SPATIAL AND NETWORK ASPECTS OF THE SPREAD OF INFECTIOUS DISEASES IN LIVESTOCK POPULATIONS

MIKHAIL CHURAKOV

SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
Doctor of Philosophy

INSTITUTE OF BIODIVERSITY, ANIMAL HEALTH AND COMPARATIVE MEDICINE

COLLEGE OF MEDICAL, VETERINARY AND LIFE SCIENCES
UNIVERSITY OF GLASGOW

SEPTEMBER 2014

Abstract

In this thesis, I focus on methodological concepts of studying infectious disease transmission between agricultural premises. I used different disease systems as exemplars for spatial and network methods to investigate transmission patterns.

Infectious diseases cause tangible economic threat to the farming industry worldwide by damaging livestock populations, reducing farm productivity and causing trade restriction. This implies the importance of veterinary epidemiological studies in control and eradication of pathogens.

Recent increase in availability of data and computational power allowed for more opportunities to study mechanisms of pathogenic transmission. Nowadays, the bottleneck is primarily associated with efficient methods that can analyse vast amounts of high-resolution data.

Here I address two livestock pathogens that differ in their epidemiology: bacteria *Streptococcus agalactiae* and foot-and-mouth disease (FMD) virus.

Streptococcus agalactiae is a contagious pathogen that causes mastitis in cattle, and thus possesses a substantial economic burden to the dairy industry. Known transmission routes between cattle are restricted to those via milking machines, milkers' hands and fomites during milking process. Additionally, recent studies suggested potential introductions from other host species: primarily, humans. However, strain typing data showed discrepancies in strain compositions of bacteria isolated from humans and bovines. In this thesis, strain-specific features of between-herd transmission of *Streptococcus agalactiae* within dairy cattle population in Denmark are investigated.

Foot-and-mouth disease (FMD) is a viral infection that affects cloven-hoofed animals and is of big importance mainly because of the trade restrictions against infected regions and countries. Control programmes against FMD usually include vaccination and culling of animals. However, the debate on the optimal control for FMD is still ongoing. In this thesis, I address questions on identification of the routes of infection and on requirements for movement recording systems to be used for efficient contact tracing during an FMD outbreak.

The thesis consists of seven chapters:

1. General introduction.

   Overview of methods for analysis of spatial and network contact structures that may affect infectious disease transmission between farms; and description of two exemplars of disease systems used for analyses.

2. Descriptive epidemiology of *Streptococcus agalactiae* in the population of Danish dairy cattle herds in 2009–2011.

   Description of available data, construction of the principal dataset for strain-specific analysis in chapters 3 and 4.

3. Spatial clustering of Danish dairy cattle herds infected with various strains of *Streptococcus agalactiae*.

   Spatial analysis of the distributions of *Streptococcus agalactiae* strains.

4. Contact patterns between Danish dairy cattle herds and their role in the spread of various strains of *Streptococcus agalactiae*.

   Network analysis of movement and veterinary contact networks for specific strains of *Streptococcus agalactiae*.

5. Estimation of possible transmission trees: application to the Darlington cluster within the 2001 FMD epidemic in the UK.

   Extension of the previous approach to reconstruct transmission trees.

6. The effect of rapid contact tracing using movement recording systems for controlling disease outbreaks.

   Assessing the effect of delays in movement tracing for post-silent spread of FMD-like pathogens.

7. General discussion.

   Overall conclusions and future perspectives.

This thesis reveals several interesting findings. Firstly, the increased understanding of strain-specific transmission characteristics of *Streptococcus agalactiae*. One of the observed strains (ST103) showed significant and consistent spatial clustering of its cases among Danish dairy cattle herds in 2009–2011.

Secondly, the network analysis of cattle movements and affiliations with veterinary practices showed that veterinary practices were exclusively associated with transmission of ST103 of

Streptococcus agalactiae. Contrastingly, movement networks appeared to be important for all the three predominant bacterial strains (ST1, ST23 and ST103).

Fourthly, the new extended approach that allows estimation of the whole transmission tree at once was proposed and tested for the Darlington cluster within the 2001 FMD UK epidemic.

Finally, in chapter 6, it was shown that mathematical modelling did not suggest any advantages of ensuring smaller delays in the post-silent control of FMD-like pathogens.

Acknowledgements

This thesis is a result of four years of work at the Institute of Biodiversity, Animal Health and Comparative Medicine, University of Glasgow. The Ph.D. project was funded by the University of Glasgow through the Lord Kelvin & Adam Smith scholarship scheme.

I am sincerely grateful to my supervisors Prof. Rowland Kao, Prof. Ruth Zadoks and Dr. Simon Rogers for their helpful advice, careful guidance and solid support. I can say that they helped me not only to grow as a scientist but also as a person.

I also wish to extend my gratitude to Jørgen Katholm and Jørgen Nielsen from the Knowledge Centre for Agriculture (Aarhus, Denmark) for providing the data for this project.

I am deeply indebted to my family for their constant support despite the distance between us.

I would like to say a big thank you to all the people who surrounded me during these four amazing years in Glasgow: collleagues in the Institute and people who became my true friends. Joaquin, Noushin, Mark, Colette, Sema, Laetitia, Jess, Cat, Anthony, Hattie, David, Julie, Yi-Hsiu, Daniel, you have supported me in so many ways, and made Glasgow feel like home.

In memory of my grandfathers, Leonard S. Borin and Vladimir V. Churakov.

Table of Contents

List of Tables

List of Figures

Chapter 1

General introduction

1.1 Infectious diseases in livestock

Infectious diseases are caused by pathogenic microorganisms (such as viruses, bacteria, parasites or fungi) and can be spread, directly or indirectly, from one host to another. Which implies their importance for human and animal populations.

Infectious diseases of livestock are of interest because of the economic implications. Presence of infection can affect farming industry directly (through decreased productivity) and indirectly (through trade restrictions).

In addition, some infectious diseases of livestock may also affect human health. According to the recent FAO report [FAO, 2013], around 70% of the new diseases that have emerged in humans over recent decades are of animal origin (wildlife and livestock).

Epidemiology is the science that studies the patterns, causes, effects, and control of diseases (or other health-related issues) in populations. The ultimate goal of any epidemiological research is to determine effective and efficient (regarding time, labour and economic resources) control programmes. Targeted control should rely on confirmed transmission mechanisms (or patterns) of particular infectious disease.

Clinical data are widely available in both veterinary and human epidemiology. However, the information on human contact patterns is not easily accessible or just not recorded, while livestock surveillance and movement recording systems (especially in Western Europe) provide more data on a fine-scale.

Advances in computer hardware also allowed for dealing with bigger datasets. And now the bottleneck is in efficient quantitative methods to deal with this amount of information.

In this chapter, I will introduce several approaches to analyse and model the spread infectious diseases, all of which have undergone major developments in the past decades. In the thesis, these approaches are applied to an endemic, production limiting disease (bovine mastitis) and a highly infectious, trade limiting disease (foot-and-mouth disease). Salient features of the two exemplar diseases will be briefly described towards the end of this chapter.

1.2 Computational methods in epidemiology

Computational methods in veterinary epidemiology include statistical analysis and mathematical modelling of infectious diseases of livestock and wildlife animals. Each of these methods is focused on one or several components of the collected data: spatial, temporal, network[1] and genetic.

Here, a brief overview of methods focusing on different components is given. Later in the thesis, most attention will be given to spatial and network aspects, and individual methods will be discussed in detail in the corresponding chapters.

1.2.1 Spatiotemporal component

One of the crucial steps in understanding the epidemiology of a considered pathogen is the analysis of spatial distribution of its cases. Presence of consistent spatial patterns is usually a reflection of specific transmission mechanisms. There are number of statistical methods to test for spatial clustering (non-random gathering in one location) of infected individuals [Cuzick and Edwards, 1990, Kulldorff, 1997]. Evidence of spatial clustering suggests the existence of either a contagious process or local environmental factors [Carpenter, 2001].

If a study period is long enough for changes over time in geographical distribution of cases, temporal component needs to be taken into account in order to analyse the dynamics of infectious diseases. Spatiotemporal clustering was used in a number of studies [Fenton et al., 2009, Ahmed et al., 2010].

1.2.2 Network component

Contact patterns between individuals are another factor that can determine the disease spread. And only rarely do those contacts that are important for disease transmission appear to be homogeneous, which complicates the analysis [Kiss et al., 2006, Kao et al., 2006, Böhm et al., 2009]. Therefore, these contacts are summarised to form networks: a set of edges (contacts) between nodes (individuals). And then, they are analysed using network analysis: methods to assess properties of these networks.

[1] In this thesis, term “network data” is used for any kind of contact structures between individuals that can be represented by a network, e.g. network of animal movements between premises.

In veterinary epidemiology, the most frequently used networks are based on animal movements. However, other types of contacts (e.g. vector) can also be summarised in the same way.

The networks of cattle, sheep and pig movements were intensively analysed using the network analysis [Kiss et al., 2006,Kao et al., 2006,Kao et al., 2007,Dubé et al., 2009,Nöremark et al., 2011].

1.2.3 Genetic component

Genetic data can provide another perspective on the epidemiology of a pathogen.

For slow-mutating organisms, genetic data may be a good marker of independent branches of transmission [García-Álvarez et al., 2011], whereas for fast-mutating organisms, it can be used to infer evolutionary history of pathogens and their transmissions between hosts (phylogenetic analysis) [Jamal et al., 2011,Malirat et al., 2011].

Genetic data hold tremendous amount of information regarding the epidemiology, virulence factors and mechanisms of pathogenesis and antibiotic resistance. The appropriate methods that deal with whole genome sequencing data are beyond the scope of this thesis. Here, low-resolution genetic data (multilocus sequence typing data) were used to make distinction between particular strains of *Streptococcus agalactiae*.

1.3 *Streptococcus agalactiae*: exemplar of endemic bacterial disease

Streptococcus agalactiae (or group B streptococcus) is one of the bacteria that can be the cause of mastitis (inflammation of the udder) in dairy cattle and lead to decreased productivity and milk quality [Keefe, 1997].

S. agalactiae is usually transmitted between cattle via milking machines or milker's hands [Keefe, 1997]. It can also be introduced from other species: the pathogen can be carried in particular by humans [Andersen et al., 2003], who are the main host of the organism, mostly as asymptomatic carriers. *S. agalactiae* is part of the normal bacterial flora colonising the gastrointestinal and genitourinary tracts of a significant proportion of the human population, but can also cause bacteremia, septicaemia, meningitis, and pneumonia [Glaser et al., 2002].

1.3.1 Bovine mastitis

In cattle, *Streptococcus agalactiae* usually colonises the udder resulting in subclinical mastitis [Keefe, 1997]. However, it was also reported to be recoverable from bovine faeces [Manning et al., 2010].

Mastitis, also known as intramammary infection (IMI), is one of the most frequent and economically important infectious diseases affecting dairy cattle, especially in developed countries [Halasa et al., 2007]. Mastitis can be caused by a number of pathogens, most frequently by *Escherichia coli*, *Staphylococcus aureus*, *Streptococcus uberis* and *Streptococcus agalactiae* [Zadoks et al., 2011].

Milk from cows suffering from mastitis has an elevated somatic cell count (SCC) and is of lower quality. However, subclinical mastitis does not create visible changes in the milk or the udder, but infected cows will produce less milk. In addition, infected cows can be a source of infection to other animals in the herd. Therefore, reliable detection of subclinical mastitis is essential in the control of the disease [Zadoks and Fitzpatrick, 2009].

Dairy associations normally advise farmers to adhere to the five point plan (initially described in [Neave et al., 1969]) in order to control bovine mastitis on their premises:

1. Record and treat all clinical cases.
2. Post-milking disinfection of teats.
3. Use dry cow therapy at drying off.
4. Cull cows with chronic mastitis.
5. Perform regular milking machine maintenance.

Amongst other causative agents of bovine mastitis, *S. agalactiae* is deemed particularly important because of its contagiousness, therefore there are mandatory control programs in some countries, e.g. in Denmark [Katholm, 2010, Mweu, 2013].

1.3.2 Control of *Streptococcus agalactiae* in the population of Danish dairy cattle herds

The cornerstone of most *S. agalactiae* control programs is that within-herd transmission is mostly due to indirect animal contacts via the milking machine whilst between-herd transmission is mostly due to movement of animals. There are possible exceptions to those rules, because other sources of *S. agalactiae* are known, notably humans [Dogan et al., 2005] and, potentially, bovine faeces [Manning et al., 2010].

Control of *Streptococcus agalactiae* on the national level in Denmark initiated in 1950s, because in 1950, 30–40% of Danish herds were infected with *S. agalactiae* [Andersen et al., 2003]. A nationwide surveillance program was based on bacteriological examination of bulk tank milk (BTM), and was accompanied with eradication program (identifying infected cows by quarter milk samples and subsequent treating or culling). The eradication program was compulsory until 1988, but then became voluntary, but still with prohibition to sell cows and pregnant heifers from herds declared to be infected. The sampling was carried out with changing intervals, but from 1995 the BTM samples were examined annually.

In 2005, the B-register (publicly available database of herds positively identified with *Streptococcus agalactiae*) was created, and owners of infected herds became obligated to disclose their status to willing buyers [Mweu, 2013]. The legislation required that every herd had to be put into the B-register if it had positive BTM sample confirmed in at least one of two subsequent tests or a submitted milk sample from one of the cows in the herd tested positive. A herd could be exclude from the B-register if all cow-level tests returned negative within the same day of testing or if four of its consecutively tested BTM samples (30 days apart) were negative.

The ban on the sale of heifers and cows was lifted in 2005 as it was argued that the risk of transmission by purchase was negligible (given the publicly available B-register). However, participation in shows or gatherings where milking was likely remained forbidden.

Bacteriological culture was used in Denmark as a standard method in the national annual surveillance for *Streptococcus agalactiae* before 2009. The PCR test was then used in the annual surveillance. The final switch from bacteriological culture to PathoProof Mastitis PCR assay [Koskinen et al., 2009] as the conventional screening tool occurred on 1 September 2010.

The annual national surveillance is performed on herd level in September–December when BTM samples are usually collected. Normally, all samples from January to September will be those not included in the yearly testing and are typically extra BTM samples from earlier positive herds (or in some cases for individual cows from the same herd, to identify and exclude infected animals for treatment or culling).

Despite the high-standard surveillance scheme and efforts in improving within-herd biosecurity, the current control of *S. agalactiae* is not perfect and is incapable to eradicate the pathogen from Danish dairy cattle population [Katholm, 2010].

1.3.3 Strain typing of *Streptococcus agalactiae*

Strain typing is used for bacterial pathogens to characterise their population composition.

Strain types of *Streptococcus agalactiae* were previously characterised by a number of typing methods: serotyping [Dogan et al., 2005], multilocus sequence typing (MLST) [Manning et al., 2010, Zadoks et al., 2011, Yang et al., 2013], randomly amplified polymorphic DNA (RAPD) [Martinez et al., 2000], ribotyping [Rivas et al., 1997] and pulse-field gel electrophoresis (PFGE) [Pereira et al., 2010].

In this thesis, MLST was used to characterise isolates from Danish dairy cattle herds. The choice of typing method was determined by the fact that MLST has sufficient discriminatory power to differentiate bacterial strains associated with various host species and also has been used intensively in previous studies [Zadoks et al., 2011, Yang et al., 2013].

MLST targets the DNA sequence variations in a set of genes (usually, housekeeping conservative genes that evolve slowly) and characterises strains by their unique allelic profiles. The technique involves PCR amplification of fragments of interest, followed by DNA sequencing.

The strain is defined as a sequence type (ST) — a number assigned to allelic profile of several allele numbers corresponding to the targeted genes (conventional number of housekeeping genes for *S. agalactiae* is seven).

The information of the observed STs is usually submitted to the public MLST database [Jolley and Maiden, 2010] (available at `http://pubmlst.org/sagalactiae/`) where new (previously not observed) STs get assigned numbers.

The population structure normally is not uniform with a limited number of the most prevalent STs (e.g. for human isolates, ST1, ST17, ST19 and ST23 are the most frequent [Bisharat et al., 2004]). Most of them are also recoverable from bovine populations [Bisharat et al., 2004, Bohnsack et al., 2004, Oliveira et al., 2006].

