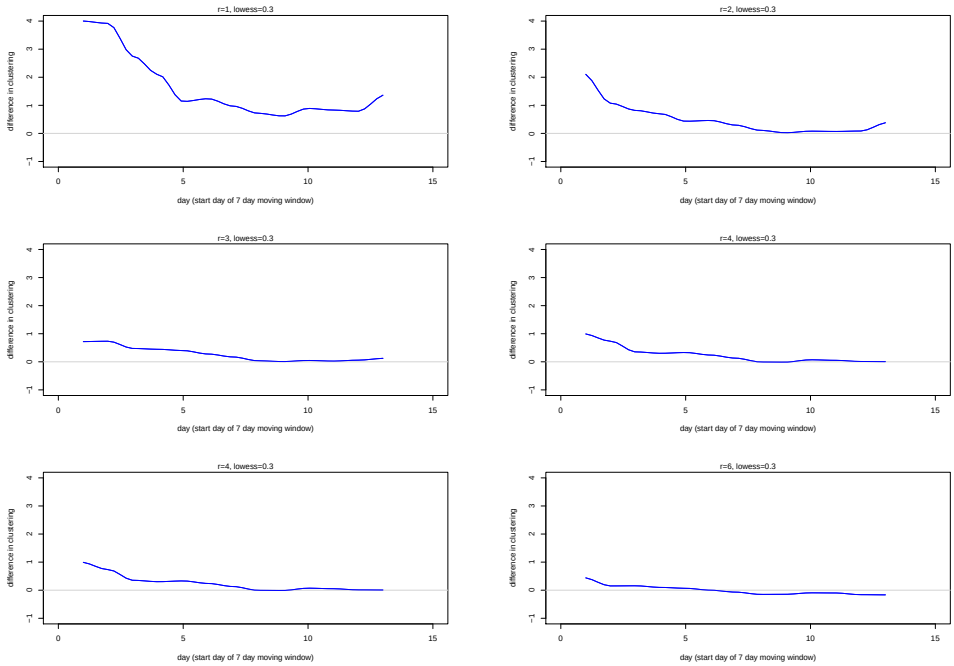
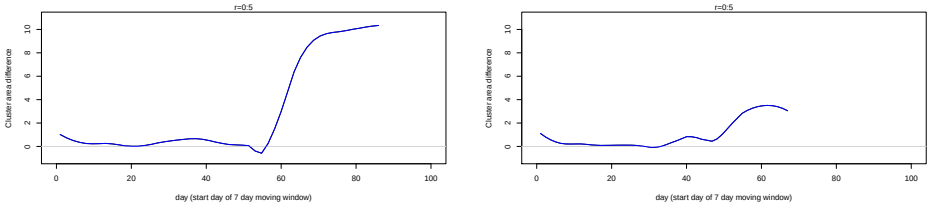
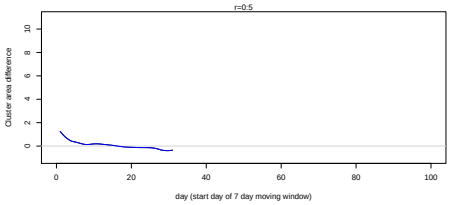


Figure 7.31: Cluster curve for the Miyazaki FMD simulated outbreak with depopulation and pre-emptive culling control methods.

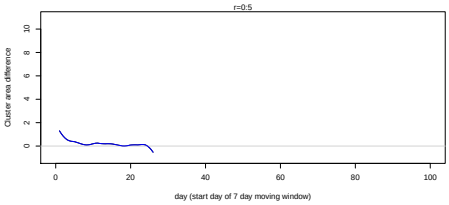


(a) No intervention

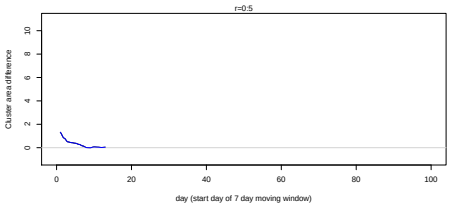
(b) Depopulation



(c) Vaccination



(d) Depopulation and vaccination



(e) Depopulation and pre-emptive culling

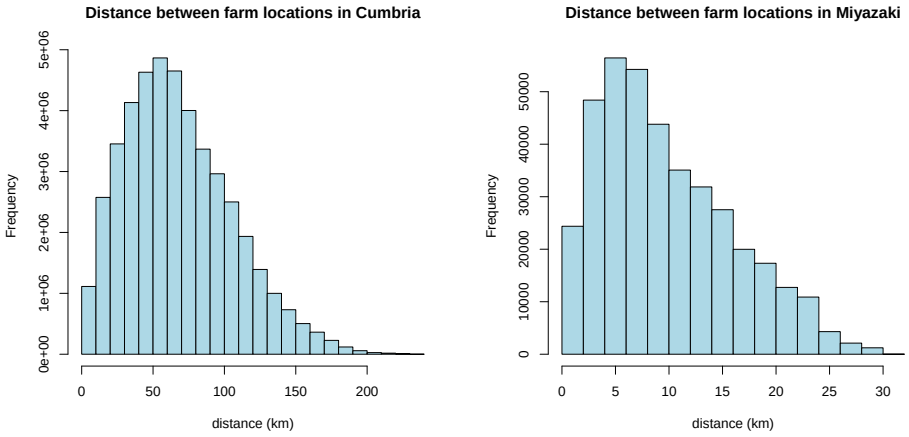
Figure 7.32: Integrated cluster curve for simulated FMD outbreaks in Miyazaki over a radius of 0:5km : (a) No intervention (b) depopulation (c) vaccination (d) depopulation and vaccination (e) depopulation and pre-emptive culling.

7.5 Discussion

Control strategies have a major impact in veterinary epidemiology. To control or eradicate a disease we generally see combinations of non-movement ban, vaccination and depopulation. We investigated five control programs; no control, depopulation, vaccination, depopulations and vaccination, and depopulation and pre-emptive culling at two geographical locations, Border counties, Great Britian and Miyazaki, Japan.

When we investigate the changes in clustering through time by the application of the cluster curve, we found that the cluster curves were similar for intervention strategies involving treatments that were applied to more than the confirmed farm (vaccination, depopulation and vaccination, and depopulation and pre-emptive culling). The integrated cluster curves told a similar story. For no control interventions, in both geographical regions, the outbreak is seen to bubble away with repeated 'bumps' of clustering. The most effective intervention methods appear to be vaccination, depopulation and vaccination, and depopulation and pre-emptive culling. For these interventions we see a bump in the early stages of the outbreak which is quickly killed off. The depopulation intervention strategy is not as effective, with a secondary outbreak within the epidemic commonly appearing. This method is more risky as there is no containment of local spread. In all cases the strongest evidence of clustering was apparent at the shortest ranges, reflecting the typical direct transmission distances for foot-and-mouth disease.

Comparing the two geographical regions we see very different structures. The spatial plots for the Border counties outbreak suggests grouping amongst the data and the outbreak is controlled without a large amount of spread (all susceptible farms were not infected). With the Miyazaki example we see that almost all farms are implicated in the outbreak in some way or another. In Figure 7.33 we looked at the distribution of the distance between the farms for the two regions. We see the distribution of the histograms are similar suggesting that there are dense areas of farms followed by farms located over larger ranges. However, when we look at the distance that the farm distribution covers these are very different. The Border county farms cover a range of 200km with the mode around 50km, while the Japanese outbreaks covers a range of 30km with a mode around 5km. In particular, only a small fraction of the Border county farm pairs are within 5km of each other. The difference in the distances between farms could be a factor in why we see significant differences in the clustering structures of the intervention strategies for the two geographical regions. This also suggest that the distance between farms can have a dramatic affect on the spread and control of a infectious disease outbreak.



(a) Border counties, Great Britain

(b) Miyazaki, Japan

Figure 7.33: Histogram of the distribution of pairwise distances between farm locations : (a)Border counties, Great Britain (b) Miyazaki, Japan.

Chapter 8

Application of the Cluster curve to Equine influenza

8.1 Introduction

Our cluster curve summarises the changes in spatial clustering through time. In the applications so far we have looked into its use with real world and simulated outbreaks of FMD in two geographical locations. In this chapter we will explore the application of the cluster curve to the 2007 outbreak of Equine influenza (EI) in Australia.

Equine influenza (EI) was first reported in horses in 1963 and is one of the major causes of respiratory infections in horses (Woodward et al., 2015) with two known subtypes. The disease is widely distributed with only eight Office Internationale des Epizooties (OIE) member countries, where passive surveillance occurs, never reporting an outbreak (Cowled et al., 2009). The major impact of EI is the secondary bacterial infection that commonly occurs.

Similar to FMD, the virus is highly contagious and can be transferred by direct and indirect methods as well as by airborne spread. Both the subtypes of EI have available vaccines, with the EI vaccine being subtype exclusive so vaccination for one does not protect for the other. Unlike FMD, slaughter is not commonly used as a control, instead, a movement ban and vaccinations are typically applied.