Strictly speaking, strain is a broader concept than ST, e.g. in studies where the observed bacterial population is relatively diverse, closely related STs (those that differ in 1–2 allele numbers) can be combined into clusters that are conventionally referred as clonal complexes [Feil et al., 2004].

In this thesis, unless stated otherwise, the epidemiological definition of host associated strain characterisation will be used: bovine and human strains[2] are identified based on the balance of primary host association (e.g. ST23 is a human strain, but ST103 is a bovine one as it was rarely recovered from humans previously).

Since there are different strains of *S. agalactiae*, associated with several host species, recent studies [Andersen et al., 2003, Zadoks et al., 2011] suggested the potential role of human strains in the re-emergence of the pathogen in bovine populations.

[2] Fish strains form a distinctive cluster and do not overlap with human and bovine strains.

1.3.4 Objectives and research questions

Unique combination of intensive surveillance program and detailed data on individual cow histories provided a detailed dataset for spatial and network modelling, and analysis of strain-specific transmission patterns.

Transmission of *S. agalactiae* between herds has not been studied on strain level before. In this thesis, I address the following question: are transmission pattern of *S. agalactiae* strain-specific? If transmission properties of different strains are distinct from each other, then novel targeted approaches should be put into place to eradicate the pathogen.

In chapter 2, available data and data processing will be described. Next, the overall question is broken down into two analyses: spatial analysis and contact structure analysis, which are presented in chapters 3 and 4, respectively.

1.4 Foot-and-mouth disease: exemplar of highly infectious viral disease

Foot-and-mouth disease (FMD) is a viral infection that affects cloven-hoofed animals and is of big importance because of the trade restrictions against infected regions and countries [Bachrach, 1968, Haydon et al., 2004].

The causative agent of FMD is a single stranded RNA virus [Domingo et al., 2002] and is subject to rapid evolution due to its high mutation rate [Domingo et al., 2003], which makes it possible to perform phylogenetic analysis on genetic samples collected during a short period of time.

FMD virus can be transmitted between farm animals via direct contact, fomites (e.g. contaminated vehicles) or aerosol [Bachrach, 1968]. Cattle are the most susceptible to FMD, while domestic pigs are very effective in propagating the disease. In sheep and goats, the clinical manifestations of infection are usually less severe than in cattle and pigs, but they can also play a significant role in dissemination of the pathogen [Gibbens et al., 2001].

FMD has low associated mortality but can decrease livestock productivity and therefore represents a serious threat to farming industry [Knight-Jones and Rushton, 2013]. Countries with circulating FMD in their livestock populations are subject to considerable international trade restrictions on animals and animal products [James and Rushton, 2002].

FMD is distributed worldwide [Rweyemamu et al., 2008] most notably in South America, Africa and Asia. In most countries FMD is endemic [Thalmann and Nöckler, 2001], but sometimes it can go out of control and affect large proportions of livestock population, like in the UK in 2001 [Gibbens et al., 2001].

Historically, FMD control programs in different countries included isolation, slaughter, destruction of virulent material and, later, vaccination [Blancou, 2002]. In Europe, preventive vaccination was banned in the early 1990s [Leforban and Gerbier, 2002] and has never been a part of the FMD control program in the UK [Haydon et al., 2004].

In the UK, traditional FMD control policies include slaughter of all infected and in-contact susceptible animals, which in 2001 (during the major epidemic, discussed in the next section) were accompanied by strict restrictions on movement of animals and vehicles around infected premises [Gibbens et al., 2001]. After slaughter, the carcasses were destroyed and the buildings were thoroughly disinfected [Scudamore et al., 2002]. Contact tracing was also performed to identify the source of the outbreak and premises that might have been affected.

1.4.1 The 2001 foot-and-mouth disease epidemic in the UK

In February 2001, the first cases of FMD outbreak were confirmed in an abattoir in Essex. The epidemic lasted for seven months with 50 newly infected premises per day at its peak in late March [Gibbens et al., 2001]. National movement ban shortly after the start of the epidemic was followed by sequentially applied culling policies that seriously affected populations of cloven-hoofed animals (including sheep, cattle, pigs and others): more than 6.5 million animals were culled on 9900 premises (but only 2030 were infected) [Haydon et al., 2004, Kao, 2002], with an estimated cost of £3–5 billion to the national economy. It was reported that rapid and complete culling of dangerous contact (DC) farms was necessary to control the epidemic [Ferguson et al., 2001b], however further studies did not fully support this statement [Kao, 2003, Matthews et al., 2003].

Large and detailed dataset collected during the 2001 FMD epidemic in the UK inspired creation of numerous epidemiological models based on ordinary differential equations [Ferguson et al., 2001a, Ferguson et al., 2001b] and stochastic simulations [Keeling et al., 2001]. Haydon et al. constructed epidemic trees to estimate the case-reproduction ratio and the possible efficacy of alternative control measures [Haydon et al., 2003]. Savill et al. showed that Euclidean distance between infectious and susceptible premises is a better predictor of transmission risk than shortest and quickest routes via road using a spatial model [Savill et al., 2006]. Later, Tildesley et al. made a comparison between model and data at the individual farm level, assessing the potential of the model to predict the infectious status of farms in both the short and long terms [Tildesley et al., 2008].

1.4.2 Objectives and research questions

In this thesis, I address questions on estimation of possible routes of infection (in application to FMD) and on efficiency of contact tracing via movements during a potential FMD outbreak. Although these questions can be formulated for other disease systems, FMDV was chosen as an exemplar of highly infectious pathogen because FMD is still a relevant livestock disease to most countries and a lot of work has been done in parameter fitting for various mathematical models of FMD transmission.

Therefore, the objectives are:

1. To investigate existing methods of transmission tree reconstruction and how they can be extended.
2. To identify the impact of faster contact tracing due to real-time movement recording systems on control of a potential FMD outbreak.

1.5 Thesis outline

In this chapter, I briefly discussed methods in veterinary epidemiology that deal with spatial and network data. Also, description of two exemplars of disease systems that will be used in the next chapters was given.

In chapter 2, I will discuss the available data and construction of the principal dataset for strain-specific analysis of *S. agalactiae*, which will be performed in chapters 3 and 4, regarding the spatial distributions and transmission patterns associated with contact networks, respectively.

In chapter 5, I will review the existing approaches of transmission tree estimation and propose an alternative method to extend one of the previously published approaches.

In chapter 6, I will assess the effect of delays in movement tracing on the final epidemic size using mathematical modelling.

And in chapter 7, I will summarise and discuss the future perspectives in spatial and network methodology in relation to veterinary epidemiology.

Chapter 2

Descriptive epidemiology of *Streptococcus agalactiae* in the population of Danish dairy cattle herds in 2009–2011

2.1 Introduction

The first stage of any epidemiological study is the descriptive analysis of collected data and seeks to summarise observations on various characteristics of infection cases (time, location, disease conditions, etc.). This is the essential part that leads to formulation of hypotheses about the drivers of the disease spread.

In this chapter I describe how and what data were collected for this study, discuss the characteristics of these data, and explain how the principal dataset that will be used for strain-specific analysis in chapters 3 and 4 was defined.

The study period for this analysis is 2009–2011, when isolates of *Streptococcus agalactiae* from bulk tank milk (BTM) samples were analysed using multilocus sequence typing (MLST), to identify distinct strains of the bacteria.

The motivation for this study was inspired by previous characterisation (by MLST types) of the population structure of *S. agalactiae* in Danish dairy cattle [Zadoks et al., 2011]. In this thesis I will focus on strain-specific features of the pathogen.

Several characteristics of these data made it unsuitable for direct analysis:

- It contained both active and passive surveillance results (i.e. those collected during the official national annual sampling of dairy herds and those performed outwith these time periods, respectively).

- Two different methods, bacteriological culture and polymerase chain reaction (PCR), were used in 2009–2011, but not at the same time during the whole study period. This is also problematic because the two tests have different sensitivity and specificity characteristics.

The following questions were highlighted as being important for the construction of the principal dataset:

1. What herds should be considered as positive?

   Bacteriological culture and PCR tests were used during the study period to detect the presence of *Streptococcus agalactiae* in BTM samples. Both tests are not perfect, and sometimes it can be difficult to say if herd was infected or not (e.g. when PCR and culture give different results).

   Also, for some herds there were test results on samples collected outwith the annual surveillance. And it is not clear how to account for this information, especially if these results are not in agreement with the official annual testings.

2. How do we assign strain types to a herd?

   Samples that defined herd's infection status were collected during the official annual surveillance, but samples for MLST could be collected on different dates. Assignment of strain types should be made with caution, especially if MLST results were based on samples collected between annual surveillance periods and can potentially be assigned to any of the two years.

3. How should we account for potential carry-over during sample collection?

   The BTM sample collection procedure was imperfect, which might occasionally have led to cross-contamination of collected samples and, hence, false identification of positive herds. Exclusion of false positives (FP) from the study dataset will be discussed further in this chapter.

2.2 Herd-level data

The surveillance herd-level data were extracted from the Danish Cattle Database by Jørgen Nielsen (Knowledge Centre for Agriculture, Aarhus, Denmark) in January 2013.

The data include information on herds (geographic location, herd size, types of animals, milking systems, etc.) and history of BTM tests, on which I'm going to focus further.

Jørgen Katholm (Knowledge Centre for Agriculture, Aarhus, Denmark) was responsible for coordination and implementation of the annual surveillance program during the study period (and for several years before and after) and he has provided additional information that isn't available in published resources.

Milk samples data were affected by carry-over during BTM sample collection (discussed later in this chapter). This issue had been addressed before this study, but here we had additional data (strain typing) that was used to resolve cases of potential (but previously not confirmed) false positive results. Therefore, the first round of carry-over corrections (that included recovery of milk pick-up routes) was not performed by the author of this thesis.

2.2.1 Bulk tank milk sample collection

BTM samples are routinely collected in September–December each year under the national surveillance program. During the collection of milk from the bulk tank, the first 30 litres of milk was routinely flushed through the milk hose and pipes in order to avoid contamination by milk residues from former sampling. Then, 60 ml of milk was extracted and stored in plastic test tubes that were immediately stored on ice. In the next 24 hours the samples were delivered to Steins Laboratory A/S (Hjaltesvej 8, 7500 Holstebro, Denmark) for examination [Andersen et al., 2003]. Later, the samples were tested for the presence of important pathogens, including *S. agalactiae*.

MLST analysis was performed on isolates from the same samples that were previously confirmed by culture. Molecular data will be discussed later in this chapter, just before the case definition, which is based on both PCR and bacteriological culture.

2.2.2 Switch from bacteriological culture to PCR in the surveillance

Bacteriological culture was the standard (conventional) screening method for *Streptococcus agalactiae* surveillance until the end of 2009.

Recently, the new real-time PCR assay the PathoProof Mastitis PCR (Finnzymes Oy, Espoo, Finland) has become available [Koskinen et al., 2009]. In 2009, it began to be used for BTM screening in Denmark, but became the standard procedure only on 1 September 2010.

The two tests were evaluated both analytically [Koskinen et al., 2009] and using latent class analysis [Mweu et al., 2012]. A cycle threshold[1] (Ct) cut-off value of 40 was proposed to

[1] The cycle threshold in the real-time PCR is defined as the number of cycles required for the fluorescent signal to exceed the background level. Ct levels are inversely proportional to the amount of target nucleic acid in the sample (i.e. the lower the Ct level the greater the amount of target nucleic acid in the sample).

be used for identification of BTM samples with *S. agalactiae*. The PCR is more sensitive, while bacteriological culture is more specific for this cut-off.

2.2.3 Study population

The number of dairy cattle herds decreased throughout the study period (see Table 2.1), which followed the trend of steady decrease (accompanied by the increase in herd size) since 1966 [Mweu, 2013] (decreased by a factor of 2 between 1990 and 2000: from 20,091 to 9,886 herds). Reports on active dairy herds are published each year in May. The list of herds to be visited during the annual surveillance is based on the list of milk producers on the last quarter of each year. Some herds can avoid annual sampling, mainly due to being absent from the testing list if they were just established, but each year the number of not tested herds is below 10 (Jørgen Katholm, personal communication).

Table 2.1: The number of dairy cattle herds tested during the annual surveillance in Denmark in 2009–2011.

Year	Tested herds	Newly registered herds	Herds excluded from the register
2009	4258	-	-
2010	4091	26	193
2011	3918	22	195

2.2.4 Timing of annual testings

National bulk tank surveillance in Denmark usually takes place in the last quarter of each year (September–December). Temporal distribution of BTM samples taken for the annual surveillance is presented in Figure 2.1.

2.2.5 Delays in collecting samples for bacteriological culture and strain typing

In 2009 and 2010, the same BTM samples were used for both tests (PCR and culture). On 1 September 2010, after PCR became the conventional method for *Streptococcus agalactiae* screening, not all the active dairy herds were tested using bacteriological culture, and some of the samples were re-collected for culture after the annual surveillance round of sample collection. For 2011, BTM samples for bacteriological culture and MLST analysis were collected in the early 2012 from all herds that were PCR positive at the 2011 annual surveillance.

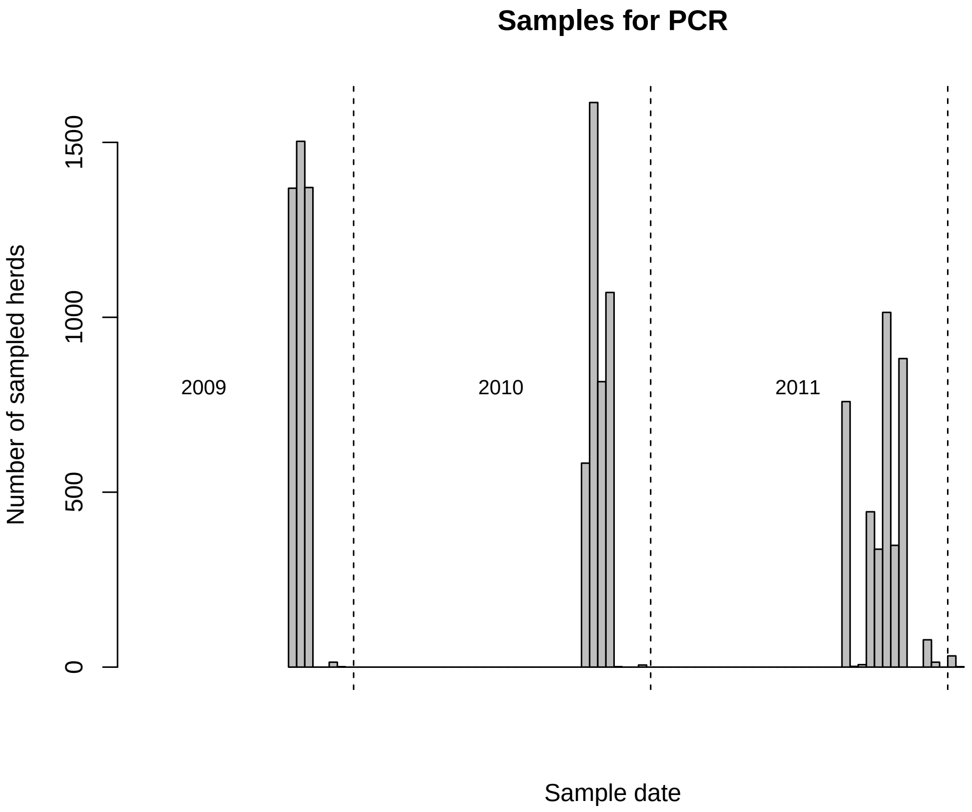

Figure 2.1: Temporal distribution of the number of herds tested during the annual bulk tank milk samples collection during the study period in 2009–2011.

Of course, in the period of around 4 months (duration of the annual samplings), herd status can change (e.g. because of treatment or introduction of a new bacterial strain). However, later in this chapter it will be discussed why this is not a problem, i.e. persistence of the same ST over time within farms (see Figure 2.4).

Distributions of the delays between annual samples (for PCR) and subsequent samples (for culture and MLST) in 2010 and 2011 (effectively, 2012) are presented in Figure 2.2. In 2009 both tests were performed on one BTM sample per herds, hence no delays.

2.2.6 Coverage of the surveillance scheme for two tests

The gradual switch in the standard screening tool from bacteriological culture to PCR affected the coverage of tests: while PCR was performed on all collected samples in 2009–2011, bacteriological culture covered 100% of herds only in 2009 and was used against only PCR positive herds in 2011. This can be observed in Table 2.2 that shows a summary of annual BTM test results for both PCR and culture, in the study period.

Therefore, the shift in standard surveillance test procedures during the study period made it

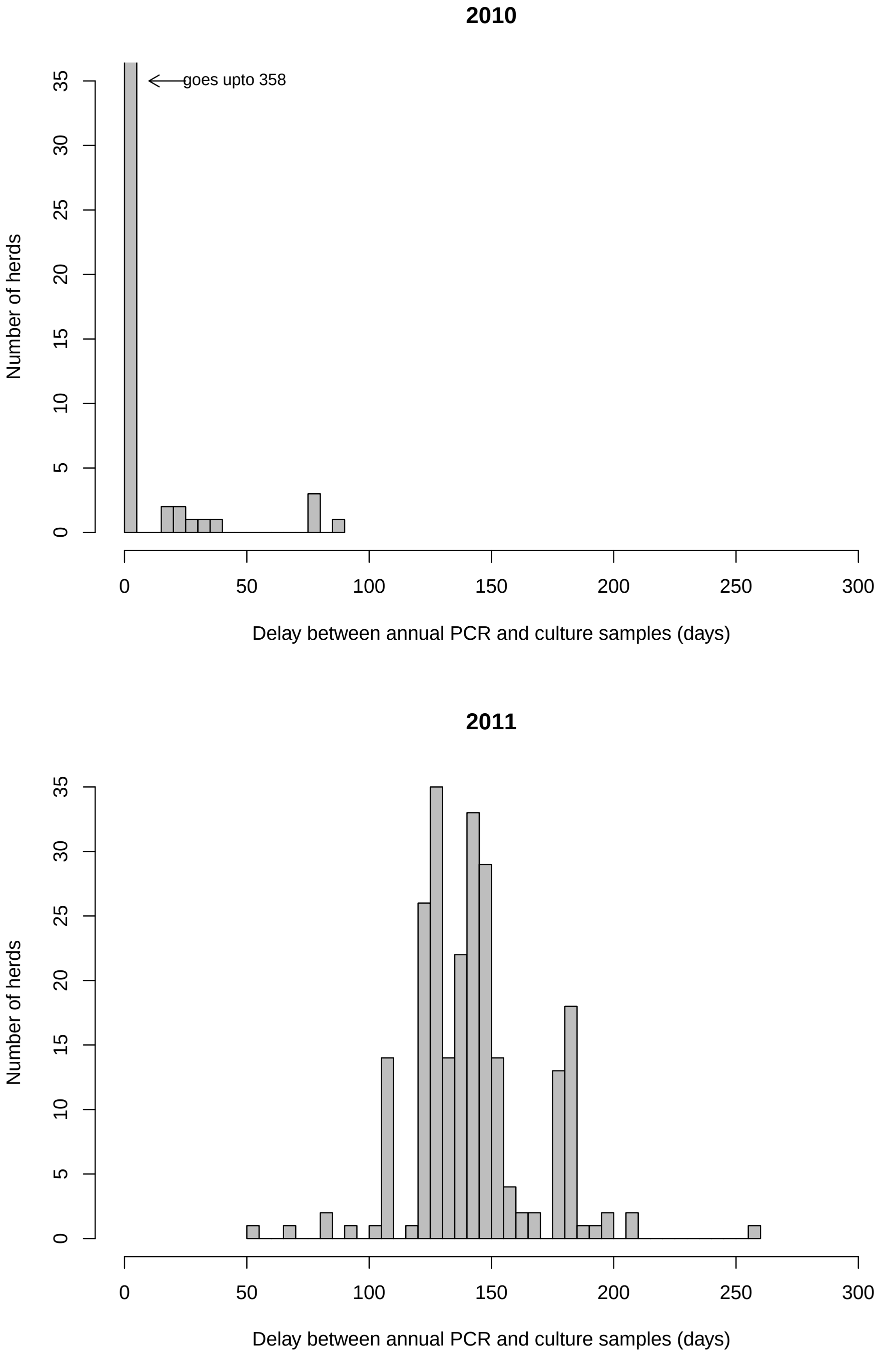

Figure 2.2: Distribution of the delays between annual PCR and culture samples in 2010 and 2011.

impossible to use either PCR or culture to define infected herds, instead, both tests should be taken into account.

Table 2.2: Summary of annual bulk tank milk test results in 2009, 2010 and 2011.

2009

Culture PCR	Positive	Negative	Not tested	Total
Positive	178	132	0	310
Negative	20	3928	0	3948
Total	198	4060	0	4258

2010

Culture PCR	Positive	Negative	Not tested	Total
Positive	131	133	1	265
Negative	11	94	3721	3826
Total	142	227	3722	4091

2011

Culture PCR	Positive	Negative	Not tested	Total
Positive	159	59	3	221
Negative	0	0	3675	3697
Total	165	75	3678	3918

2.2.7 Comparison of bacteriological culture and PCR tests

Sensitivity and specificity of both tests were assessed in [Mweu et al., 2012]. The real-time PCR assay showed higher sensitivity but lower specificity than the culture test. Which, in turn, implied the preference of PCR over culture as a conventional BTM test for detection of *Streptococcus agalactiae*. However, confirmation by bacteriological culture was advisable for herds with high Ct values [Mweu et al., 2012].

Even though PCR is more sensitive and can detect growth-inhibited and nonviable bacteria, both methods can yield negative results for true positive herds. Bacteriological culture is also not 100% specific (e.g. it can occasionally detect other bacterial species). These facts corroborate the idea that both tests should be used in conjunction to provide a more reliable case definition.

2.2.8 Carry-over during sample collection

Cross-contamination of BTM samples can occur during milk sample collection as a result of milk residues in the sampling equipment from a previously sampled infected herd [Andersen et al., 2003]. This may lead to false positive (FP) results when BTM samples are tested at the laboratories (especially using very sensitive tests such as PCR). In order to minimise this kind of errors, a number of herds that fell under suspicion (based on previous test history and subsequent test results) were scrutinised by people from the Danish Knowledge Centre for Agriculture. Milk collection routes were recovered, which allowed to identify the order of sampled herds and, ultimately, establish pairs: a suspected carry-over herd and its source.

Corrections for carry-over

First round of corrections for carry-over had been introduced to the Danish Cattle Database before the data for this project were extracted, thus I will at first discuss how it changed the data, and then I will report on the corrections that were made specifically for this thesis (by myself) given the availability of strain typing data.

A previous report [Mweu, 2013] had different numbers of PCR positive herds in 2009–2011: 310 (out of 4258), 270 (out of 4091) and 243 (out of 3918), respectively. This was before the carry-over issue was addressed by the Danish researchers, but subsequently 5 herds in 2010 and 22 in 2011 were confirmed as false positives and removed from the dataset, thus the drop to 265 in 2010 and to 221 in 2011 (see Table 2.2). However, strain typing data were not used to verify potential carry-over events. But the presence of the same strain in BTM samples from subsequent herds could corroborate the fact of carry-over, similarly two different strains indicate unlikely carry-over.

Confirmation by strain typing data

Here, previously unavailable MLST results were incorporated to confirm several potential FP herds due to carry-over. Note that only pairs of herds that had isolates available were assessed.

The decisions that I made to include/exclude data are summarised in Figure 2.3. We expect PCR Ct values to be higher for false positive herds than for the sources of cross-contamination (less pathogenic content, thus more cycles needed to detect the DNA). Also, cross-contaminated samples should hold the same strain of bacteria (i.e. the same sequence type (ST)). Herds that did not meet any of these criteria were assumed to be true positives (see Table A.1). For the remaining herds, those that were persistently infected with the same ST in other years (herds 8 and 188) were excluded from the list of false positive herds. On the other hand, if the source farm was persistently infected with the same strain (herds 217, 304, 323), the next herd (in the milk truck pick-up route) was treated as a likely carry-over. Herd 303 was also treated as FP, despite its source had ST0 in 2011 (ST could not be identified).

These corrections resulted in the exclusion of 14 herds from the list of positive herds in 2009, one herd in 2010, and four in 2011 (see Table A.1). The four herds (297, 303, 304, 323) that

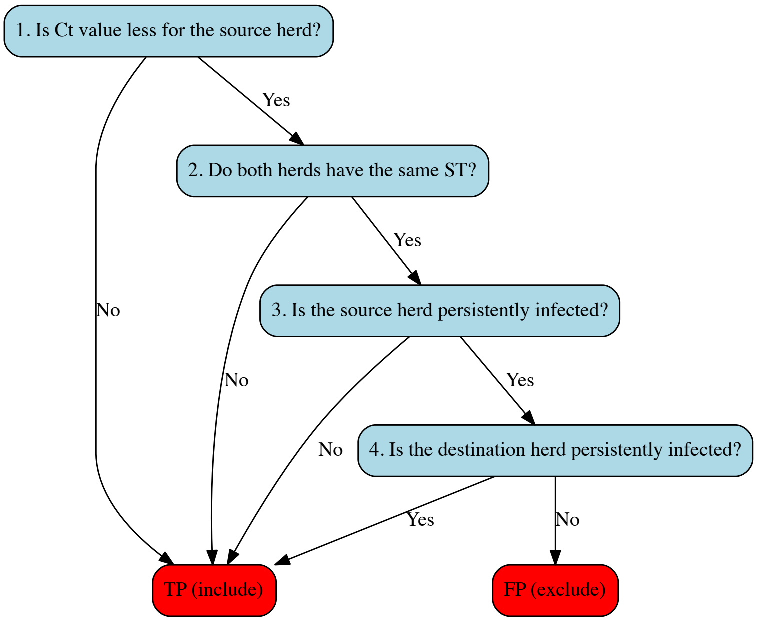
1. Is Ct value less for the source herd?
Yes
No
2. Do both herds have the same ST?
Yes
No
3. Is the source herd persistently infected?
Yes
No
4. Is the destination herd persistently infected?
Yes
No
TP (include)
FP (exclude)

Figure 2.3: Decision flowchart to include/exclude annual surveillance data for potentially false positive herds due to bulk tank milk carry-over. A pair of source–destination herds is assessed based on the PCR and MLST results for particular year.

are now confirmed for 2011 were amongst the previously excluded 22 herds: 6 PCR positive (including the aforementioned four) and 16 PCR negative herds.

2.3 Molecular data

2.3.1 Data generation

At first, collected BTM samples were cultured in Denmark. Then, isolates were sent to the Moredun Research Institute, where conventional MLST was performed [Jones et al., 2003] or material was prepared for high-throughput MLST (HiMLST) at Streeklab Haarlem, The Netherlands [Boers et al., 2012].

There are potential failures at almost every stage in the process of identification of strain type. MLST can be performed only on cultured isolates, hence we can not process growth-inhibited or nonviable bacteria. Also, bacteria can die during transportation or in the lab. Furthermore, even the process of allele sequencing can be unachievable that will restrict us from getting one of the allele numbers and, in turn, the final sequence type (ST) number[2].

[2] As it was discussed previously, sequence type (ST) is defined based on 7 allele numbers that represent sequences of 7 conservative house-keeping genes.

2.3.2 Consistency of sequence types

Potential within-herd variability of bacterial strains implies that several STs can be identified from the same BTM sample; and individual animal milk samples can be of use to test the hypothesis of multiple STs in one herd. However, in this study, one sequence type was assumed to be representative for the whole herd based on the ongoing work on analysis of within-herd distribution of ST in dairy herds in Denmark and Finland (Ruth Zadoks, personal communication).

Furthermore, in this study 6 herds had two BTM samples collected in 2009 (in October–November and then in the end of December). And all of them showed consistent MLST results: same ST for both isolates (one ST1, three ST23 and two ST103).

Also, for several arbitrarily chosen herds, repeated MLST results in 2011–2012 were assessed (Figure 2.4). The BTM samples were collected in February–June 2011 and in February–March 2012. Only one of 16 herds failed to show consistent ST for both samples, this indicates that late BTM sample collection for MLST analysis in 2011 is unlikely to impose biases on the results presented in this thesis.

The only herd that had different MLST results is potentially a subject to carry-over in the lab, because previously processed isolate had the same rare ST (it was given temporary number 999 as it had not been observed before this study), and the observed herd had consistent ST23 in 2009 and 2010, thus was assigned with ST23 for 2011.

2.3.3 Available strain typing results for 2009–2011

The distribution of collection dates of BTM samples used for MLST is presented in Figure 2.5 and is different from the distribution of sample collection dates for the national annual surveillance (Figure 2.1). The differences can be attributed to:

- Delay in collection of samples for MLST in 2011.
- Several samples for MLST from the same herd within the same year.

The summary of available MLST results is given in Table 2.3. Calendar years were used here to differentiate herds based on the sample collection dates because year assignment based on farming seasons might be ambiguous, e.g. result from sample collected in May 2010 can be used to assign ST for 2009 or 2010.

‘Not existent’ entries are records where an isolate was reported to exist at some point, but it wasn’t sent to the laboratory. For ‘No growth’ records an isolate could not be obtained (e.g. because it was dead), also one isolate from 2011 with incomplete allelic profile (sequencing

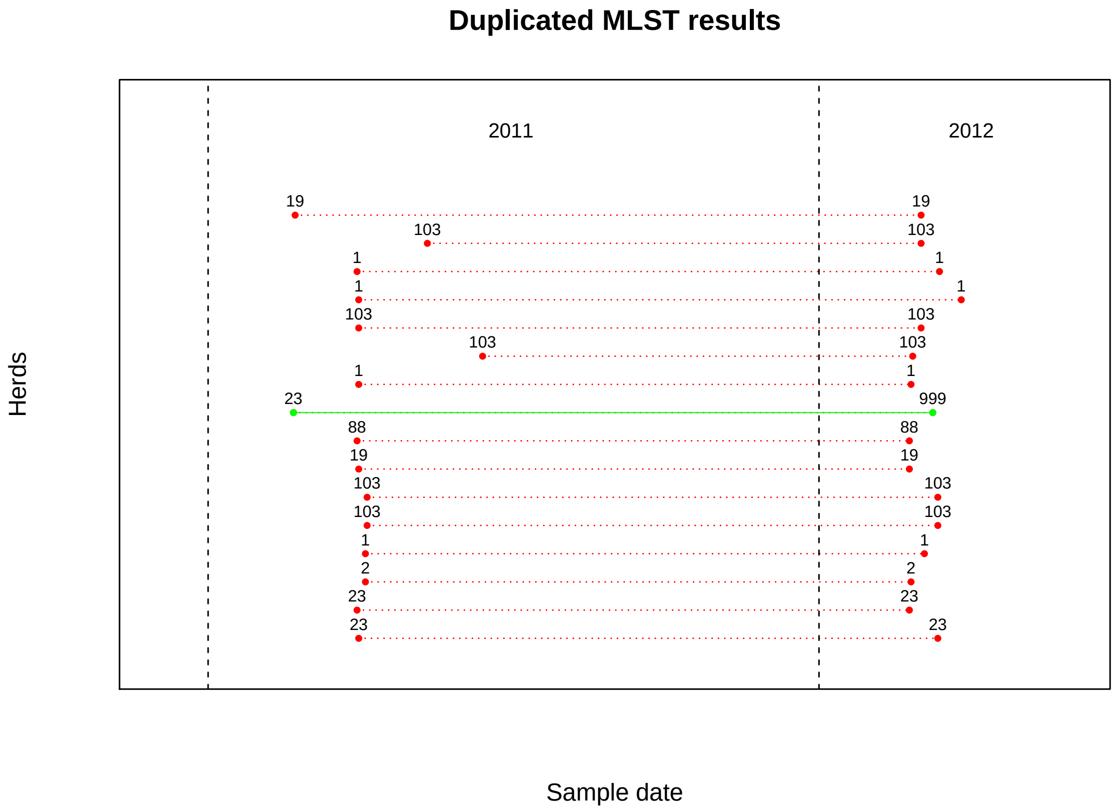

Figure 2.4: Repeated strain typing results for 16 herds in 2011. Only one herd (in green) had inconsistent results due to suspected carry-over in the lab.