In this chapter we will describe the background behind EI and its 2007 outbreak in Australia. We then apply the cluster curve to better understand the epidemic.

8.2 Disease epidemiology

Equine influenza is a highly contagious respiratory disease with two known subtypes identified in horses. The disease is characterised by a fever, nasal discharge, a dry cough, lethargic demeanour, and a loss of appetite (Firestone et al., 2011; Timoney, 1996). The major complication associated with this disease, more common than with other respiratory infections, is the likelihood of secondary bacterial infections. These infections cause serious and occasionally life threatening consequences for the horse, generally the severity being greater in the young and old (Firestone et al., 2011; Cowled et al., 2009).

The virus is transmitted by direct and indirect contact and has been reported to have been transmitted by airborne spread. It has an incubation period of one to three days, with secretions of the virus occurring in as little as 24hrs, and can continue for up to 10 days. The virus is easily killed by the cold, sunlight, heat, and most common disinfectants, however, it can survive in the soil for up to two days and in water for up to two weeks. Long range spread of the virus has been recorded to have occurred over 32m via coughing and up to 8km by aerosolised wind-borne spread (Firestone et al., 2011).

The two strains of the virus do not cross-react and therefore, immunisation and antibodies for one do not protect against the other (Timoney, 1996).

8.3 Australian epidemic history

Before the 2007 outbreak, Australia was one of only three countries with a significant equine population to have remained EI free (Firestone et al., 2011). The outbreak was first linked to imported horses. Due to swift and efficient intervention methods the outbreak was restricted to only two eastern states; New South Wales (NSW) and Queensland (QLD), but not before nearly 10,000 premises were infected (Cowled et al., 2009).

The outbreak was traced back to a quarantine facility and was brought into the country on the 8th of August, 2007 via imported infected vaccinated horses from Japan. The exact method of how the disease spread out of the quarantine facility is unknown, but it is hypothesised to have spread undetected via fomites (such as clothing). The local spread first infected horses competing in an equestrian event at Maitland, near Newcastle NSW, over the weekend of the 17th to 19th August. Following the event, several horses were then transported long distances while harbouring the disease. At least one infected horse later spread the disease, the following weekend, at another horse event in Narrabri showground almost 400km away. The majority of cases were then linked to these two events or local spread, over a distances of about 28km,

in the first few weeks of the outbreak. To eradicate the disease a series of control methods were implemented including zone-based movement restrictions, contact-tracing followed by quarantine of suspected premises, targeted vaccination and on-farm biosecurity measures (Firestone et al., 2011).

During the outbreak, various dogs (different ages and breeds) that were kept in the vicinity of infected horses were also noted to have a respiratory infection (Kirkland et al., 2010). Equine influenza virus in canines has also been reported in the US (Crawford et al., 2005) and England (Daly et al., 2008).

The last case occurred in December 2007. In December 2008 Australia regained its EI free status after an extensive surveillance program observed no new cases (Cowled et al., 2009).

8.4 Literature review

Cowled et al. (2009) carried out a descriptive and cluster analysis on the 2007 EI outbreak. They found that the outbreak consisted of three key phases. The first was a dispersion phase where there was substantial spatial scattering of a few infected horses. This is why most of the final area affected by EI was determined by a few cases in the days before detection. The dispersion of the disease rapidly decreased and had mainly ceased by 1st September after the implementation of movement bans on 25th August. The second phase was characterised by local spread. In this phase minimal dispersion occurred and instead there was a large increase in the number of IPs. The movement bans restricted the long range dispersion, but transmission to neighbouring farms [1] continued. It was during this phase that vaccination control methods were implemented. The third phase was the disease fade out stage. In their cluster analysis Cowled et al. (2009) found 37 epidemiologically linked premise clusters. They found that urban clusters generally had a longer epidemic duration and shorter distances of disease spread. However, surprisingly they also found little difference in the incidence rates, cumulative incidence and reproduction rates between rural and urban regions.

Firestone et al. (2011) carried out a case control study of 200 horse premises to investigate intrinsic premise factors and biosecurity compliance factors. Intrinsic factors remain unchanged once the epidemic begins, such as descriptions of the locations and types of premises in which horses were most at risk. Biosecurity compliance factors relate to methods used to reduce the risk of contamination of a premise. They found that the most significant intrinsic factor was the proximity of a premise to its nearest IP. They found an increase likelihood of new cases

[1] We use the word 'farm' throughout this chapter as short hand for a premises with horses. There are, of course, farms without horses, and premises with horses that would not usually be described as a farm.

occurring within a 5km radius of IPs. Their analysis also found that the 10km buffer zones used were appropriate for local spread containment. With the biosecurity compliance factors, two were found significant and used in the final model. These were the presence of a footbath and daily monitoring for clinical signs. The results of their study suggest the compliance with on-farm biosecurity controls prevented the spread of EI onto a premise in high risk areas.

8.5 Data

8.5.1 Infected premises data

Information on disease control activities within Australia is recorded via the Animal Emergency Management Information Systems (ANEMIS) software. With the EI outbreak, all premises involved within NSW were given a unique ANEMIS record that contained information on the location, disease status, horse population, laboratory test results, notes on clinical signs, dates of visits, dates on status changes and the dates of first clinical signs (Cowled et al., 2008).

The majority of information that makes up the IP database was created by running progressive enquires of ANEMIS throughout the epidemic. This was then cleaned to remove duplicates and incorrect entries; the resulting dataset contained 5944 IPs. This was then combined with information on identified suspect premises (SPs) that were likely to be infected. Any SPs that were tested and came back with negative laboratory results were deleted. A total of 160 IPs were added to the IP database bringing the total to 6104 IPs. An additional 212 IPs were added to the database after laboratory results determined these premises to be historically serotype positive: antibodies suggest there was a previous infection even though clinical signs are no longer present. These farms were classed as resolved rather then infected as the active infection had passed. Altogether this brings the total number of identified IPs in the final dataset to 6316 (Cowled et al., 2008).

A limitation of this dataset outlined by Cowled et al. (2008) is that the list of IPs may be under estimated, as once movement restrictions were lifted any farms with sero-positive laboratory results and a status date after this were not included. This was because the authors took a conservative approach where if the origin of the horse, during active infection, was unknown the event was not included.

8.5.2 Horse population data

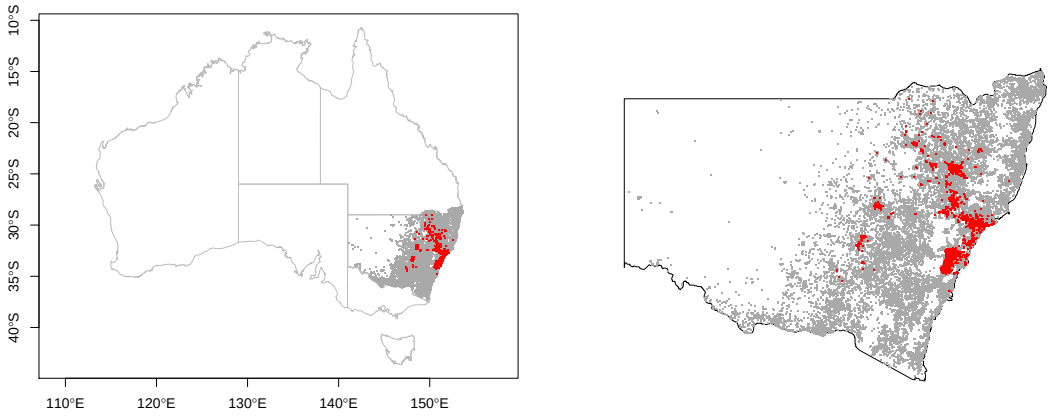
When evaluating an outbreak it is important to know where the susceptible farms (i.e. premises with horses) are located. Before the outbreak there was no official government record of horse location and population within Australia. During the outbreak the NSW local disease control centre compiled a horse population database. They pooled resources from the Rural Lands Protection Board (RLPB) database, the Australian Horse Industry Council Database (AHIC), the Equine Influenza Registration Database (EIRDB), the ANEMIS data base, and other smaller databases such as travelling horse statements and the yellow pages. This pooling of resources produced a dataset with 102,000 records. This was later cleaned and reduced down to 51,615 records. Cleaning the data involved the removal of duplicates and incomplete records (for example when there was insufficient information to geocode the data) (Cowled et al., 2008).

Despite the efforts that went in to compiling this dataset, its creators concede that it almost certainly does not capture the entire horse population (Cowled et al., 2008).

8.6 Application of the Cluster curve

For our application of the cluster curve to the equine influenza outbreak we focus our attention on premises that had the status infected, resolved or suspect. An infected premises was any premises that had one or more resident horses with a laboratory confirmed positive PCR, or a premises where blood sampled un-vaccinated horses were sero-positive for antibodies with evidence of active infection. A suspected premises was any premises with no identified link to a confirmed farm or other exposure, but where horses displayed clinical signs or a premises where inconclusive reactions had been detected. A resolved premises was any premises which was an infected premises, a dangerous contact premises, or a suspected premises on which an investigation has concluded there is no longer an active infection. As the outbreak progressed a farm's status changed as the active infection passed. At the end of the outbreak all infected farms were classed as resolved. The spatial distribution of these cases is shown in Figure 8.1.

During the outbreak information was collected on all farms within NSW and the surrounding areas. The location of theses farms, shown in grey in the spatial plot, was used to create an estimated farm density used in the computation of the inhomogeneous K-function and hence the cluster curve. This kernel estimate was implemented using a bandwidth of $h_x = 0.275^\circ$ (25km) (smoothing in the x direction) and $h_y = 0.303^\circ$ (28km) (smoothing in the y direction), chosen using Scott's (Scott, 1992) rule of thumb bandwidth selector. The resulting intensity function is shown in Figure 8.2.



(a) Australia (b) NSW

Figure 8.1: Map showing the spatial plot of EI cases. Grey indicates the location of all susceptible farms within NSW and red all susceptible, infected and resolved EI farms

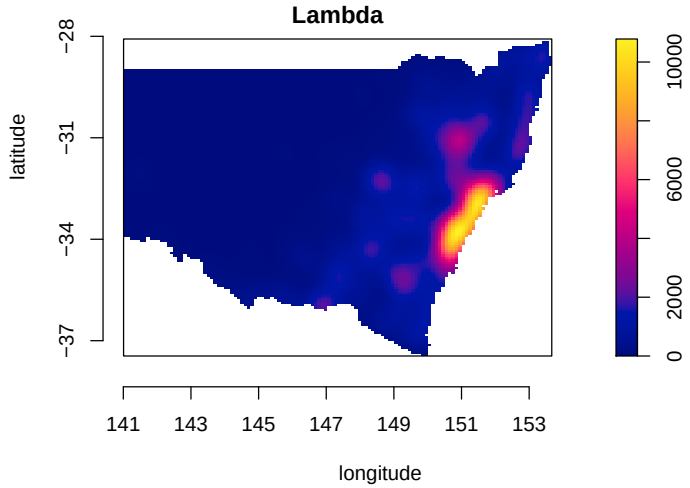


Figure 8.2: Estimated density for horse farms

For our dataset we classed day 1 of the outbreak as the 19/07/2007. In total our dataset ranged from day 1 to day 851, however after day 158 there were only 3 cases, one on each of the days 231, 457, and 851. These three events have a status of resolved, with the last point occurring after Australia regained its EI-free status. This suggests that these farms

were discovered to have had EI after the active infection had passed. For our analysis we only focus on days prior to 160. Our date record for each event is based on either the date of the first clinical signs or, if this is unknown, the earliest visit date recorded in ANEMIS at which the property was given a status minus one day.

In Figure 8.3 and 8.4 we display the spatial-temporal distribution of events every 20 days. In the first 20 days we see only a few cases, however, by day 40 we see a substantial increase in the number of cases. From day 40 to day 100 most of the spread and events appear to have occurred. As expected, we see the majority of cases occur within the high density areas for the horse population.

Epidemic curve

The epidemic curves are shown in Figure 8.5 with the daily curve on the left and 7-day moving window on the right. We can see that after the initial few cases the virus spread rapidly. The number of new cases remained large for around 40 days and then died off rapidly.

Cluster curve

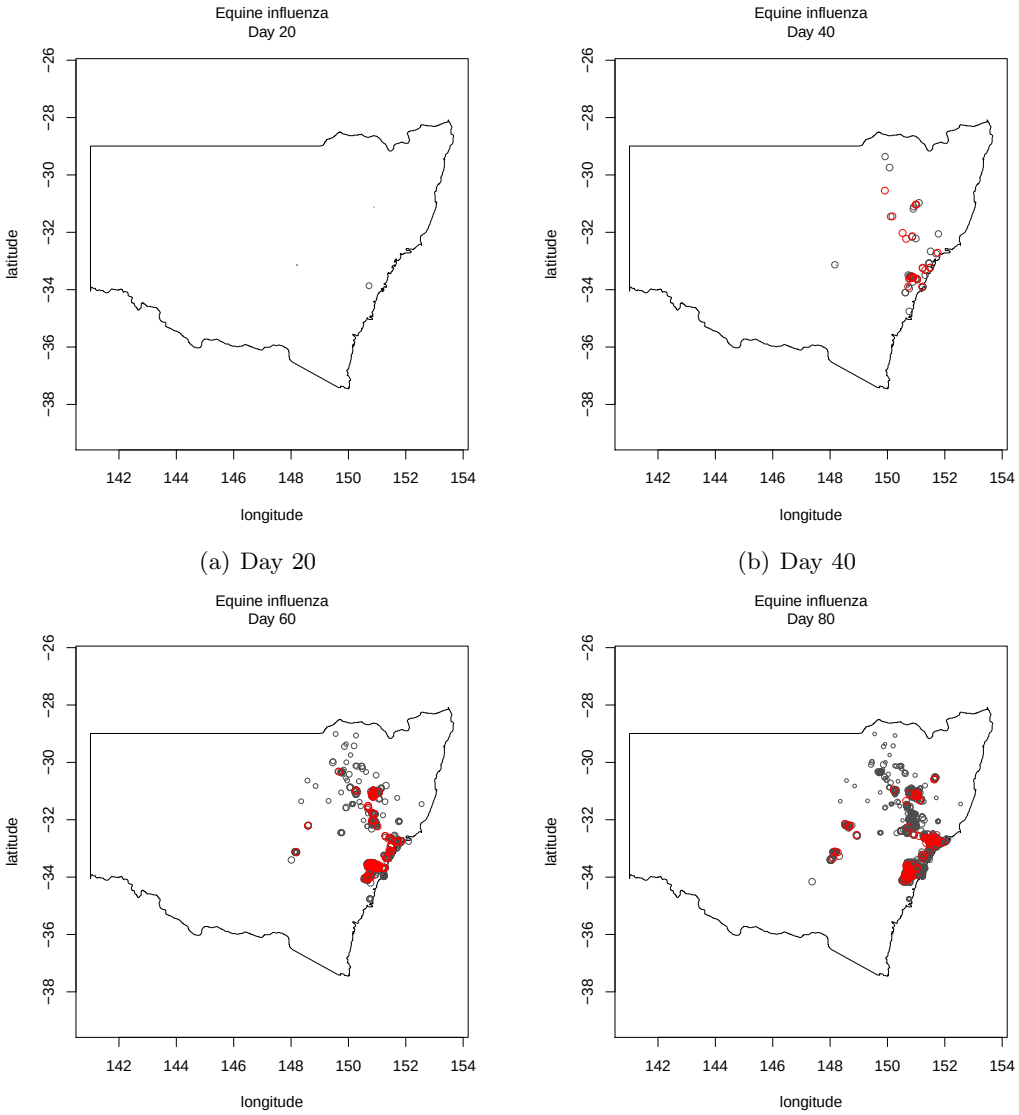
The shapefiles for this dataset used longitude and latitude. To keep the units consistent with the previous work we used a scaling conversion of 1° longitude and latitude equals 92km.

The application of cluster curve to the equine data is shown in Figure 8.6. We evaluate this at radiuses of 1:5km at 1km increments. Since our intention is that the cluster curve be displayed alongside the epidemic curve, we illustrate this at a 2km radius (Fig 8.7).

From the analysis of these curves we see a similar structure at all radiuses monitored. The beginning and end of the outbreak is characterised by a significant degree of clustering. In the last 50 days of the outbreak the degree of clustering is at its greatest.

The initial spread of the outbreak was determined by the movement of horses from two shows (day 31-33 and day 38-40). We see that following these shows there was a period of significant local clustering as the initial infection from the shows created pockets of local clustering around their farms.