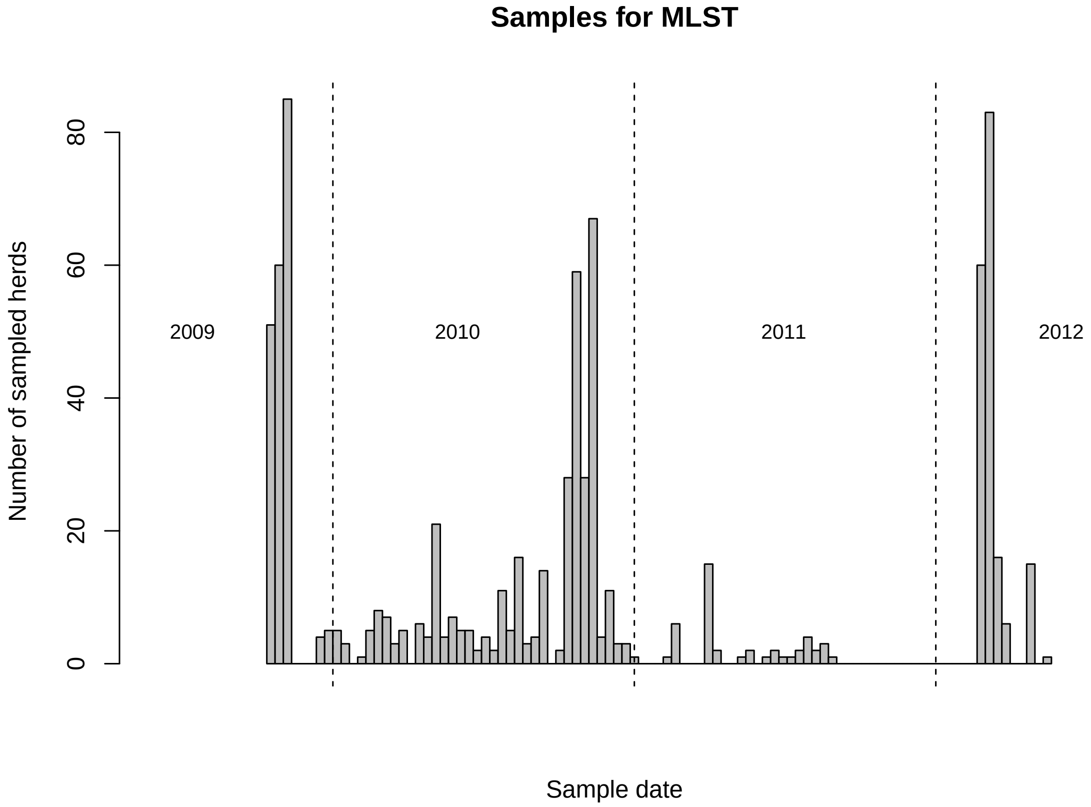

Figure 2.5: Distribution of bulk tank milk samples taken for strain typing in 2009–2011.

of one of the alleles failed) was put into this category. ‘Not S. ag.’ records mean that the cultured material contained some other bacterial species, not *Streptococcus agalactiae*. Hence, for these three categories ST0 was used to account for positive herds with unidentified strains.

Some of the molecular data had to be discarded as it arrived from false positive herds (due to carry-over during the sample collection).

Table 2.3: Available strain typing data.

Calendar year	2009	2010	2011
Total records (one per herd)	198	198	183
Not existent	5	3	-
No growth	1	6	5
Not S. ag.	1	-	4
Removed as FPs	14	1	9
Known ST	177	188	175
Unknown ST (ST0)	7	9	9

New sequence types

The public MLST database [Jolley and Maiden, 2010] (available at `http://pubmlst.org/sagalactiae/`) is mainly used by scientists to determine allelic profiles of previously observed STs. If a new allelic profile that is not present in the database is identified, it can be submitted to assign an ST number that will be used in the future.

Several STs obtained from the MLST results were absent from the public MLST database at the time of their detection, and thus had no officially assigned ST. They were assigned temporary ST numbers (Table 2.4) that will be used for further analysis. Most of the new STs were single locus variants (SLVs) of one of the previously reported STs (i.e. their allelic profiles differed in only one position).

Table 2.4: Newly observed sequence types in the course of this study. SLV is a single locus variant (one allelic position is different), DLV is a double locus variant.

Temporary ST	Year	Herds with ST	Relationship with observed STs
801	2010	1	DLV of ST1
923	2010	1	SLV of ST23
908	2011	2	SLV of ST8
910	2011	1	SLV of ST10
919	2011	1	SLV of ST19
926	2011	1	SLV of ST26
999	2011	1	completely new type

2.4 The principal dataset

2.4.1 Case definition

Both bacteriological culture and PCR tests can be used (independently or in combination) to determine herd infectious status.

Case definition for a particular year was based only on the annual surveillance test results. Additional samples were taken occasionally and mostly from previously infected herds (for confirmation or after farmer's request aiming to get negative result to get rid of the positive infection status), and therefore can introduce undesired biases (because the fraction of herds that have additional results is potentially biased).

Coverage of the bacteriological culture test in 2010 and 2011 was quite low (9% and 6% of registered dairy herds, respectively), suggesting that it can only be used as additional indicator of *Streptococcus agalactiae* infection.

There are three straightforward ways to define cases for further analysis (based on the official annual surveillance results only):

1. Herds confirmed by both PCR and culture
2. Herds that were tested positive by PCR (regardless the culture results)
3. Herds confirmed by at least one test (PCR or culture)

The summary of the three potential datasets is given in Table 2.5 (note that corrections for carry-over were taken into account).

Table 2.5: Three possible datasets based on various case definitions. Note that corrections for carry-over were taken into account in contrast to Table 2.2.

Year	PCR and culture	PCR only	PCR or culture
2009	168	300	316
2010	130	264	275
2011	159	221	221

The case definition that relies on positive results of both tests provides a relatively small dataset, unnecessarily leaving out a large amount of molecular data.

In 2009, all the registered herds were tested by both, PCR and culture, unlike in further years. In 2010, coverage of bacteriological culture was far from 100%, but around 100 PCR negative herds were tested. In 2011, only PCR positive herds were tested using bacteriological culture (and also several months after the official annual surveillance). Thus, the information on the number of culture positive PCR negative herds was incomplete. Previous statements confirm that PCR had consistent coverage for the whole study period (unlike culture) and might be used on its own as the evidence of *Streptococcus agalactiae* infection [Mweu et al., 2014].

However, use of herds with culture positive but PCR negative results is important because, while PCR is a more sensitive test [Mweu, 2013], culture's specificity was almost 100%, making it almost a definite proof of bacterial presence in the sample and, thus, confirmation of infection.

Therefore, it was decided to use the 'PCR or culture' dataset that added to the 'PCR only' dataset 16 and 11 herds for 2009 and 2010, respectively. Notably, most of added herds (i.e. PCR negative but culture positive) had meaningful (not ST0) MLST results (16 out of 16 for 2009, and 9 out of 11 for 2010).

For the dataset based on the 'PCR only', biases (under-reporting) are expected to be consistent for three years. For 'PCR or culture' dataset, the number of 'missing' (those that could have been included if the culture coverage was 100%) herds increases from 2009 to 2011 (from zero in 2009). However, the 'PCR or culture' dataset (compared to the 'PCR only') increases the number of defined cases, thus, will provide more power for further analysis.

2.4.2 Assignment of sequence types

There can be cases when ST was not obtained, but there was enough evidence to count the corresponding herd as positive (culture positive or PCR positive). All these herds were assigned with ST0 that will be treated as a different ST (from ST1, ST2, etc.) in further analysis. Although, ST0 could have been assigned to herds that have one of the previously observed STs.

Assignment of sequence type for positive herd can be ambiguous if BTM samples for MLST and annual testing (by PCR) were collected at different dates: there is a possibility that at the time of annual screening ST associated with the observed herd can be different. However, based on the previously mentioned evidence that ST is usually persistent for a herd, it was decided to use available information on ST even if it was from a different date. This also applies for 2011 when samples for MLST were collected several months after the annual screening.

Negative herds that had MLST records were converted to positive (based on the fact that presence of MLST record implies the positive result of bacteriological culture: bacteria had to be grown in the lab to obtain an isolate) and added to the dataset (9 herds for 2010 and one for 2011, see column (1) in Table 2.6). Also, for herds with no molecular data from the annual BTM samples, additional MLST results based on samples collected between annual surveillance periods were used. Thus, extra 25 herds (for the whole study period) were assigned with STs (column (2) in Table 2.6).

Table 2.6: Additional sequence type assignments: (1) herds that were negative by both PCR and culture (during the annual surveillance) but had MLST results were added to the dataset; (2) additional MLST results outwith annual surveillance were used to treat missing STs for positive herds; (3) for some persistently infected herds with only two MLST results, the missing ST was inferred.

Year	Negative herds with MLST (1)	Additional ST used (2)	Inferred ST (3)	New number of herds
2009	-	9	14	316
2010	9	15	8	284
2011	1	1	8	222

Inferring missing sequence types

For some herds and for some years, STs could not be obtained in the lab, which led to the assignment of ST0 (that is an indication of unidentified strain type). In some cases, missing data could be inferred based on the molecular results from other years, given that for Danish dairy herds the median duration of infection with *S. agalactiae* was previously estimated to be 2 years [Mweu et al., 2012].

For herds that were confirmed positive for each of three years but had MLST results only for two of them, I attempted to infer the missing ST. It appeared that for herds that had two (out of three: 2009–2011) consistent STs among ST1, ST23 and ST103, it was common to observe the same ST for the third year. The same did not apply to ST2, ST19, ST314, ST461 and ST627 (see Table 2.7). It was decided to substitute missing data for herds with consistent ST1, ST23 and ST103. And also for herds with two consistent results of ST314 and ST461, because for them the missing ST was in 2010 (the middle year). Herds with two consistent results of ST2, ST19 and ST627, as well as herds with inconsistent STs were not corrected and remained to have ST0. The numbers of inferred STs are presented in column (3) in Table 2.6.

The number of herds with inferred STs was small enough to affect the clustering tests results (in Chapter 3) and also did not bias the network analysis results (in Chapter 4). More details can be found in the next section.

Table 2.8 describes the final version of the principal dataset.

2.4.3 Sensitivity analysis

Although the process of preparation of the principal dataset was described and justified in the previous sections, the sensitivity analysis was performed to reveal if results of further

Table 2.7: Observed sequence type combinations to support inference for missing strain typing data: for each yearly ST pattern (e.g. [?, 1, 1]) the numbers of herds with particular ST are presented (e.g. for pattern [23, ?, 23], 11 out of 16 herds had [23, 23, 23] profile). NA for ST was used if the herd was negative in a particular year, and ST0 if the ST was not identified but the herd was positive. Most of herds infected with ST1, ST23 and ST103 in at least two out of three years, had the same ST in the third year, suggesting that missing data can be substituted with consistent ST (from the other two years). The same did not apply to other strains: ST2, ST19, ST314, ST461, ST627.

?, 1, 1	
ST	Number
NA	4
0	3
1	12

1, ?, 1	
ST	Number
NA	2
0	4
1	12

1, 1, ?	
ST	Number
NA	5
0	4
1	12
26	1
103	1

?, 23, 23	
ST	Number
NA	3
0	4
23	11

23, ?, 23	
ST	Number
NA	1
0	3
23	11
630	1

23, 23, ?	
ST	Number
NA	1
NA	2
0	1
23	11
999	1

?, 103, 103	
ST	Number
NA	1
0	1
103	5
588	1

103, ?, 103	
ST	Number
NA	3
0	5
103	5

103, 103, ?	
ST	Number
NA	1
0	3
103	5
296	1

2, 2, ?	
ST	Number
NA	1
0	1
23	1

19, 19, ?	
ST	Number
NA	4
0	1
19	1

314, ?, 314	
ST	Number
NA	1
0	1
314	1

461, ?, 461	
ST	Number
0	1

627, 627, ?	
ST	Number
0	1

Table 2.8: The principal dataset.

Year	Number of herds	Known ST	Unknown ST (ST0)	Percentage known
2009	316	195	121	61.7%
2010	284	185	99	65.1%
2011	222	163	59	73.4%

analyses depend on the choices made.

For this purpose, the conservative dataset (cases were defined by PCR only, no corrections for STs were made) was used. Results of K-function analysis (methods are described in the next Chapter 3) for the conservative dataset are presented in Figure A.1 and show significant clustering of ST103 throughout the whole study period, which is in agreement with the results for the principal dataset shown in Figure 3.4.

Results of the network analysis (Chapter 4) depended on a small number of “effective” links, i.e. links corresponding to contacts between source farm that was infected with a particular strain in the previous year and newly infected farms (farms infected with the same strain in the current year that were not infected in the previous year). Therefore, inferred and assigned STs might have changed the results dramatically. However, closer investigations on which herds were affected by the inference of missing STs showed that they could not change the final results. For the movement network only one herd with ST profile (461,?,461) participated in the “effective” link in 2010. Hence, the results on ST level for major STs (1, 23, 103) stayed the same, and for CC103 the “effective” link was one of four and could not change the outcomes significantly. The veterinary networks were denser and, thus, there were more cases when herds with inferred STs participated in “effective” links: one for ST23, one for ST103, three for ST314 and two for ST461 (all in 2010). But they also could not affect the results as their contribution in the number of “effective” links was not significant: one out of 11 for ST23, one out of 10 for ST103, and five out of 22 for CC103 (i.e. ST103, ST314 and ST461 combined together).

2.5 Molecular characterisation of isolates

Population composition of *Streptococcus agalactiae* isolated from annual BTM samples in 2009–2011 was consistent between years with respect to the three predominant strains: ST1, ST23 and ST103 (Figure 2.6). Also, it shows a steady decrease of ST10 and ST19. Note that the graphs only show identified STs (i.e. ST0 was not included).

eBURST diagrams [Feil et al., 2004] are used to visualise genetic similarities between STs: nodes (STs) are connect if they are closely related and numbers on edges represent the numbers of differences in allelic profiles of corresponding STs. eBURST diagram of available

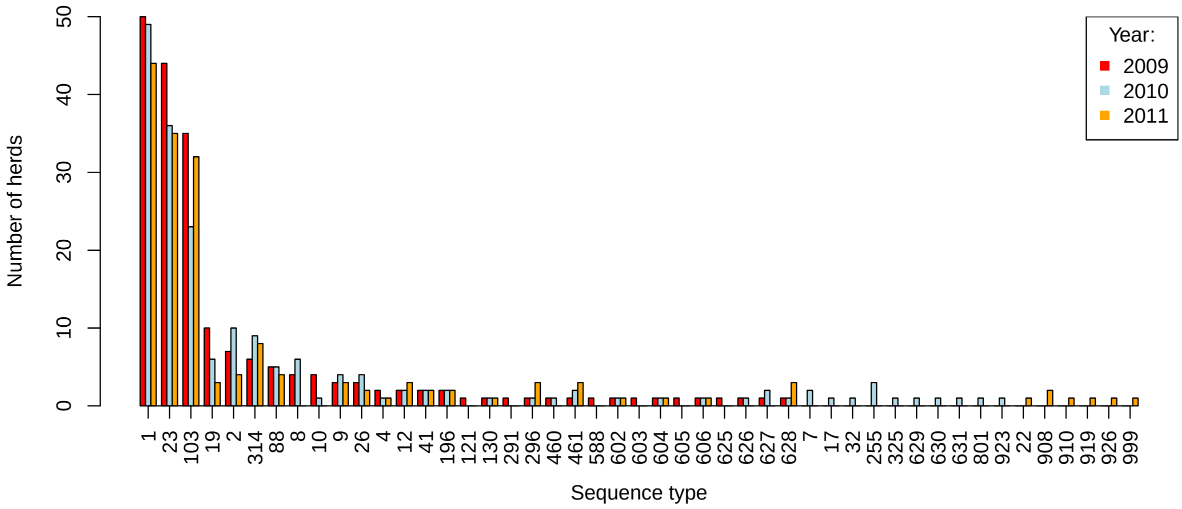

Figure 2.6: Population composition of *Streptococcus agalactiae* isolated from bulk tank milk in 2009–2011.

isolates is shown in Figure 2.7.