For the control of EI, movement restrictions and vaccination were applied. The initial application of vaccination occurred as ring vaccination (towards the end of the outbreak blanket vaccination was also applied) from the 28th September (day 73). It took approximately two weeks for immunity to develop. We therefore, would expect to see the implications of vac-



(a) Day 20

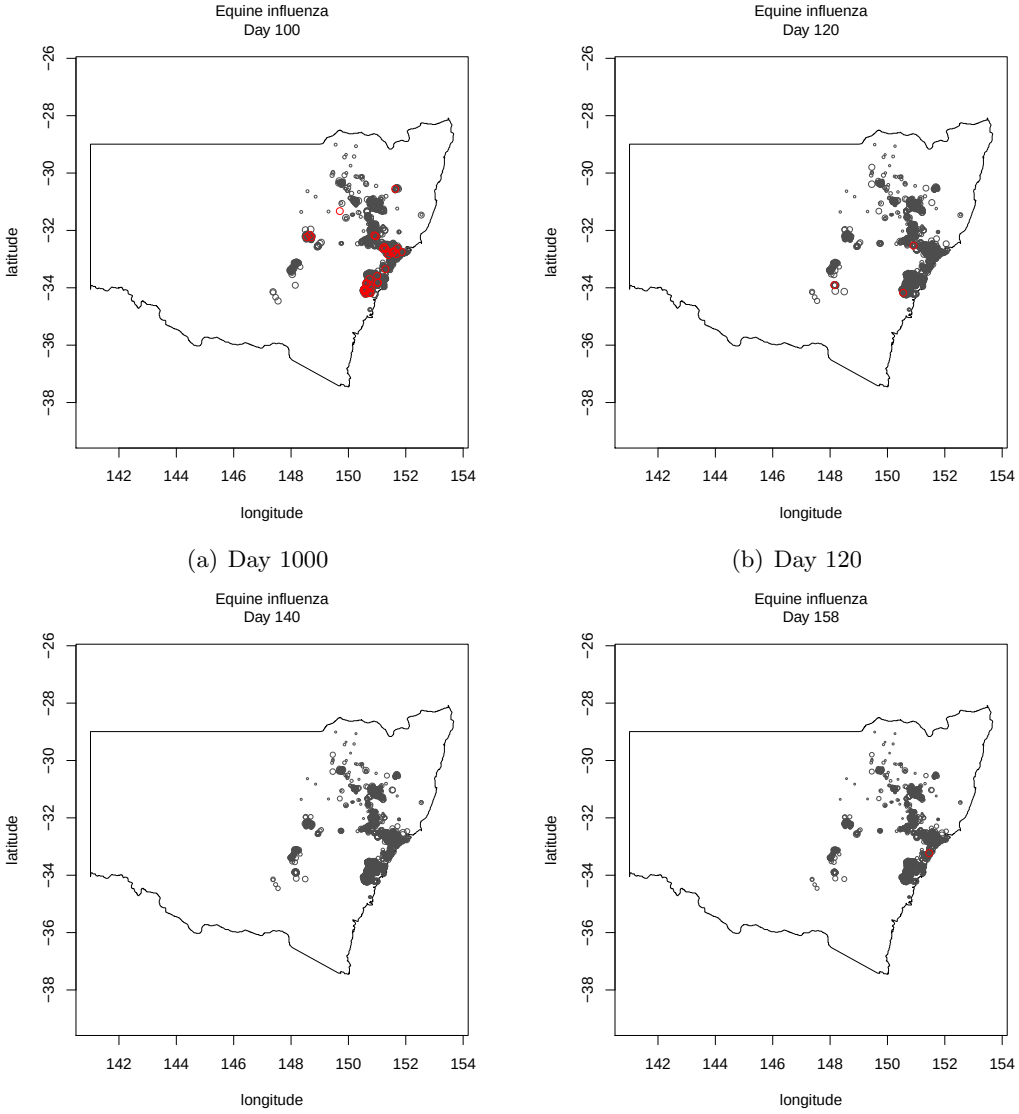
(b) Day 40

(c) Day 60

(d) Day 80

Figure 8.3: Spatial plots of the outbreak of EI, Australia 2007, for 20 day windows. Red represents new cases recorded on the day and grey the previous cases. The plotting size indicates how recently the case occurred, with large symbols indicating very recent events and small symbols earlier ones.

cination after day 90 (Cowled et al., 2009; Perkins et al., 2011). The period in which the vaccination method takes affect coincides with our increased degree of clustering. This is most likely due to local spread being contained within the buffer rings.



(a) Day 1000

(b) Day 120

(c) Day 140

(d) Day 158

Figure 8.4: Spatial plots of the outbreak of EI, Australia 2007, for 20 day windows. Red represents new cases recorded on the day and grey the previous cases. The plotting size indicates how recently the case occurred, with large symbols indicating very recent events and small symbols earlier ones.

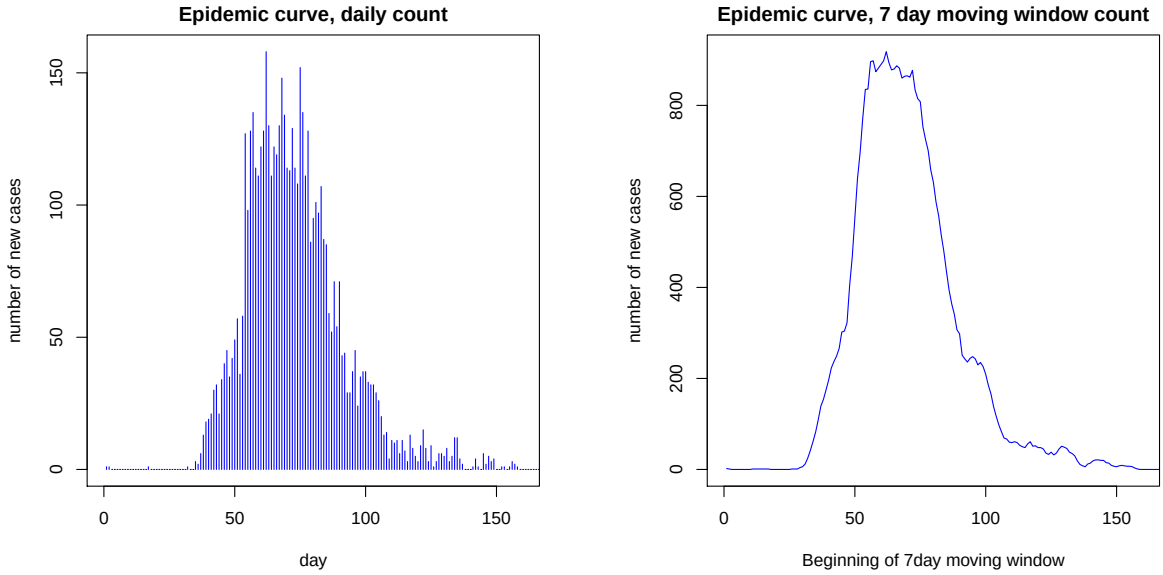
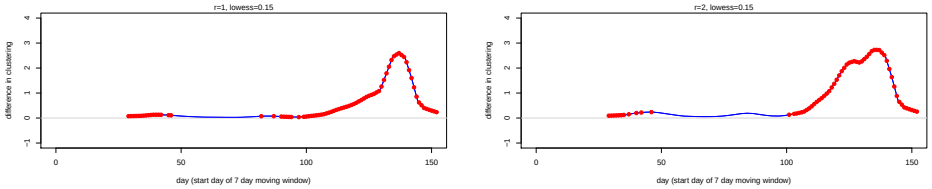
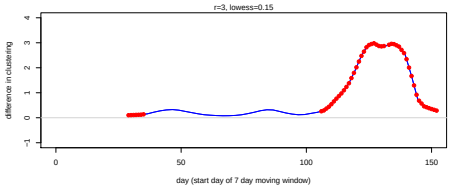


Figure 8.5: Epidemic curve of EI Australia 2007. Left: daily epidemic curve, right: 7-day moving window epidemic curve.

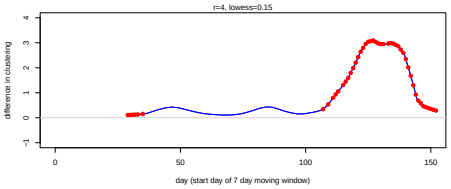


(a) 1km

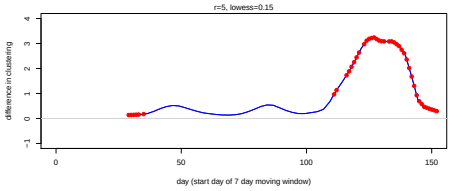
(b) 2km



(c) 3km



(d) 4km



(e) 5km

Figure 8.6: Cluster curve for EI outbreak at 1:5km at 1km intervals

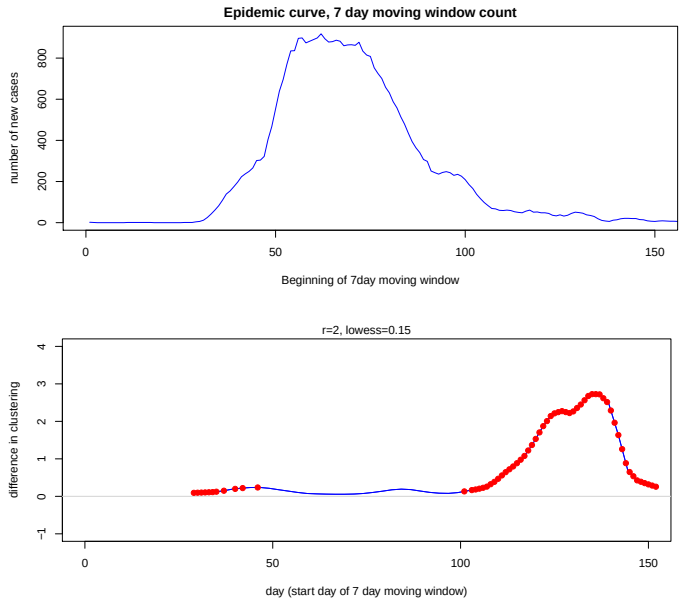


Figure 8.7: Epidemic curve and cluster curve for 2km radius

8.7 Discussion

In this chapter we continued our illustration of the use of the cluster curve with its application to the 2007 epidemic of equine influenza in Australia.