2.5.1 Clonal complexes

A considerable proportion of bacterial populations belongs to a limited number of clusters of closely related genotypes, referred as clonal complexes [Feil and Spratt, 2001]. Historically, clonal complexes are defined by a CC founder (the most representative ST) and its SLVs. However, in this thesis slightly altered definition will be used to account for other relevant STs that are not SLVs.

Clonal complexes are particularly of interest in this study as transmission properties are expected to be shared between bacterial subpopulations with similar genotypes (i.e. members of the same CC).

CCs used in this study are described in Table 2.9. It was decided to assign DLVs and TLVs (two and three differences in allelic profiles, respectively) to existing CCs[3] to minimise the number of subpopulations but preserve the biological relevance of the definition. Eight groups of closely related STs were identified, only ST999 was not classified into any of the clonal complexes.

[3] In case they were not SLVs to other STs that were CC founders.

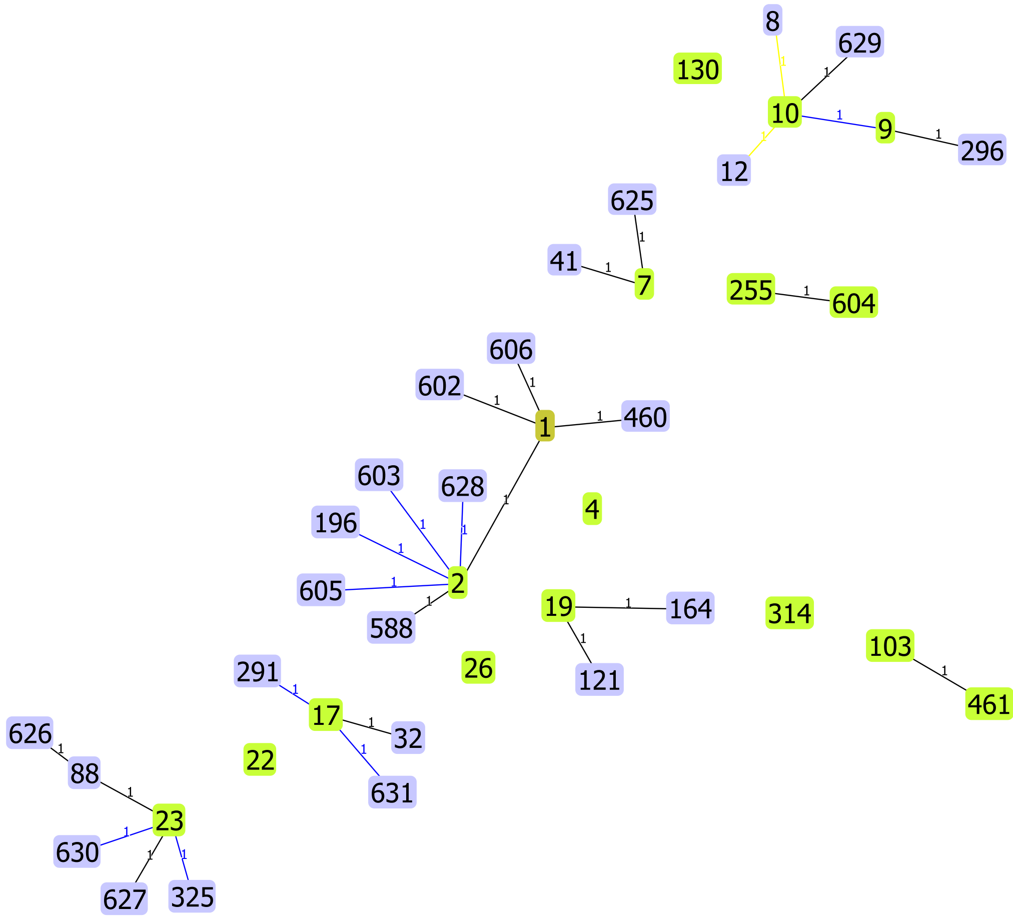

Figure 2.7: eBURST diagram of the bovine *Streptococcus agalactiae* isolates obtained from Denmark in 2009–2011: 41 previously observed STs, 7 new STs are not shown. Numbers on the edges show distance between isolates (number of different allele numbers in their allelic profiles: one is SLV, two is DLV, etc.). Primary STs of connected groups are in green.

2.6 Discussion

In this chapter, the process from the data collection to the choice of the dataset for further analysis was discussed. The explicit case definition allowed us to treat inconsistent (between study years) infection data based on two different testing techniques. The final dataset adequately summarises the available data, account for known errors and maximises the amount of the data taken into account.

2.6.1 Data preparation

Rarely raw data are suitable for analysis and have to go through substantial cleaning and editing first. It takes time to scrupulously go line by line and check the data for errors and inconsistencies, but if done properly it minimises the risk of introduction of manual errors. All the data manipulations must be transparent and well-documented. However, careful data

Table 2.9: Definition of clonal complexes used in the analysis.

Clonal complex	SLVs	DLVs	TLVs	Infected in 2009	Inf. in 2010	Inf. in 2011
CC1	2, 4, 460, 602, 606	196, 588, 603, 605, 628, 801		66	64	54
CC7	41, 625	255, 604		4	8	3
CC10	8, 9, 12, 629, 910	296, 908	130	15	16	13
CC17	32, 291, 631		22	1	3	1
CC19	121, 164, 919			11	6	4
CC23	88, 325, 627, 630	626		51	46	39
CC26	923, 926			3	5	3
CC103	461	314		42	34	43

manipulations do not guarantee the absence of errors, the data can have imperfections that could have been introduced earlier, in the process of data collection and data management. In most of the complex cases, there can be no single perfect decision. But whatever decisions are made by a researcher, they must be corroborated by evidence, and all the potential biases or errors must be acknowledged.

Two pieces of data that did not entirely match each other were used in this study: the annual surveillance test results and MLST data on the collected isolates.

At first, the annual surveillance data were used to make a decision on how to classify the status of the herds: herd that tested positive by either PCR or bacteriological culture was considered positive. This allowed for consistent case definition that maximised the data available for the strain-specific analysis.

The list of positive herds was complemented by molecular data to determine what strains were present on farms. Evidence of the ST consistency over time allowed to make use of MLST data based on samples that were collected outwith the official annual surveillance periods.

Some of the data had to be corrected. Herds that had negative PCR and culture test results along with identified ST were added to the dataset. Also, for some cases, STs were inferred based on the observation that predominant STs stayed consistent for most of the herds.

2.6.2 Potential biases

It is important to acknowledge the potential biases.

Errors can be introduce from the very first stage of data collection. The issue with carry-over during BTM sample collection had been addressed before but here extra data (MLST) were used to classify true and false positive herds.

Growth-inhibited and nonviable bacteria were assigned with ST0 that is treated as different to any other ST. However, this is not strictly correct, ST0 can potentially contain known STs (e.g. if for some reason culture with ST103 failed to grow, it will be treated as ST0).

2.6.3 Conclusion

The process of data cleaning and data preparation was explained in this chapter. The final dataset is consistent regarding the use of annual surveillance data and maximises the use of strain typing information. The latter is important given the strain-specific focus of the overall research questions.

Chapter 3

Spatial clustering of Danish dairy cattle herds infected with various strains of *Streptococcus agalactiae*

3.1 Introduction

Epidemiological studies are often focused on spatial components of the distribution of disease.

The spatial distribution of cases depends on transmission mechanisms of the pathogen. In the simplest case, infected individuals appear close to each other, forming distinctive groups (i.e. clusters), due to the point-source nature of the disease, however more complex dissemination strategies can also result in unusual occurrences of cases. For example, the spread of an infectious disease such as foot-and-mouth disease (FMD) is better described by a network of contacts, however, Gerbier and Chadeouf [Gerbier and Chadoeuf, 2000] determined clustering of cases within 1–3 km for retrospective data from the 1967 to 1968 UK foot-and-mouth disease outbreak.

Clusters of disease can occur at various levels: e.g. in animal populations, transmission via a wildlife or human reservoir implicates clustering at larger scale than localised common-source exposure.

Therefore, it is important to identify clusters of cases or other anomalies in the spatial distributions of infected individuals, which might be useful in determining the mechanisms that drive the spread. The clustering of disease events may help in identifying a common environmental factor or source of exposure.

In veterinary epidemiology, the unit of interest is often not an individual animal but a herd or a farm, which can be represented by a point on a plane.

Spatial clustering analysis was previously used in epidemiological studies for a number of pathogens and disease systems, including BVDV in Danish cattle [Ersbøll and Ersbøll, 2009] and *Staphylococcus aureus* in British cattle [García Álvarez et al., 2011].

In this study, spatial distributions of herds identified with particular strains of *S. agalactiae* were assessed. Differences between strains regarding their spatial clustering will support the hypothesis of presence of strain-specific transmission patterns [Zadoks et al., 2011].

3.1.1 Objective

The focus of this study is to determine whether different strains of *S. agalactiae* are associated with specific and distinct spatial patterns.

3.2 Materials and methods

3.2.1 Data

The principal dataset described in the previous chapter was used here. Information about locations of dairy herds was also extracted from the Danish Cattle Database.

Given that the population composition of *Streptococcus agalactiae* isolated from annual BTM samples in 2009–2011 was dominated by three strains: ST1, ST23 and ST103 (Figure 3.1), the further analysis will be focused on them and the corresponding clonal complexes.

For positive herds with no assigned ST, it was decided to treat them as unidentified STs (i.e. ST0), different from all known Danish STs. Effectively, they were excluded from the strain-specific analysis.

3.2.2 Spatial clustering methods

Here I focus exclusively on spatial clustering, i.e. non-random spatial distribution of objects where Euclidian distance is used (opposed to social and network clustering, where proximity of objects is defined as a network distance between two nodes). It was decided not to focus on other distance metrics here because:

1. Euclidian distance measures are available, and they are easy to work with. Measures of other metrics are rarely available (in our case, network distance can be used for two available networks, but this will be addressed in the next chapter and using a different approach).

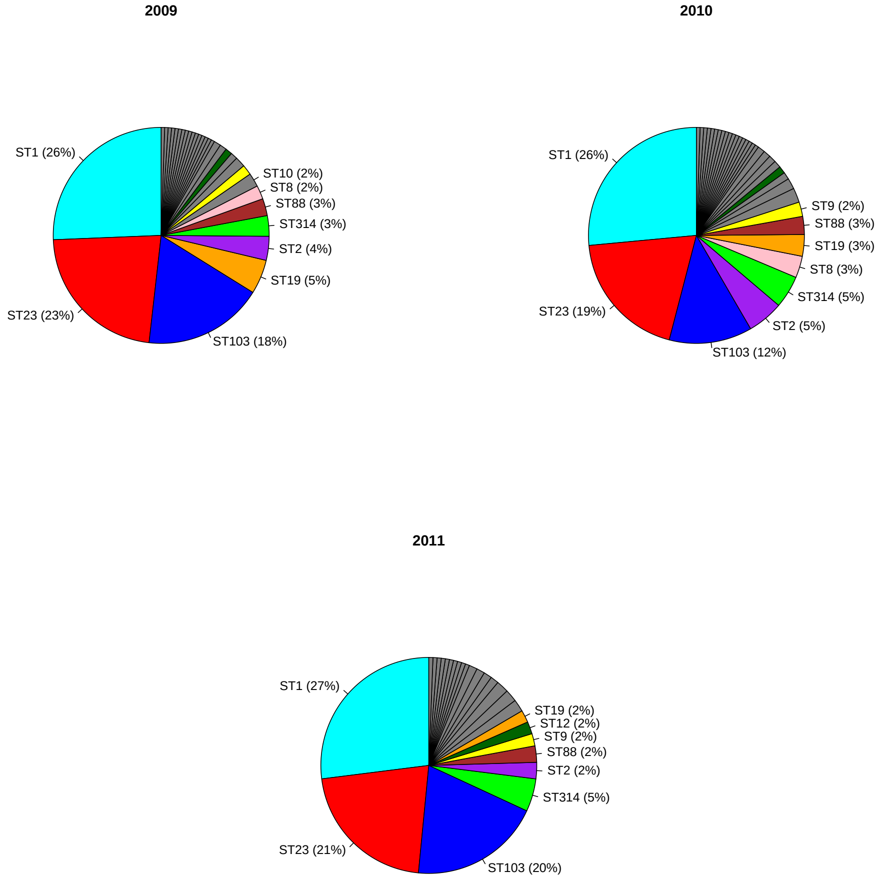

Figure 3.1: Predominant strains in the population composition of *Streptococcus agalactiae* isolated from bulk tank milk samples in 2009–2011.

2. Euclidian distance may be significant (and has been previously used) for those pathogens that can spread locally (e.g. FMD). Thus, we are testing for potential local transmission of *S. agalactiae* or transmission mechanisms that are effectively equal to local.

3. Even though landscape features (coastal line, lakes, urban areas) might be a problem (as herd loactions do not follow a homogeneous Poisson process), most of clustering methods take into account the underlying spatial structure (we are not simulating random point process but reassign STs between study herds, thus restricting possible spatial distributions).

Most of clustering detection methods test the null hypothesis that there is no clustering in the spatial data. The corresponding test statistic is measured and if it is significantly large (or small, depending on the statistic used), the null hypothesis is rejected.

Several clustering tests were compared in [Song and Kulldorff, 2003]. Most of them evaluate the presence of clustering, but do not have the ability to identify the locations of clusters.

It is important that tests for spatial randomness adjust for inhomogeneous background populations, i.e. spatial clustering should be measured relative to the population, not to the landscape features. This is accomplished by selecting appropriate controls for the cases. A similar cluster would be expected for the controls and cases under the null hypothesis, i.e. if there is no spatial disease-association.

The choice of methods for this study was based on the following considerations:

- Tests for spatial randomness should adjust for inhomogeneous background populations (because Denmark has substantial and irregular coastline).
- The tests should reflect several different paradigms rather than be modifications of one approach.
- Preference was given to the tests that were used previously for epidemiological analyses.

Here, the spatial clustering will be analysed using three tests that adjust for the underlying population at risk: Cuzick–Edwards' kNN test, K-function and the spatial scan statistic.

A comprehensive review of spatial clustering methods can be found in [Carpenter, 2001].

Cuzick–Edwards' kNN test

The Cuzick–Edwards' k nearest neighbours (kNN) test [Cuzick and Edwards, 1990] is a significance test to detect the possible clustering of groups formed by a particular node (the cluster centre) and its k nearest neighbours within the overall population.

The test statistic is the overall number of cases when two herds belong to the same class (e.g. have the same ST) and one of them is among another's k nearest neighbours and it is given by:

$$T_k = \sum_{i=1}^{n} \delta_i d_i^k, \tag{3.1}$$

where δ_i equals 1 if i has the considered ST, and 0 otherwise; d_i^k is the number of herds with the same ST among k nearest neighbours of i.

In order to test the clustering hypothesis, the test statistic of the observed distribution is compared against those of the randomly simulated ones, for different numbers of k. Spatial clustering for a particular k is indicated if T_k for the observed distribution is higher than for the 97.5%[1] of the simulated distributions. Random distributions were obtained by reassigning STs and infection statuses for the dairy herds.

The kNN test does not implicitly use spatial distance between nodes, and is not restricted to circular clusters. There is no obvious connection between neighbours and kilometres: it depends on the density of points. Therefore, the kNN test and other spatial clustering tests that are based on explicit Euclidian distance (e.g. K-function) measure different kinds of clustering, and agreement/disagreement of their results will be an indication of clustering consistency or clustering features (e.g. spatial distance is not important for likelihood of transmission, it is more likely to get infected from the nearest neighbour than any other farm).

As it was previously mentioned, one of the advantages of this test is that it adjusts for the presence of heterogeneously distributed population at risk. It also performs well in detecting multiple clusters [Song and Kulldorff, 2003].

The kNN method was previously used by García Álvarez et al. [García Álvarez et al., 2011] to test spatial clustering of dairy cattle herds infected with particular STs of another bacterial pathogen causing bovine mastitis *Staphylococcus aureus*.

K-function

The K-function is another measure of global[2] clustering. It is defined as the expected number of events (i.e. cases; in our case, herds with a particular strain) within a certain distance, divided by the density of the points (or nodes). Usually, clustering is detected by the difference between the observed K-function and completely spatially random point process. But for countries with irregular coast lines and other landscape features (e.g. lakes, urban areas), the underlying population at risk cannot be simulated using the random process. This is a case for Denmark, where dairy herds are situated heterogeneously throughout the country. In such conditions, a generalisation of the K-function for inhomogeneous point patterns, proposed by Baddeley et al. [Baddeley et al., 2000], can be used. Here, I used the inhomogeneous K-function implemented in the **spatstat** R package [Baddeley and Turner, 2005] to define the difference between inhomogeneous K-functions for cases and the entire population at

[1] The standard way to report results of Cuzick–Edwards' kNN test is to plot T_k for the observed distribution along with the 95% envelope of T_k for simulated distributions. Thus, if the observed distribution is significantly different from the simulated ones (i.e. T_k is outside of the 95% envelope), there are two options: significant clustering (if the observed T_k is higher) or dispersal (if the observed T_k is smaller). Hence, the cut-off threshold for significant clustering was 97.5%. The threshold of 95% will not guarantee that the observed distribution is significantly different from the random.

[2] Here by global I understand a measure for the whole population, rather than for individual cluster (like in the spatial scan statistic). For K-function it is a measure of all groups of nodes within a particular distance.

risk (as proposed in [Ersbøll and Ersbøll, 2009]):

$$D(s) = K_{inhom}^{case}(s) - K_{inhom}^{pop}(s), \tag{3.2}$$

where s is distance (scale of clustering).

Monte Carlo randomisation was used to permute locations of cases (implemented by randomly reassigning STs for the considered herds using *rlabel* function) in order to generate random distributions. The observed difference function $D(s)$ and 95% simulation envelopes for random simulated distributions are plotted against the distance s to detect clustering. Deviations from the null hypothesis can be detected by distances where $D(s)$ lies outside the 95% simulation envelope, in this case the null hypothesis of no clustering is rejected.

K-function has been used in a number of epidemiological studies [Ersbøll and Ersbøll, 2009, Ahmed et al., 2010, Mweu et al., 2014].

Spatial scan statistic

The spatial scan statistic [Kulldorff, 1997], contrastingly to previously mentioned methods, is able to locate a cluster and test its significance. It identifies the most likely clusters over all circular windows, which vary in radius and the circle centroid position.

If $L_{j(i)}$ is the likelihood under the alternative hypothesis that there is a cluster for circular window[3] with centre in i and radius of j, and L_0 is the likelihood under the null hypothesis (no cluster), then:

$$\frac{L_{j(i)}}{L_0} = \left(\frac{D_{j(i)}}{U_{j(i)}\frac{C}{N}}\right)^{D_{j(i)}} \left(\frac{C - D_{j(i)}}{C - U_{j(i)}\frac{C}{N}}\right)^{C - D_{j(i)}}, \tag{3.3}$$

where C is the total number of cases, $D_{j(i)}$ is the number of cases within circular window of i, N is the total population size and $U_{j(i)}$ is the population size within circular window of i.

The null hypothesis of no clustering is rejected when the test statistic[4]:

$$T = \max_{i,j}\left(\frac{L_{j(i)}}{L_0} Ind\left(D_{j(i)} > C - U_{j(i)}\frac{C}{N}\right)\right) \tag{3.4}$$

is large. As for the kNN method, 97.5% of the simulated distributions was used to determine the threshold for significant clustering.

The spatial scan statistic is good at detecting localised clusters [Song and Kulldorff, 2003].

[3] Alternatively, circular window can be substituted by node i and its j closest neighbours, which will make this method similar to Cuzick–Edwards' kNN. But in this thesis the definition with circular windows was used.

[4] $Ind()$ is an indicator function, i.e. equals 1 if the condition is true, and 0 otherwise.

3.3 Results

3.3.1 Spatial clustering of all infected herds

At first, the total population of infected herds was tested for spatial clustering (using K-function method). Cases of *Streptococcus agalactiae* in the total population of dairy cattle herds were mostly unclustered (Figure 3.2) with indication of clustering in 2010 for distances between 15–45 km.

3.3.2 Spatial clustering of individual sequence types

Here we test if spatial distributions of particular STs are different.

For spatial clustering analysis of individual STs only positive herds identified in the previous chapter (those that showed evidence of bacterial presence by either PCR or culture test during the national bulk tank surveillance) were used, i.e. spatial clustering within the infected population was tested.

Alternatively, one can consider distributions of herds with a particular ST among all the herds regardless their infection status (the results for K-function analysis are presented in Figure A.2 in the Appendix).

The two approaches usually yield the same results, the differences may appear only if the case population is clustered (thus, subpopulations for individual strains might be clustered within the total population while not clustered within the case population). In our case, the case population is not clustered (Figure 3.2), hence the two methods are effectively equivalent. The choice of the approach to use in this thesis was motivated by the fact that in previous studies [García Álvarez et al., 2011] clustering of particular strains was also tested against the infected population.

Locations of herds infected with strains that dominated in the study population (i.e. ST1, ST23, ST103; see Figure 3.1) were tested for spatial clustering in the population of other positive herds (herds with the selected individual ST were considered as cases and other herds were considered as controls). Other STs (that had enough cases for the analysis) showed no significant clustering (see Figure A.3 in the Appendix). Some of the most infrequent types did not allow tests to be performed due to the small case population size.

Cuzick–Edwards' kNN

The Cuzick–Edwards' kNN test was implemented as a program in Java. 500 random ST distributions (preserving the same ST composition) were generated to compare their test

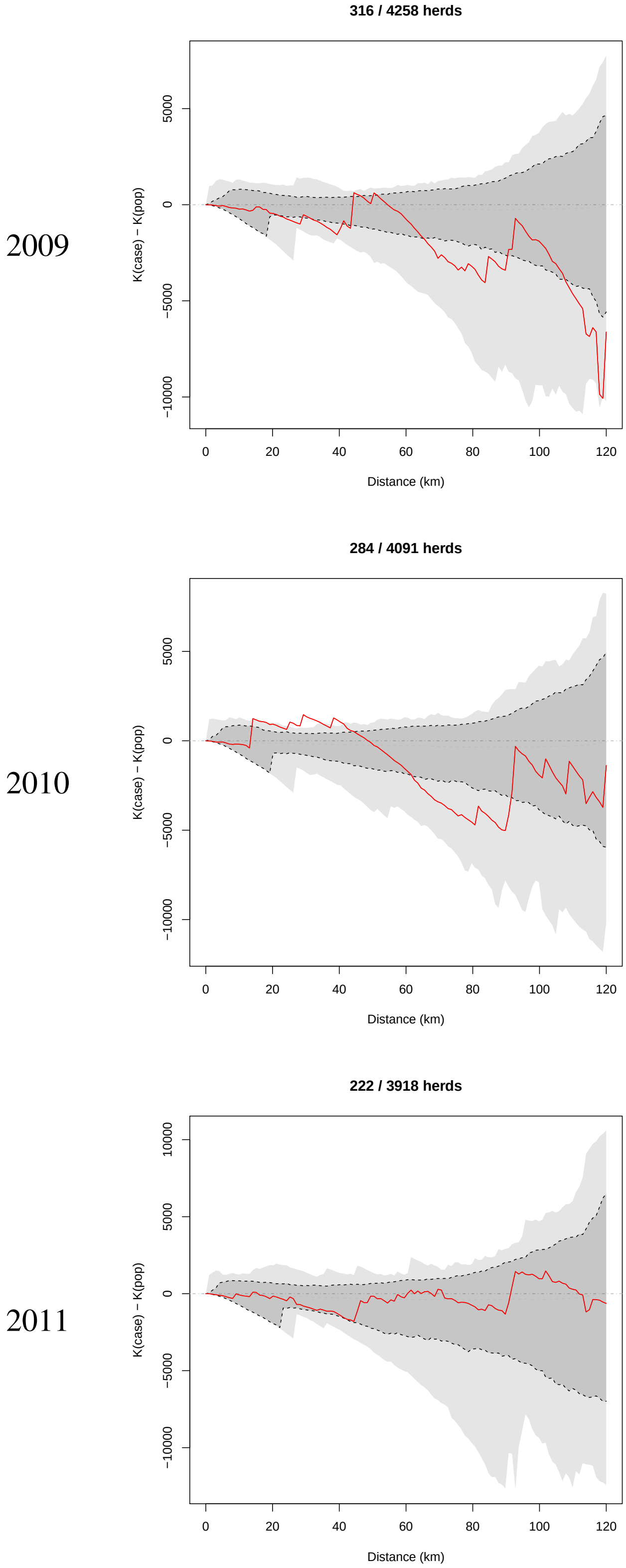

Figure 3.2: Results of K-function method for all infected herds: positive herds were not clustered in the overall population (apart from 2010 at a certain scale). Clustering is detected when the estimation of the difference function $D(s)$ for the observed distribution of cases (red line) is significantly higher than for the randomly generated distributions (dark gray envelope). Numbers on top represent the number of cases and the total population at risk. Light gray area defines extreme values of $D(s)$ over all simulated distributions.

statistic against the one for observed ST distribution for number of neighbours k from 1 to 50 (given the case population size). Results of Cuzick–Edwards' kNN test (Figure 3.3) showed that only herds infected with ST103 were clustered for all three years. Herds with ST1 were also clustered in 2010, but not in other years. Spatial distribution of ST23 was not significantly different from random.

K-function

Results of K-function analysis presented in Figure 3.4 corroborated those of the kNN test: among the observed strains, ST103 was the only strain that was clustered in 2009–2011. And ST1 once again was clustered at a relatively small scale (around 10 km) in 2010.

The spatial scan statistic

Locations of clusters were identified using the spatial scan statistic implemented in SaTScan software (Figure 3.5). Only ST103 has clusters with log-likelihood greater than 10 for each year. ST1 has a significant cluster in 2010 (21 out of 60 herds), and ST23 in 2010 (although quite small — 4 herds). Other clusters have small values of log-likelihood (less than 6) and can be treated as insignificant.

3.3.3 Spatial clustering of clonal complexes

Clonal complexes (CCs) defined in the previous chapter were also tested for spatial clustering (using K-function). Case was defined if ST form the herd belonged to the considered CC.

No significant spatial clustering was observed for either of 8 CCs[5] (see Figures A.4 and A.5 in Appendix). Small number of cases for some CCs made it impossible to run the analysis (e.g. one herd for CC17 in 2009). Interestingly, CC103 showed no significant clustering, while ST103 and a strain that combined ST103 and ST461 (SLV of ST103) did (see Figure 3.6). Although, contribution of ST461 was limited (1, 2 and 3 cases in 2009–2011, respectively) and this was mainly due to the clustering of ST103.

[5] CC19 did show the evidence of spatial clustering, but this was on a very small scale only and due to the fact that two herds were close to each other in some sparse region (the overall population was 4 cases).

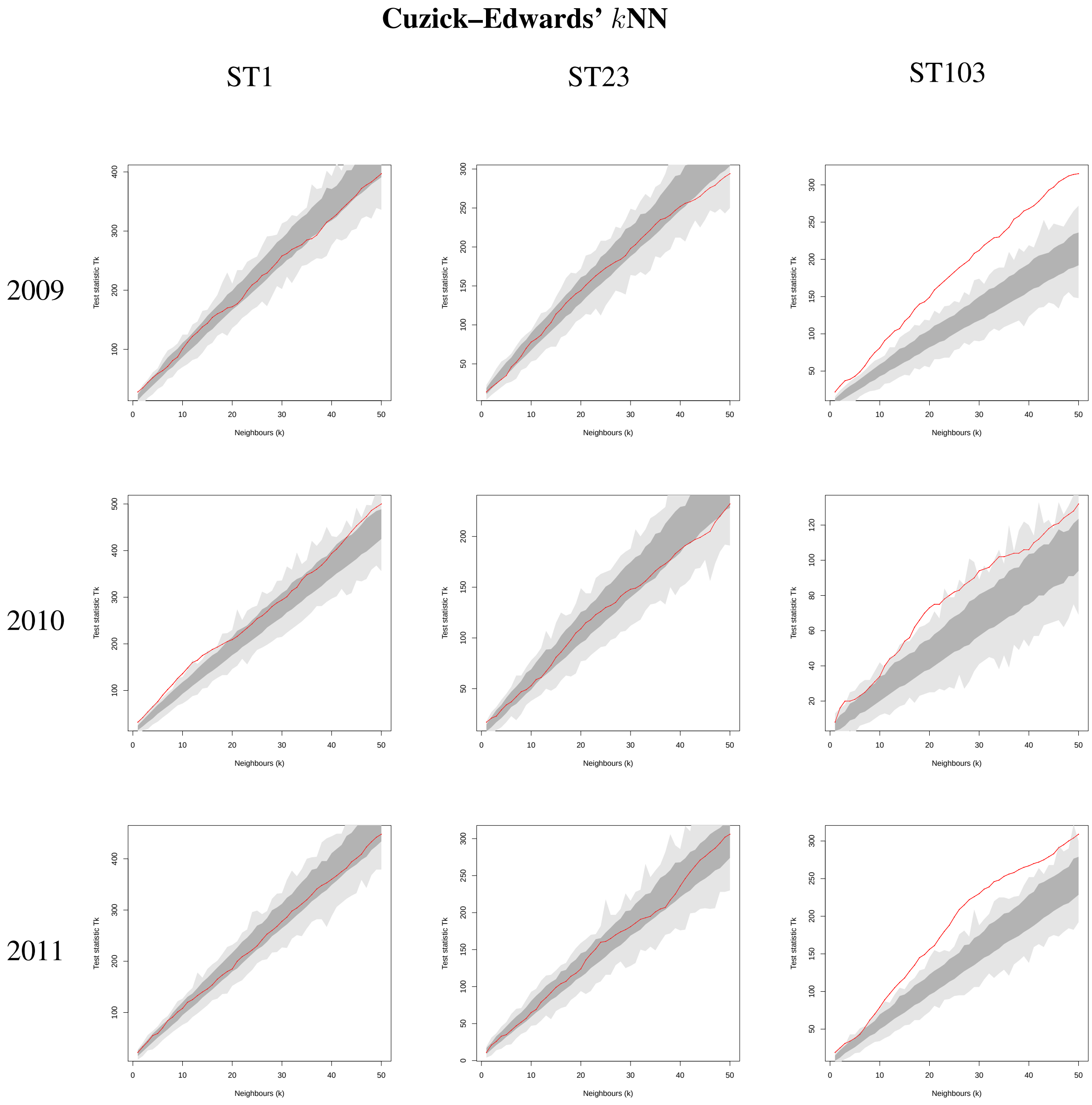

Figure 3.3: Results of Cuzick–Edwards' kNN test for three predominant STs: clustering is detected when the test statistic for the observed ST distribution (red line) is significantly higher than the values of the test statistic for randomly generated ST distributions (shown in light gray, 95% confidence intervals are in dark gray).