The outbreak lasted 158 days. It was contained to two eastern states (NSW, QLD) and affected nearly 10,000 premises. The initial spread was determined by the infection of two horse shows on days 31-33 and 38-40. A method of movement restrictions and vaccination was implemented to eradicate the disease. The vaccinations were initially applied as buffer rings from day 73. With a two week time lag between vaccination and immunity, the peak of the outbreak occurred before the vaccination method could have an effect (Cowled et al., 2009; Perkins et al., 2011).

In previous analyses of the 2007 equine influenza outbreak, the data were found to be highly clustered with 37 epidemiologically linked clusters (Cowled et al., 2009). This resonates with our results where the cluster curve suggests significant grouping of events in the beginning and latter stages of the outbreak at radiuses of 1:5km. We see that the period preceding the horse shows was characterised by significant clustering via local spread. The degree of clustering is significant, however, not large, as long range transmission occurred from horses returning to their farms after travelling to attend the shows. We saw a vast increase in the degree of clustering difference after the immunity of vaccination took effect. This suggested that the vaccination and movement restrictions were successful in controlling spread over ranges greater than 10km, with the majority of infection occurring as a result of local spread within the buffer rings.

The findings in the latter stages of the epidemic, once the vaccination intervention method had taken effect, mirror the results from the simulated epidemics in Chapter 7. Once a vaccination method is in effect, we observed a period of significant clustering during which local spread occurs within the buffer rings before the outbreak is effectively shut down.

Chapter 9

Conclusion and discussion

In epidemiology we are interested in the analysis of disease events in regard to possible risk factors. While many of the methods of analysis are equivalent for human and veterinary cases, veterinary data has the added difficulty of 'human factor'. This is the need for humans to first detect and then report the disease, and once detected, humans generally intervene. This process requires multiple steps for an event to be recorded and therefore, the true quantity of animal disease can be under-reported. Once a disease is notified to the appropriate authorities a method of intervention is generally applied; this can be as little as movement restrictions to much more severe measures like the depopulation of farms. These methods limit and control the spread which can dramatically alter the progression of disease.

In this thesis we have looked into tools to address data quality issues as well as improving the analysis of the spatial spread through time. To detect low level anomalies, where the number of cases is less then expected, we looked at the application of exceedance probabilities in spatio-temporal models for routinely recorded areal disease counts. We illustrated the use of this via the endemic outbreak of foot-and-mouth disease in Viet Nam over the time frame January 2006 to December 2008. We also investigated the use of second-order properties of spatial point processes to create our cluster curve and integrated cluster curve for the evaluation of changes in spatial clustering through time for event data. The application of the cluster curve was applied in different interactive visual displays and to real and simulated epidemic outbreaks of foot-and-mouth disease and an epidemic outbreak of equine influenza.

Exceedance probabilities

With any analysis, the reliability of the results depends on the quality of the data. While residual analysis is an effective approach to detecting anomalies for many types of models (such as linear regression with fixed effects), with complex models this methodology can be problematic. Hierarchical models with random effects can provide great flexibility in the modelling process, but can also mask extreme events in a standard residual analysis. Our aim was to demonstrate a wider use of exceedance probabilities (already previously used to detect unusually high levels of relative risk and detection of anomalous clusters) in the detection of unexpectedly low counts of reported cases of disease in routinely recorded animal health data. Exceedance probabilities can be applied at any level of a model and describe the probability that a random term lies in the extreme tails, in our case the lower tail of the specified distribution. Any term that has a high probability of being in the lower tail suggests a significantly lower number of cases than expected.

In our application of exceedance probabilities to the endemic outbreak of foot-and-mouth disease in Viet Nam over a three year period (2006-2008) we saw that the use of a general method of residual analysis (squared Pearson residuals) did not indicate significant anomalies in the data from any of the regions of the country. In the analysis of the lower tail exceedance probabilities, however, we found several provinces where the number of foot-and-mouth disease positive communes was unusually low given the provincial characteristics. Interestingly, the provinces identified as anomalous in 2006 formed a geographical cluster. Our model did not include a representation of spatial correlation and therefore, this is an aspect of the data and not an artefact of the model. We found that areas identified over the time period as possible anomalies coincided with major movement pathways of cattle within Viet Nam. These could potentially be explained by unmeasured covariates – we recognise that the number and resolution of the explanatory variables available is a limitation of our study. Nonetheless, it is possible that the unexpectedly low counts arise from under-reporting of foot-and-mouth disease incidence.

Overall, the results of our analysis warrant further investigation by the relevant authorities, and demonstrate the practical usefulness of exceedance probabilities.

The cluster curve

Our tools; the cluster curve and integrated cluster curve use second-order properties of spatial point processes, specifically the inhomogeneous K-function, to investigate changes in spatial clustering through time. We did not aim to detect clusters of disease in space and time –

there are already several existing methods to do so. Our aim was to instead provide a simple means of describing how the pattern of spatial clustering varies with time. The cluster curve is calculated by evaluating over a sequence of time windows the difference between the K-function computed from the data within the window and the theoretical K-function under the 'no clustering' assumption. We intend that the cluster curve be used alongside epidemic curves.

To illustrate the motivation behind our cluster curves we simulated two sets of pairs of outbreak datasets one based on a spatially homogeneous underlying population and the other, a spatially inhomogeneous population. For each pair of outbreaks, one was generated by a Poisson point process and the other by a cluster Poisson point process. Experiments on simulated data demonstrated the potential for the cluster curve to uncover interesting properties of the spatio-temporal epidemiology that are not apparent in the epidemic curve alone.

We proposed several methods for interactive displays of the cluster curve. These are the use of sliders to easily switch between radiuses of clustering, and the creation of a web interface. The use of a web application will allow the tool to be used by a wider audience as no coding is required. Researchers can upload their own data and then apply the cluster curve.

Application of the cluster curve to epidemic outbreaks of foot-and-mouth disease and equine influenza

We applied our cluster curve to two epidemics of foot-and-mouth disease, and an outbreak of equine influenza. These are all examples of an epidemic in a country previously holding foot-and-mouth disease or equine influenza free status. The 2001 outbreak of foot-and-mouth disease in England infected over 2000 premises and lasted 241 days before eradication via non-movement bans, depopulation and pre-emptive culling of any susceptible premise. The 2010 outbreak of foot-and-mouth disease in Japan resulted in 292 farms becoming infected and was controlled with depopulation and buffer ring vaccination. The 2007 outbreak of equine influenza in Australia affected nearly 10,000 premises before being effectively controlled and eradicated through various methods of vaccination (primarily ring vaccination).

To illustrate how the cluster curve can inform the user on the changing spatial structure of an epidemic we compared the features uncovered by the cluster curve with knowledge gained from previous analyses. For example, we know that foot-and-mouth disease is a highly infectious disease and that the English outbreak was determined to have very high levels of localised spatial spread within 3km of an infected premises and the majority of localised spread occurring within 10km. In the analysis of the cluster curves we found the majority of the outbreak is characterised by significant clustering occurring at a higher degree at distances

of less than 4km. Therefore, conclusions drawn from previous extensive analyses can easily be obtained quite directly from evaluation of the cluster curve.

Intervention methods

In the real world animal diseases are not left to run their natural course but instead intervention methods are put into action. Furthermore, there is far more flexibility in controlling an animal disease than a human one, since options like depopulation and pre-emptive culling are available. With the control of disease in veterinary scenarios we generally see a combination of the methods, for example non-movement bands with depopulation and vaccination or depopulation and pre-emptive culling. The three epidemics we looked at used: vaccination, depopulation, vaccination and depopulation, and pre-emptive culling to control and eradicate the disease. We investigated the impact of the control methods on the changes in spatial clustering through time. This was done by simulating two sets of outbreaks using the software InterspreadPlus, one based on the geographical and population density of the border counties, United Kingdom and the other Miyazaki, Japan. For each set we simulated five datasets, four with the applied intervention strategies (depopulation, vaccination, depopulation and vaccination and depopulation and pre-emptive culling) and one left without intervention so that we could see the natural progression of the disease.

When we investigate the changes in clustering through time by the application of the cluster curve, we found that methods where treatments were applied to more then the confirmed farm (vaccination, depopulation and vaccination and depopulation and pre-emptive culling) had similar curve structures, and for the integrated cluster curve we saw similar results. We found a significance degree of clustering at the smaller radiuses.