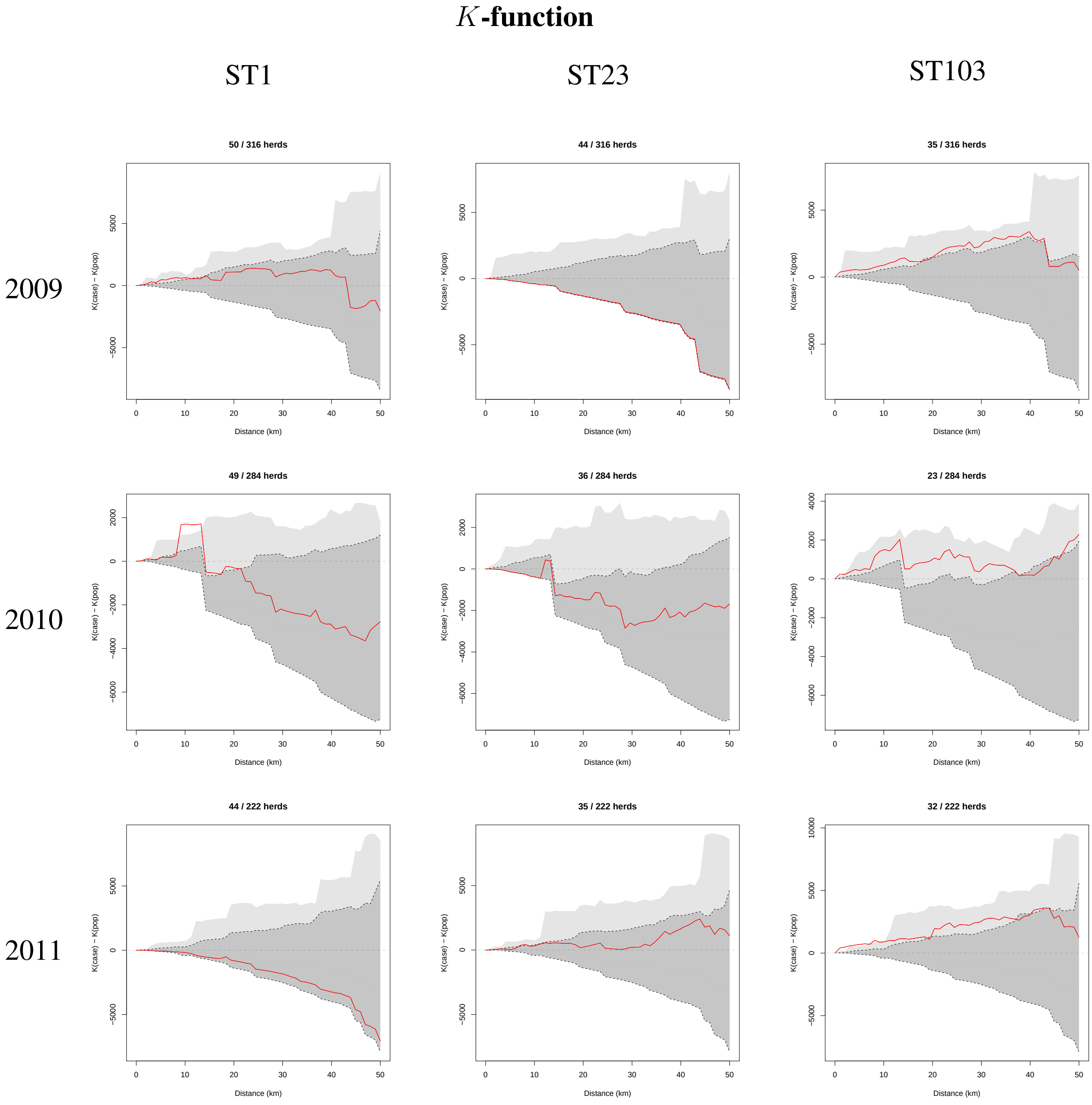

Figure 3.4: Clusters of predominant STs (ST1, ST23, ST103) detected using K-function in 2009, 2010 and 2011. Clustering is detected when the K-function estimate for the observed ST distribution (red line) is significantly higher than those for randomly generated ST distributions (shown in light gray, 95% confidence intervals are in dark gray).

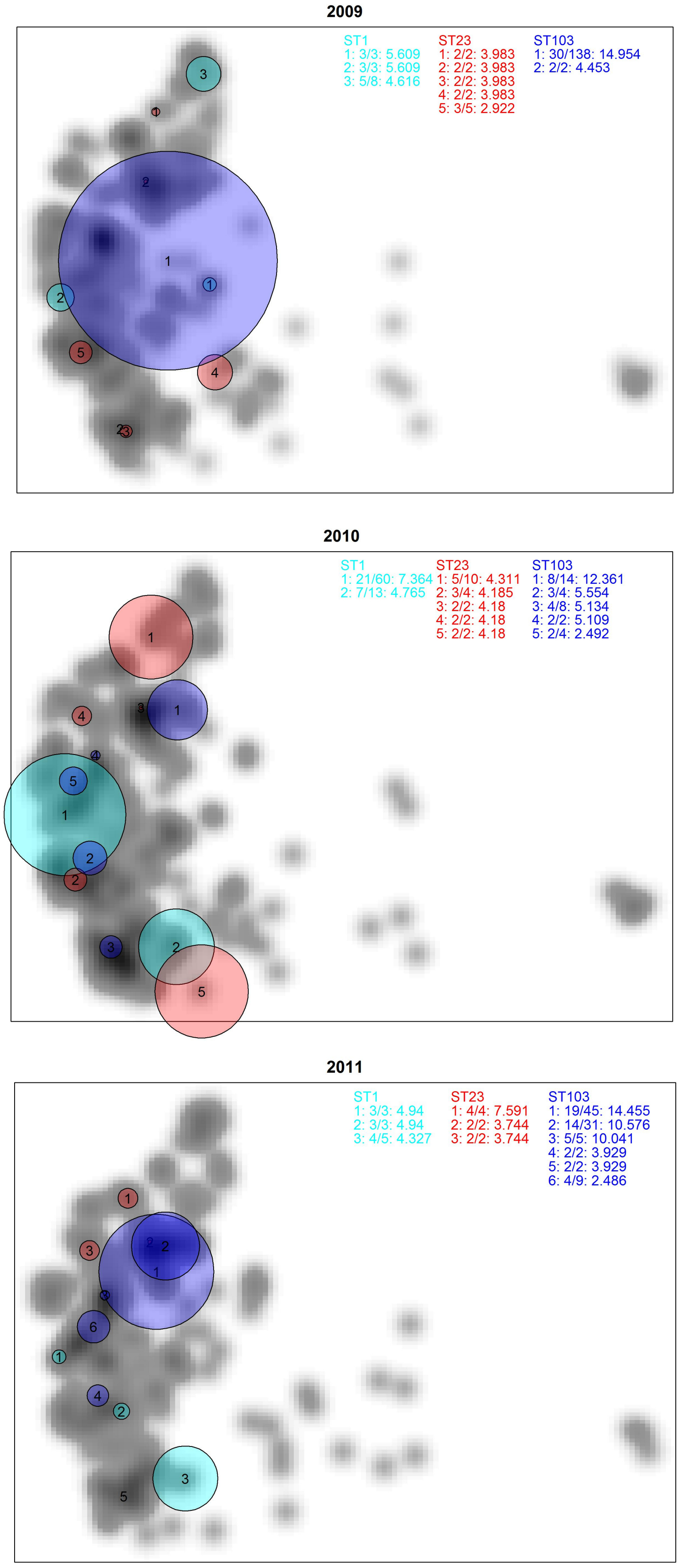

Figure 3.5: Clusters of herds with the three predominant STs detected using the spatial scan statistic (SaTScan) in 2009–2011. Density plots in the background show densities of all positive herds. Colours used for clusters: ST1 (cyan), ST23 (red), ST103 (blue). Numbers in circles indicate the rank of clusters according to their log-likelihood and refer to the table (top right corner of each figure) with information on clusters: number of cases in cluster, size of cluster and the log-likelihood.

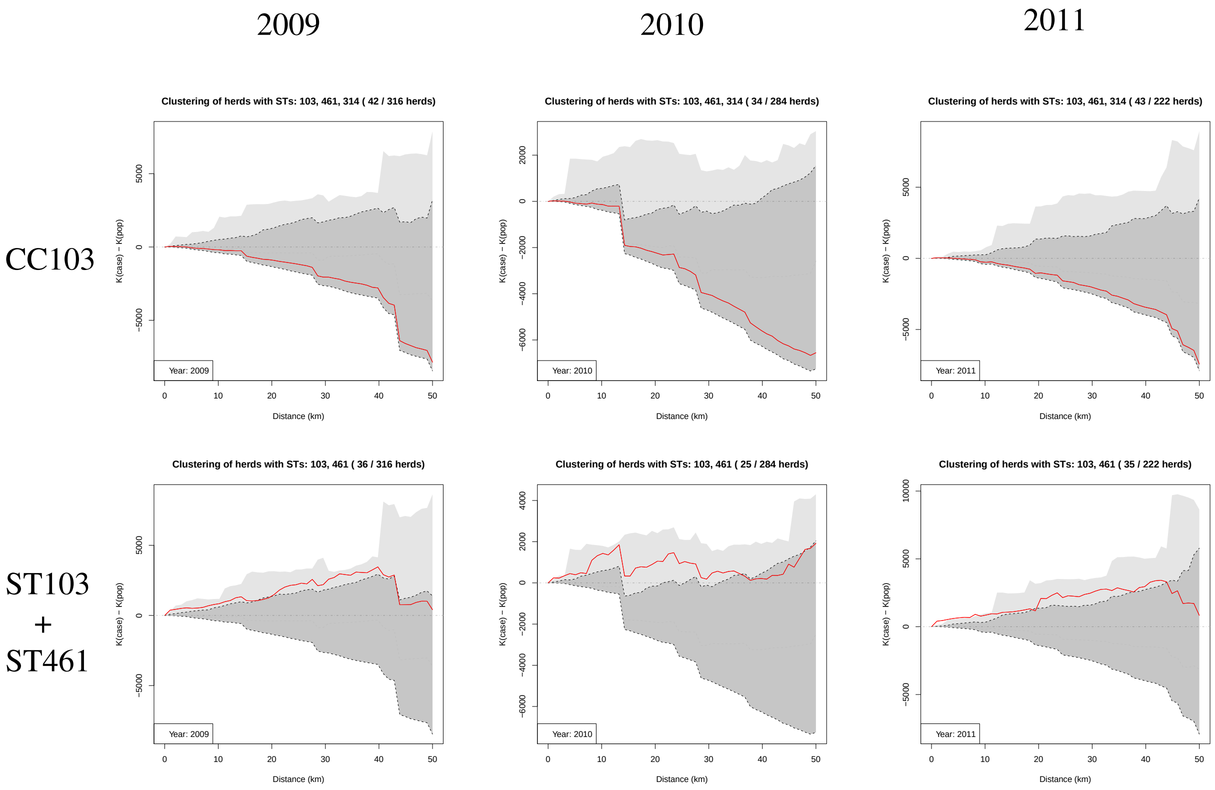

Figure 3.6: Results of K-function method for clonal complexes of ST103: spatial clustering was detected for ST103 + ST461, but not for CC103 (ST103, ST461, ST314). Clustering is detected when the estimation of K-function for the observed distribution of cases (red line) is significantly higher than for the randomly generated distributions (dark gray envelope). Numbers on top represent the number of cases and the total population at risk.

3.4 Discussion

3.4.1 Statistical power

Certain minimum number of cases among herds is needed to answer the question if they are spatially clustered or not. If the set of considered cases is too small (e.g. for relatively infrequent ST) then the results of clustering methods will be meaningless (some software for spatial clustering will not even work in this case). Therefore, use of CC level analysis (inclusion of genetically closely related strains) is a way to increase statistical power of the outputs.

3.4.2 Multiple testing problem

Cuzick–Edwards' kNN and K-function tests may suffer from multiple testing if the test is repeated for different numbers of k: the null hypothesis can be rejected for some numbers of k by chance. However, our results for ST103 show significant clustering for almost all of the k values, suggesting that effect of the multiple testing does not change the conclusions about the clustering of herds infected with ST103.

3.4.3 Comparison of different clustering methods

The three clustering methods are conceptually different (use relatively different paradigms). K-function is based on Euclidian distance, whether Cuzick–Edwards' kNN takes into account a preset number of nearest neighbours regardless the actual distances. SaTScan assesses properties of individual circular windows rather than global overall statistics. Therefore, it allows identification of clusters positions.

There are difficulties in comparing the results of clustering tests. K-function cannot be directly compared with the kNN method because they use different scales (explicitly spatial and "in nearest neighbours", respectively). For dense/sparse regions the results will be the most different because the kNN disregards density of points, it takes into account only their order from the centre of the considered cluster. SaTScan, unlike the other two methods, measures its test statistic for one cluster only, not the total measure for all the clusters of the same scale. This leads to detection of small localised clusters that are missed by other methods. They might be interesting in outbreak investigations but are not very helpful to determine overall transmission mechanisms of the considered strain.

However, all the three considered methods confirm spatial clustering of ST103 throughout the study period. Also, the methods agree in clustering of ST1 in 2010 for certain scales (although the scales were incomparable). This agreement of several methods affirms that detected clustering is not due to an artefact of the data that was incorrectly captured by one of the test.

3.4.4 Spatio-temporal clustering

Temporal components of the distribution of disease are also of interest in epidemiological studies. And spatio-temporal clustering can be assessed using the extensions of the discussed methods [Ahmed et al., 2010]. However, this is more relevant to disease outbreaks rather than endemic diseases like bovine mastitis. And since the temporal data for this project are restricted to the yearly periodicity due to the annual surveillance scheme, this chapter focused on spatial aspects only.

3.4.5 Spatial clustering of ST103

The results show that ST103 is different among three other frequent sequence types, because it was persistently clustered for the study period, while others were not. It is not clear what this difference can be attributed to, but these findings suggest that ST103 has transmission mechanisms that are different from those of other major STs.

Interestingly, ST1 and ST23 are quite common in humans, whereas ST103 was previously rarely reported in humans, but was common in some cattle studies [Yang et al., 2013]. One might hypothesise that transmission of ST103 is based on cattle movements, however it is not clear how movements result in spatial clustering, and further investigations are needed to test this hypothesis.

3.4.6 Spatial clustering of clonal complexes

Use of CCs for spatial clustering analysis does not only increase statistical power but is also biologically reasonable. STs that have closely related genotypes are expected to share transmission properties. Another evidence to use CC level analysis is that in a number of cases ST and its SLVs were recovered from the same herd (Ruth Zadoks, personal communication).

However, none of the 8 CCs that were considered in this study showed significant spatial clustering. Interestingly, even CC103 was not clustered in any of the study years, in contrast to its founder ST103. Exclusion of ST314 (DLV of ST103) from the analysed CC leads to detectable clustering for ST103 and ST461 grouped together, but this is due to the limited contribution from ST461. To conclude, spatial clustering analysis has not revealed matching transmission patterns for ST103 and its closely related ST314, suggesting that transmission properties of *S. agalactiae* are more likely to be ST specific rather than CC specific.

3.4.7 Further analysis

Spatial clustering tests enable researchers to identify non-randomness in the distribution of pathogenic subpopulations. However, these tests could not reveal the mechanisms that led to these spatial distributions, and other methods should be utilised for explanation of the observed patterns.

Differences in spatial clustering for ST103 suggest that its possible transmission routes via cattle movements or other mechanisms should be investigated using appropriate methods.

Chapter 4

Contact patterns between Danish dairy cattle herds and their role in the spread of various strains of *Streptococcus agalactiae*

4.1 Introduction

Heterogeneity of contacts between individuals complicate the nature of disease transmission: it is harder to predict outcomes of an outbreak and, therefore, control the disease.

One of the most important types of contacts within livestock populations are described by trade networks or, more generally, movement networks. Relocation of live animals between agricultural holdings is an important driver for spreading infectious diseases [Keeling et al., 2001, Green et al., 2006, Ortiz-Pelaez et al., 2006].

Traceability of livestock is implemented via registration of animal movements. Recorded movements of animals compose a directed complex network that can be used in epidemiological studies.

However, contact networks (in epidemiological context) are not restricted to live movements and can, for example, represent associations between neighbouring farms, between farms sharing equipment, between farms having common on-farm visitors, etc. Basically, any kind of contact that can be potentially associated with infection transmission.

Network analysis is a widely-used approach for investigating the properties of such contact networks. It has been used to estimate the potential size of epidemics [Kao et al., 2006], investigate the course of previous outbreaks [Ortiz-Pelaez et al., 2006], inform mathematical models of outbreaks [Dubé et al., 2009] and to assess the effect of interventions [Robinson and Christley, 2007].