The conclusions found by analysing the different structures of the intervention methods provides knowledge of the characteristics for a particular method. This knowledge can then be applied to real world examples, and enables us to draw conclusions about possible reasons behind certain structure. For example, with the equine influenza outbreak in Australia we see a large degree of clustering in the latter stage (after day 100) of the outbreak. We know that from day 90, immunity from vaccination starts to take effect, therefore, creating buffer zones. From the analysis of vaccination buffer rings on the simulated data we would expect a large degree of clustering for a period (as cases are trapped within the buffer boundary) before the disease is eliminated. What we expected for a vaccinated outbreak based on simulations is what is seen in the real world case of equine influenza.

Future work

With our analysis of endemic foot-and-mouth disease in Viet Nam, possible future work includes the analysis of current data to investigate if the pattern of under-reporting is persisting. We might also develop models that specifically account for under-reporting on the cattle routes, so as to be able to estimate the magnitude of the effect.

Our cluster curve provides a promising tool for analysing the changes in spatial clustering through time. However, there are always opportunities for improvement. Future work could include the investigation of mathematical expressions for approximating confidence intervals for the K function so as to avoid the need for computationally expensive envelopes. Also, further consideration would be helpful into the use of the best scaling factor (denominator) for both the cluster curve and the integrated cluster curve. There is no right or wrong answer for this issue, but it may prove that some choices facilitate interpretation.

To fully investigate the impact of intervention methods on spatial clustering through time, further investigation is required into different methods of outbreak simulations and different veterinary infectious diseases. This would help to assess whether the results we found in our investigation carry over into the wider field of animal diseases and controls.

Bibliography

Bachrach, H. (1968), 'Foot-and-mouth disease', *Annual Reviews in Microbiology* **22**(1), 201–244.

Baddeley, A., Diggle, P., Hardegen, A., Lawrence, T., Milne, R. and Nair, G. (2014), 'On tests of spatial pattern based on simulation envelopes', *Ecological Monographs* **84**(3), 477–489.

Baddeley, A., Moller, J. and Waagepetersen, R. (2000), 'Non- and semi-parametric estimation of interaction in inhomogeneous point patterns', *Statistica Neerlandica* **54**(3), 329–350.

Baddeley, A. and Turner, R. (2005), 'spatstat: An R package for analyzing spatial point patterns', *Journal of Statistical Software* **12**(6), 1–42.
URL: *http://www.jstatsoft.org/v12/i06/*

Baddeley, A., Turner, R., Møller, J. and Hazelton, M. (2005), 'Residual analysis for spatial point processes (with discussion)', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(5), 617–666.

Baddeley, A., Turner, R. and Rubak, E. (2015), *spatstat: Spatial point pattern analysis, model-fitting, simulation, tests*. R package version 1.41-1.
URL: *http://CRAN.R-project.org/web/packages/spatstat*

Baxter, M. (2014), 'Large regional map', electoralcalculus.co.nz.

Beck, L., Lobitz, B. and Wood, B. (2000), 'Remote sensing and human health: new sensors and new opportunities', *Emerging Infectious Diseases* **6**(3).

Becker, R., Chamber, J. and Wilks, A. (1988), *The new S Language*, Wadsworth and Brook/Cole.

Bergevoet, R. and van Asseldonk, M. (2014), 'Economics of eradicating foot-and-mouth disease epidemics with alternative control stratergies', *Arch. med. vet.* **46**(3).

Besag, J. (1974), 'Spatial interaction and the statistical analysis of lattice systems', *Royal Statistical society* **36**(2), 192–236.

Besag, J., York, J. and Mollie, A. (1991), 'Bayesian image restoration, with two applications in spatial statistics (with discussion)', *Annals of the institute of statistical mathematics* **43**, 1–59.

Bessell, P., Shaw, D., Savill, N. and Woolhouse, M. (2008), 'Geographic and topographic determinants of local FMD trasmission applied to the 2001 UK FMD epidemic.', *BMC Veterinary Research* **4**(40).

Bessell, P., Shaw, D., Savill, N. and Woolhouse, M. (2010), 'Statistical modeling of holding level susceptibility to infection during the 2001 foot and mouth disease epidemic in Great Britian', *International journal of infectious diseases* **14**.

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

Bivand, R., Pebesma, E. and Gomez-Rubio, V. (2008), *Applied Spatial Data Analysis with R*, Springer.

blogspot.co.nz (2010), 'Japan Miyazaki Tokyo', http://aesisgoinjapan.blogspot.co.nz/2010/11/history-of-miyazaki-japan.html.

Bokund, A., Halasa, T., Christiansen, L. and Enoe, C. (2012), Influence of livestock markets on the spread of FMD, Optimizing the control of foot-and-mouth disease in Denmark by simulation.

Bowman, A. and Azzalini, A. (1997), *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations.*, Oxford: Oxford University Press.

Branscum, A., Perez, A. and Thurmond, M. (2008), 'Bayesian spatiotemporal analysis of foot-and-mouth disease data from the republic of turkey', *Epidemiol. Infect.* (136), 833–842.

Brooks, S. and Gelman, A. (1998), 'Alternative methods for monitoring convergence of iterative simulations', *Journal of computational and graphical statistics* **7**, 434–455.

Brunsdon, C. and Comber, L. (2015), *An Introduction to R for spatial analysis and mapping.*, Sage.

by Rstudio, S. (2015), 'Shinyapps.io', http://www.shinyapps.io/.

Carlin, B. and Louis, T. (2000), *Bayes and Empirical bayes methods for data analysis*, 2 edn, New York: Chapman & Hall/ CRC Press.

Carpenter, T. (2001), 'Methods to investigate spatial and temporal clustering in verterinary epidemiology', *Preventive Veterinary Medicine* **48**, 303–320.

Centers for Disease Control, . (1990), Guideline for investigating clusters of health events, Morbidity and Mortality Weekly report RR-11, US Department of Health and Human services, Atlanta, GA.

CGIAR-CSI (2013), *NASA Shuttle Radar Topographic Mission (SRTM) 90 metre Digital Elevation Database*. Release 4.1.
URL: *http://www.cgiar-csi.org/data/srtm-90m-digital-elevation-database-v4-1*

Chang, W. e. a. (2015), 'Shiny: Web application framework for R', http://shiny.rstudio.com/. An RStudio project.

Christensen, J. (2001), 'Epidemiological concepts regarding disease monitoring and surveillance', *Acta vet. scand.* **94**.

Clark, P. and Evans, F. (1954), 'Distance to nearest neighbor as a measure of spatial relationships in populations', *Ecology Society of America* **35**(4), 445–453.

Cleveland, W. (1979), 'Robust locally weighted regression and smoothing scatterplots', *J. American statistical association* **74**.

Cocks, P., Abila, R., Bouchot, A., Benigno, C., Morzaria, S., Inthavong, P., Van Long, N., Bourgeois-Luthi, N., Scoizet, A. and Sieng, S. (2009), FAO ADB and OIE SEAFMD study on cross-border movement and market chains of large ruminants and pigs in the Greater Mekong sub-region, Technical report, FAO ADB and OIE SEAFMD, Bangkok, Thailand.

Cowled, B., Garner, G. and Moloney, B. (2008), NSW EI data set.

Cowled, B., Ward, M., Hamilton, S. and Garner, G. (2009), 'The equine influenza epidemic in Austrlia: Spatial and temporal decriptive analyses of a large propagating epidemic.', *Preventive Veterinary Medicine* **92**.

Crawford, P., Dubovi, E., Castleman, W., Stephenson, I., Gibbs, E., Chen, L., Smith, C., Hill, R., Ferro, P., Pompey, J., Bright, R., Medina, M.-J., Group, I. G., Johnson, C., Olsen, C., Cox, N., Klimov, A., Katz, J. and Donis, R. (2005), 'Transmission of equine influenza virus to dogs', *Science* **310**.

Cuzick, J. and Edwards, R. (1990), 'Spatail clustering for inhomogeneous populations', *Journal of the Royal Statistical Society* **52**(1), 73–104.

Daly, J., Blunden, A., MacRae, S., Miller, J., Bowman, S., Kolodziejek, J., Nowotny, N. and Smith, K. (2008), 'Transmission of equine influenza virus to English foxhounds', *Emerging infectious diseases* **14**(3).

David, F. and Barton, D. (1966), 'Two space-time interaction tests for epidemicity', *British journal of preventive & social medicine* **20**(1), 44–48.

Davies, G. (2002), 'Foot and mouth disease', *Research in Veterinary science* **73**, 195–199.

Davies, T. and Hazelton, M. (2013), 'Assessing minimum contrast parameter estimation for spatial and spatiotemporal log-Gaussian Cox processes', *Statistica Neerlandica* in press.

De La Rocque, S., Michel, V., Plazanet, D. and Pin, R. (2004), 'Remote sensing and epidemiology: examples of applications for two vector-borne diseases', *Comparative immunology, microbiology and infectious diseases* .

Diggle, P. (2000), Spatial statistics for environmental epidemiology, Technical report, Medical statistic unit, Lancaster University.

Diggle, P. (2003), *Statistical Analysis of Spatial Point Patterns*, second edn, Hodder Arnold.

Diggle, P. (2005), 'Spatio-temporal point processes: methods and applications', *Johns Hopkins University, Dept. of Biostatistics Working Papers* .

Diggle, P., Rowlingson, B. and Su, T. (2005), 'Point process methodology for on-line spatio-temporal disease surveillance', *Environmetrics* **16**, 423–434.

Diggle, P., Thomson, M., Christensen, O., Rowlingson, B., Obsomer, V., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Kamgno, J., Remme, J., Boussinesq, M. and Molyneux, D. (2007), 'Spatial modelling and the prediction of Loa Loa risk: decision making under uncertainty', *Annals of tropical medicine and parasitology* **101**(6), 499–509.

Dixon, P. (2012), 'Ripley's K function', Written for the Encyclopedia of Environmetrics, 2nd ed.

Dunn, P. and Smyth, G. (1996), 'Randomized quantile residuals', *Journal of Computational and Graphical Statistics* **5**(3), 236–244.

Durr, P. and Gatrell, A. (2004), *GIS and Spatial analysis in veterinary science*, CABI publishing.

Ederer, F., Myers, M. and Mantel, N. (1964), 'A statistical problem in space and time: do leukemia cases come in clusters?', *Biometrics* **20**(3), 626–638.

Ellis-Iversen, J., Smith, R., Gibbens, J., Sharpe, C., Dominguez, M. and Cook, A. (2011), 'Risk factors for transmission of foot-and-mouth disease during an outbreak in southern England in 2007', *Veterinary Record* .

Firestone, S., Schemann, K., Toribio, J.-A., Ward, M. and Dhand, N. (2011), 'A case-control study of risk factors for equine influenza spread onto horse premises during the 2007 epidemic in Australia', *Preventive Veterinary Medicine* **100**.

Francois, G. and Raphael, P. (1999), 'On explicit formulas of edge correction for Ripley's K-function', *Journal of vegetation science* **10**, 433–438.

Gabriel, E., Diggle, P. and Rowlingson, B. (2014), *stpp:Space-Time Point Pattern simulation, visualisation and analysis*. R package version 1.0-4.
URL: *http://CRAN.R-project.org/package=stpp*

Gabriel, E., Rowlingson, B. and Diggle, P. (2013), 'stpp: An R package for plotting, simulationd and analyzing spatio-temporal point patterns.', *Journal of Statistical Software* **53**.

Gail, M. H. (1991), 'A bibliography and comments on the use of statistical models in epidemiology in the 1980s', *Statistics in Medicine* **10**(12), 1819–1885.

Gatrell, A., Bailey, T., Diggle, P. and Rowlingson, B. (1996), 'Spatial point pattern analysis and its application in geographical epidemiology.', *Royal Geographical society* **21**(1), 256–274.

Geering, W. and Lubroth, J. (2002), 'Preparation of foot-and-mouth disease contingency plans', *FAO animal health manual* **16**.

Gelman, A., Jakulin, A., Pittau, M. and Su, Y.-S. (2008), 'A weakly informative default prior distribution for logistic and other regression models', *The Annals of Applied Statistics* pp. 1360–1383.

Gilks, W., Richardson, S. and Spiegelhalter, D. (1996), *Markov chain Monte Carlo in practice*, Vol. 2, CRC press.

GitHub (2015), 'rungithub', https://github.com/.

Glass, G., Cheek, J., Patz, J., Shields, T., Doyle, T., Thoroughman, D., Hunt, D., Enscore, R., Gage, K., Irland, C., Peters, C. and Bryan, R. (2000), 'Using remotely sensed data to identify areas at risk for hantavirus pulmonary syndrome', *Emerging Infectious Diseases* **6**(3).

Gleeson, L. (2002), 'A review of the status of foot and mouth disease in South-East Asia and approaches to control and eradication', *Revue scientifique et technique-Office international des épizooties* **21**(3), 465–472.

Graham, A., Atkinson, P. and Danson, F. (2004), 'Spatial analysis for epidemiology', *Acta Tropica* .

Green, P. and Silverman, B. (1994), *Nonparametric regression ang generalized linear models: a roughness penalty approach*, Chapman and Hall.

Grubman, M.J. abd Baxt, B. (2004), 'Foot-and-mouth disease', *Clin. Microbiol. Rev* **17**.

Hay, S., Snow, R. and Rogers, D. (1998), 'From predicting mosquito habitat to malaria seasons using remotely sensed data: practice, problems and perspectives', *Parasitology Today* **14**(8).

Hay, S., Tucker, C., Rogers, D. and Packer, M. (1996), 'Remotely sensed surrogates of meteorological data for the study of the distribution and abundance of arthropod vectors of disease', *Annals of tropical medicine and parasitology* **90**(1), 1–19.

Hayama, Y., Muroga, N., Nishida, T., Kobayashi, S. and Tsutsui, T. (2012), 'Risk factors for local spread of foot-and-mouth disease, 2010 epidemic in Japan', *Research in Veterinary Science* **93**, 631–635.

Haydon, D., Kao, R. and Kitching, R. (2004), 'The UK foot-and-mouth disease outbreak- the aftermath', *Nature review Microbiology* **2**.

Hazelton, M. and Marshall, J. (2009), 'Linear boundary kernels for bivariate density estimation', *Statistics & Probability Letters* **79**(8), 999–1003.

Hijmans, R. and van Etten, J. (2013), *raster: Geographic data analysis and modeling.* R package version 2.1-16.
URL: *http://CRAN.R-project.org/package=raster*

Hodges, J., Carlin, B. and Fan, Q. (2003), 'On the precision of the conditionally autoregressive prior in spatial models', *Biometrics* **59**, 317–322.

Hosmer, D. and Lemeshow, S. (2000), *Applied logistic regression*, Wiley, New York.

Hosmer, D.W. nd Hosmer, T., Le Cessie, S. and Lemeshow, S. (1997), 'A comparison of goodness-of-fit tests for the logistic regression model', *Statistics in Medicine* **16**, 965–980.

Jacoby, W. (2000), 'Loess: a nonparametric, graphical tool for depicting relationships between variable', *Electoral studies* **19**.

Jin, X., Carlin, B. and Banerjee, S. (2005), 'Generalized hierarchical multivariate CAR models for areal data', *Biometrics* **61**, 950–961.

Keeling, M. (2005), 'Models of foot-and-mouth disease', *Proceedings of the royal society* **272**, 1195–1202.

Kirkland, P., Finlaison, D., Crispe, E. and Hurt, A. (2010), 'Influenza virus transmission from horses to dogs, Australia', *Emerging Infectious Diseases* **16**(4).

Knowles, N., He, J., Shang, Y., Wadsworth, J., Valdazo-Gonzalez, B., Onosato, H., Fukai, K., Morioka, K., Yoshida, K., Cho, I., Kim, S., Park, J., Lee, K., Luk, G., Borisov, V., Scherbakov, A., Timina, A., Bold, D., Nguyen, T., Paton, D., Hammond, J., Liu, X. and King, D. (2012), 'Southeast Asian foot-and-mouth disease viruses in Eastern Asia', *Emerging Infectious Diseases* **18**(3).

Knox, E. and Bartlett, M. (1964), 'The detection of space-time interactions', *Applied Statistics* pp. 25–30.

Kulldorff, M. (1997), 'A spatial scan statistic', *Communications in statistics- theory and methods* **26**(6), 1481–1496.

Kulldorff, M. (2001), 'Prospective time periodic geographical disease surveillance using a scan statistic', *Journal of the Royal statistical society, A* **164**, 61–72.

Kulldorff, M. (2015), 'SaTScan: Software for the spatial, temporal, and space-time scan statistics', http://www.satscan.org/.

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

Kulldorff, M. and Information Management Services, I. (2009), *SaTScan v8.0.1: Software for the spatial and space-time scan statistics.*
URL: *http://www.satscan.org/*

Kulldorff, M. and Nagarwalla, N. (1995), 'Spatial disease clusters: detection and inference', *Statistics in medicine* **14**, 799–810.

Lawson, A. (2006), *Statistical methods in spatial epidemiology*, second edn, john wiley and sons LTD.

Lawson, A. (2010), 'Hotspot detection and clustering: ways and means', *Environmental and Ecological Statistics* **17**(2), 231–245.

Lawson, A. B. (2013), *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*, Vol. 32, CRC Press.

Lawson, A., Browne, W. and Vidal Rodeiro, C. (2003), *Disease mapping with WinBUGS and MLwiN*, John Wiley and Sons Ltd.

Lawson, A. and Williams, F. (2001), *An introductory guide to disease mapping*, John Wiley and sons, LTD.

Lawson, A. and Zhou, H. (2005), 'Spatial statistical modelling of disease outbreaks with particular reference to the UK foot and mouth disease (FMD) epidemic of 2001', *Preventtive veterinary medicine* **71**, 141–156.

Le, V., Nguyen, T., Park, J.-H., Kim, S.-M., Ko, Y.-J., Lee, H.-S., Nguyen, V., Mai, T., Do, T., Cho, I.-S. and K-W., L. (2010), 'Heterogeneity and genetic variations of serotypes O and Asia 1 foot-and-mouth disease viruses isolated in Vietnam', *Veterinary Microbiology* **145**, 220–229.

Lunn, D., Thomas, A., Best, N. and Spiegelhalter, D. (2000), 'WinBUGS – a bayesian modelling framework: concepts, structure, and extensibility', *Statistics and Computing* **10**, 325–337.

Madin, B. (2011), 'An evaluation of Foot-and-Mouth Disease outbreak reporting in mainland South-East Asia from 2000 to 2010', *Preventive Veterinary Medicine* **102**, 230–241.

Mantel, N. (1967), 'The detection of disease clustering and a generalized regression approach', *Cancer research* **27**(2 Part 1), 209–220.

Marcon, E., Puech, F. et al. (2003), Generalizing Ripleys K function to inhomogeneous populations, Technical report, mimeo.

Nagarwalla, N. (1996), 'A scan statistic with a variable window', *Statistics in medicine* **15**, 845–850.

Naus, J. (1965), 'The dstribution of the size of the maximum cluster of points on a line', *Journal of the American Statistical Statistical Association* **60**(310), 532–538.

Nishiura, H. and Omori, R. (2010), 'An epidemiological analysis of the foot-and-mouth disease epidemic in Miyazaki, Japan, 2010', *Transboundary and Emerging Diseases* **57**, 396–403.

Ntzoufras, I. (2009), *Bayesian modelling using WinBUGs*, John Wiley and Sons, Inc, Hoboken, New Jersey.

Ohser, J. (1983), 'On estimators for the reduced second moment measure of point processes', *Statistics: A Journal of Theoretical and Applied Statistics* **14**(1), 63–71.

Perkins, N., Webster, W., Wright, T., Denney, I. and Links, I. (2011), 'Vaccination program in the response to the 2007 equine influenza outbreak in Australia', *Australian Veterinary Journal* **89**.

R Development Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org/*

Richards, K., Hazelton, M., Stevenson, M., Lockhart, C., Pinto, J. and Nguyen, L. (2014), 'Using exceedance probabilities to detect anomalies in routinely recorded animal health data, with particular reference to foot-and-mouth disease in Viet Nam', *Spatial and Spatio-temporal Epidemiology* **11**, 125–133.

Rinaldi, L., Musella, V., Biggeri, A. and Cringoli, G. (2006), 'New insights into the application of geographical information systems and remote sensing in veterinary parasitology', *Geospatial Health 1* .

Ripley, B. (1976), 'The second-order analysis of stationary point processes', *Journal of applied probability* **13**(2), 255–266.

Ripley, B. (1977), 'Modelling spatial patterns', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 172–212.

Rstudio (2012), 'Rstudio: Integrated development environment for R', http://www.rstudio.org/. Retrieved 12 April 2015.

Salman, M. e. (2003), *Animal disease surveillance and survey systems: methods and applications*, Iowa State Press.

Sanson, R. (1993), The development of a decision support system for an animal disease emergency, PhD thesis, Massey University, New Zealand.

Sanson, R., Stevenson, M. and Moles-Benfell, N. (2006), Quantifying local spread probabilities for foot-and-mouth disease, International Symposium on Veterinary Epidemiology and Economics.

Scott, D. (1992), *Multivariate Density Estimation. Theory, Practice and Visualization*, New York.

Seaman, D. and Powell, R. (1996), 'An evaluation of the accuracy of kernel density estimators for home range analysis', *Ecology* **77**(7), 2075–2085.

Sharma, P. and Baldock, C. e. (1999), *Animal Health in Southeast Asia. Advances in the Collection, management and use of animal health information*, ACIAR Monograph.

Shimshony, A. (1988), 'Foot and mouth disease in the mountain gazelle in Israel', *Rev. sci. tech. Off. int. Epiz* **7**(4).

Silverman, B. (1986), *Density estimation for statistics and data analysis*, CRC Press.

Sobrino, F. and Domingo, E. (2001), 'Foot-and-mouth disease in Europe', *European Molecular Biology Organization* **2**(6).

Spiegelhalter, D., Best, N., Carlin, B. and van der Linde, A. (2002*a*), 'Bayesian measures of model complexity and fit', *Journal of the Royal Statistical society. Series B (statistical Methodology)* **64**(4), 583–639.

Spiegelhalter, D., Best, N., Carlin, B. and Van Der Linde, A. (2002*b*), 'Bayesian measures of model complexity and fit', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(4), 583–639.

Spiegelhalter, D., Best, N., Carlin, B. and van der Linde, A. (2014), 'The deviance information criterion: 12 years on', *Journal of the Royal Statistical society. Series B (statistical Methodology)* .

Stevenson, M., Morris, R., Lawson, A., Wilesmith, J., Ryan, J. and Jackson, R. (2005), 'Area-level risks for BSE in British cattle before and after the July 1988 meat and bone meal feed ban', *Preventive Veterinary Medicine* **69**, 129–144.

Stevenson, M., Sanson, R., Stern, M., O'Leary, B., Sujau, M., Moles-Benfell, N. and Morris, R. (2013), 'Interspread plus: a spatial and shochastic simulation model of disease in animal populations', *Preventive Veterinary Medicine* **109**, 10–24.

Sugiura, K., Ogura, H., Ito, K., Ishikawa, K., Hoshino, K. and Sakamoto, K. (2001), 'Eradication of foot and mouth disease in Japan', *Rev. sci. tech. Off. int. Epiz.* **20**(3), 701–713.

Sumption, K., Rweyemamu, M. and Wint, W. (2008), 'Incidence and distribution of foot-and-mouth disease in Asia, Africa and South America; combining expert opinion, official diseaseinformation and livestock populations to assist risk assessment.', *Transboundary and emerging diseases* **55**.

Tango, T. (2010), *Statistical methods for disease clustering*, Springer.

Timoney, P. (1996), 'Equine influenza', *Comp. Immun. Microbiol. Infect. Dis.* **19**(3).

Unkel, S., Farrington, C., Garthwaite, P., Robertson, C. and Andrews, N. (2012), 'Statistical methods for the prospective detection of infectious disease outbreaks: a review', *Journal of the royal statistical society* .

USGS Earth Resources Observation and Science Center (2011), *MODIS Reprojection Tool*. Release 4.1.
URL: *https://lpdaac.usgs.gov/tools/modis reprojection tool*

Wakefield, J., Best, N. and Waller, L. (2000), *Spatial epidemiology methods and applications*, chapter Bayesian approaches to disease mapping, pp. 104–127.

Waller, L., Carlin, B., Xia, H. and Gelfand, A. (1997), 'Hierarchical spatio-temporal mapping of disease rates', *Journal of the American Statistical Association* **92**(438), 607–617.

Wand, M. and Jones, M. (1995), 'Kernel smoothing'.

Ward, M. and Carpenter, T. (2000), 'Techniques for analysis of disease clustering in space and in time in veterinary epidemiology', *Preventive Veterinary epidemiology* **45**, 257–284.

Wikimedia.org (2015), 'Japan Regions and Prefectures', https://commons.wikimedia.org/wiki/File:Japan-Regions-and-Prefectures.png.

Woodward, A., Rash, A., Medcalf, E., Bryant, N. and Elton, D. (2015), 'Using epidemics to map H3 equine influenza virus determinants.', *Virology* **481**.

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: Kate Richards

Name/Title of Principal Supervisor: Professor Martin Hazelton

Name of Published Research Output and full reference:

Richards, K.K., Hazelton, M.L., Stevenson, M.A., Lockhart, C.Y., Pinto, J. and Nguyen, L. (2014). Using exceedance probabilities to detect anomalies in routinely recorded animal health data, with particular reference to foot-and-mouth disease in Viet Nam. Spatial and Spatio-temporal Epidemiology 11, 125-133.

In which Chapter is the Published Work: Chapter 4

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate: 70%

  and / or
- Describe the contribution that the candidate has made to the Published Work:

Kate Richards Digitally signed by Kate Richards Date: 2015.06.12 13:20:22 +12'00'

Candidate's Signature

12/06/2015

Date

Martin Hazelton Digitally signed by Martin Hazelton DN: cn=Martin Hazelton, o=Massey University, ou, email=m.hazelton@massey.ac.nz, c=NZ Date: 2015.06.12 13:26:52 +12'00'

Principal Supervisor's signature

12/06/2015

Date