The structure of live movement networks (as well as other contact networks) is not constant and can change over time due to a multitude of things (e.g. legislative restrictions to move animals, disease outbreaks, fluctuations in farmers' behaviour). Dynamics of changes in networks can also be addressed using network analysis.

4.1.1 Network terminology

The terminology of network analysis is very similar to that of graph theory. Individuals of interest are usually called nodes or vertices, and contacts between them are referred as edges or links. Edges can be directed or undirected, depending on the nature of the network. We assume that $E_{ij} = 1$, if there is a link from i to j, and $E_{ij} = 0$ otherwise.

Two nodes are connected if there is a path between them (i.e. it is possible to move from one node to the other by following one or more edges). Direction of edges might be taken into account to define a directed path (note that the fact that B is reachable from A by a directed path does not necessarily mean that A is reachable from B).

4.1.2 Network measures

Contact networks are mainly analysed through descriptive network measures. The basic ones include graph density (frequency of edge presence between any two nodes), degree distribution (distribution of number of edges for each node), clustering coefficient (frequency of triangles of edges) and average path length (average number of edges that are needed to reach one node from another).

For undirected networks, the size of the largest connected component (every node is reachable from any other) is another useful measure. Two similar definitions are used for directed networks. A weak component is a part of a network where all nodes are in contact with each other either directly or via other nodes, not taking the direction of the contact into account. A strong component is a part of the network where all nodes can reach each other through directed links, either directly or via other nodes. The giant strong component (GSC) is the largest strong component in the network. The giant weak component (GWC) is the largest weak component in the network. GWC and GSC have previously been used to estimate upper and lower bounds of the potential epidemic size, respectively [Kao et al., 2006].

More information on network measures relevant for disease control can be found in [Nöremark et al., 2011].

4.1.3 Network methods in veterinary epidemiology

Application to bacterial infections of dairy cattle

García Álvarez et al. [García Álvarez et al., 2011] used a network-based approach to test if the distributions of individual sequence types of *Staphylococcus aureus* (a bacterial pathogen that also causes mastitis in cattle) can be explained by movements and other contact networks.

Nodes for the network were dairy cattle herds from the study population (a cluster of 44 farms in Somerset, UK). Milk samples were collected from each study herd at two timepoints (May 2007 and October 2007) and then analysed using MLST to identify sequence types (ST) of *S. aureus*.

Cattle movements related to the study cluster were used to construct a network for the analysis. A link between two nodes (study farms) was added to the network if:

- One study farm bought cattle that previously had been on another study farm (type I link).
- Two (or more) study farms bought cattle that previously had been on a common location outside the study population (type II link).

Also, non-reportable local contacts collected via questionnaire (on-farm visitors, sharing agricultural equipment, etc.) were used to construct other risk-potential networks where two nodes were connected by a link if there was a particular relationship between the nodes (e.g. two farms shared a veterinary practice).

Then, the constructed networks of risk-potential linkages between the study farms were tested to explain the transmission of individual bacterial strains.

If transmission of a particular strain is indeed associated with one of the risk-potential networks (a significant proportion of new cases are connected through the observed network with herds infected at the previous surveillance period), it is expected that newly infected (NI) herds with this strain were exposed (i.e. connected) to the herds infected with the same strain more than the persistently susceptible (PS) herds considered. It is also expected that the observed network was different from the random ones regarding the exposure to infection of NI farms.

In order to test the hypothesis that risk-potential networks can be used to predict strain distributions, random networks with the same number of nodes and edges as the observed one were generated. Then, the mean exposure to infection of NI and PS farms was measured for the observed and simulated networks, to test if the observed network is different from random (i.e. can explain the distribution of the considered strain).

This study by García Álvarez et al. found that constructed movement network played a role in the transmission of two out of three most predominant STs of *S. aureus* between the dairy farms in the study population.

4.1.4 Objective

In the previous chapter it was shown that the spatial distribution of herds infected with ST103 was different from other STs, which suggests diverse transmission patterns between the bacterial strains. Here, the aim was to assess differences between strains of *Streptococcus agalactiae* regarding the relatedness of their distributions to the networks of contacts between the Danish dairy cattle herds.

Thus, the objective is to identify the role of cattle movements and other between-herd contacts in the spread of certain strains of *S. agalactiae* and to determine whether there is a difference between specific bacterial subpopulations.

4.2 Materials and methods

4.2.1 Epidemiological data

The same principal dataset as in previous chapters was used.

Based on two consecutive annual testings (t_1 and t_2) and for a particular strain of *S. agalactiae*, each dairy cattle herd that was registered for both years can be categorised into one of four groups (see Table 4.1).

Table 4.1: Definition of herd states based on bulk tank samples at two time-points t_1 and t_2.

Strain in BTM at t_1	Strain in BTM at t_2	Herd state
No	No	Persistently susceptible (PS)
Yes	No	Recovered (R)
No	Yes	Newly infected (NI)
Yes	Yes	Persistently infected (PI)

Since the information on strain types for herds that were infected before 2009 or after 2011 was not available, only states of herds in 2010 and 2011 (based on testings in 2009–2010 and 2010–2011, respectively) could be defined.

It is also possible to consider the overall bacterial population (all strains, including unknown ST0) and obtain the strain-unspecific information on how many herds of each type there were in 2010 and 2011. Although, these NI herds will be only those that were susceptible in the previous year, e.g. a herd that was previously infected with another strain will not be treated as NI (as opposed to the strain specific definition given by Table 4.1).

Strain specific definition of NI herds (based on the presence of particular ST in the BTM[1]) yields the numbers of herds given in Figure 4.1. This is not equivalent to the classification of strain-unspecific NI herds by ST they acquired.

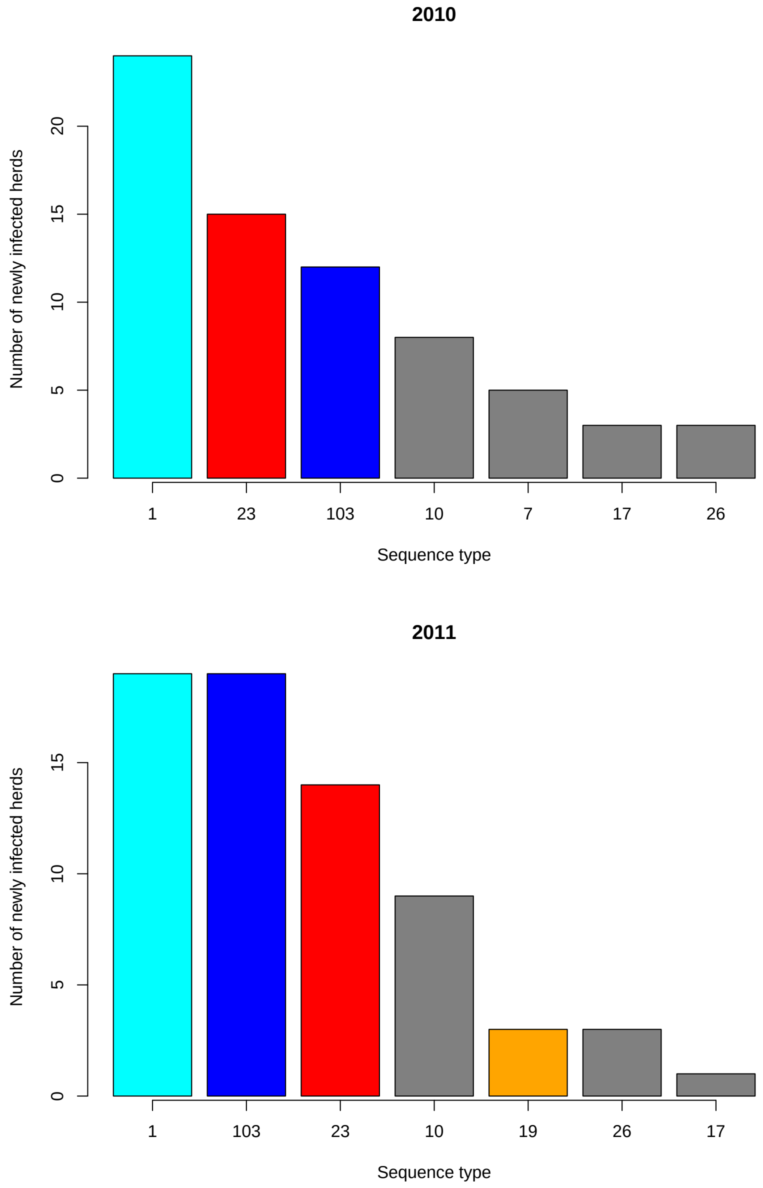

Figure 4.1: Number of newly infected herds with different sequence types in 2010 and 2011.

[1] For example, if herd was infected with ST23 previously, but obtained ST103 this year, it will be NI for ST103, although it was previously infected (but with another strain).

4.2.2 Movement data

Cattle movement data were obtained from the Central Herd Register (Danish Veterinary and Food Administration, Glostrup, Denmark) that captures all cattle movements within Denmark on a daily basis. This includes movements between all types of cattle herds: dairyand non-dairy.

Exclusion of movements that do not impose risk of infection is a common practice before network analysis. In this case, only movements recorded as "sold alive" were kept for further analysis. Intensity of farm-to-farm movements is presented in Figure 4.2.

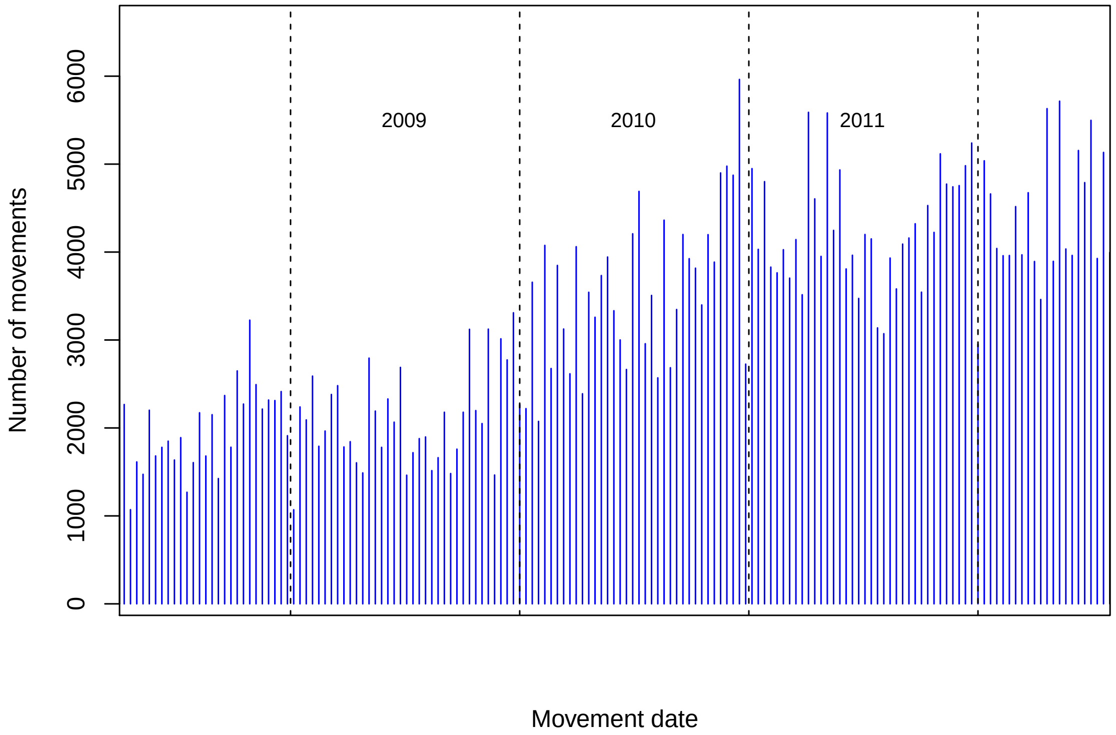

Figure 4.2: Intensity of farm-to-farm cattle movements in Denmark (sold alive).

4.2.3 Veterinary practice data

Information on affiliation of dairy cattle herds with registered veterinary practices is recorded in Denmark. Each herd has a list of veterinary practices and corresponding start and end dates. Data on veterinary practices were used to construct a contact network where a link connects to nodes if the corresponding herds shared the same veterinary practice during the time period of interest.

Although veterinarians can potentially play a role in transmission of *S. agalactiae*, this contact network was primarily considered as a marker of other important connections between dairy herds such as sharing farming equipment, pastures, milk hauliers, etc. that were assumed to coincide with sharing of veterinary practice [Olofsson et al., 2014].

4.2.4 Analysis method

Only registered dairy cattle herds (i.e. those that deliver milk to the factory) were of interest in this study. However, for movement network analysis, it was also important to incorporate the information about all the cattle that moved between two dairy herds via other agricultural holdings to account for possible indirect transmissions.

The aim was to build networks with nodes that represent registered dairy herds and:

- Directed links that represent if there were any cattle moved from one herd to another regardless of the exact route (whether it was a direct farm-to-farm movement or via intermediate nodes, e.g. markets).
- Undirected links that connect herds that shared veterinary practice at some point during the study period.

Movement network construction

Here, we are focused on data that can potentially be associated with disease transmission routes. In this case, it is the number of cows that were in one herd and then appeared in another herd (potentially travelling via intermediate dairy or non-dairy farms, markets, etc.) within a time period of interest.

Dates of first PCR samples in each year (19 October 2009, 18 October 2010, 30 August 2011) determined two time periods used for construction of movement networks. The networks for 2009–2010 and 2010–2011 were defined by the following rule: we add an edge to the network if during the considered period there are records that a cow left one herd (start node) and entered another one (end node) with the additional condition that these events should be ordered chronologically. Thus, we capture potential transfer of bacteria from one herd to another taking into account cases when a cow could have visited other intermediate agricultural premises.

It was assumed that the chance that a cow gets infected with *Streptococcus agalactiae* outside a dairy herd (e.g. when on a market or communal pasture) is negligible.

Veterinary practice network construction

The veterinary network has the same nodes as the movement network — all the registered dairy cattle herds. The same time periods were used to define two networks for 2009–2010 and 2010–2011: a link connects two herds if they shared veterinary practice at some point during the specified period. Thus, the networks were undirected, as opposed to the movement networks.

Network analysis

The constructed networks were used to address the following questions:

- Can distributions of any of the *S. agalactiae* subpopulations be explained by the structure of the contact networks?
- Are there any consistent patterns for the *S. agalactiae* strains that are associated with either contact network?

The overall hypothesis behind this was that some strains are mainly transmitted by cattle movements while others spread via other routes.

The approach by García Álvarez et al. [García Álvarez et al., 2011] was adopted to test this hypothesis.

Let E_{ij} represent the pairwise exposure within the observed contact network (i.e. E_{ij} = 1 if there is a link from i to j, and i was infected at t_1; otherwise E_{ij} = 0). Then, the mean exposure to infection of NI farms is given by:

$$E_{NI} = \frac{\sum\limits_{j=1}^{n} \sum\limits_{i=1, i\neq j}^{n} E_{ij}\delta_j}{\sum\limits_{j=1}^{n} \delta_j}, \tag{4.1}$$

where $\delta_j = 1$ if farm j is NI at t_2, otherwise $\delta_j = 0$.

The numerator is the total “amount” of potential infections received by NI herds via network edges: it is the number of all edges from infected at t_1 farms to NI farms at t_2. The denominator is the number of NI farms at t_2.

Similarly, the mean exposure to infection of PS farms (E_{PS}) is:

$$E_{PS} = \frac{\sum\limits_{j=1}^{n} \sum\limits_{i=1, i\neq j}^{n} E_{ij}\gamma_j}{\sum\limits_{j=1}^{n} \gamma_j}, \tag{4.2}$$

where $\gamma_j = 1$ if farm j is PS at t_2, otherwise $\gamma_j = 0$.

The next step is to simulate random networks. In the original paper, the number of nodes and edges stayed the same but edges were shuffled within the network. The states of nodes were also kept the same. However this approach generates perfectly random networks, it would be useful to preserve some other graph properties (particularly, degree distribution) so that the random sample reflects the structural traits of the observed network (i.e. generated random networks are similar to the real-world network). For this purpose, the method to generate random networks was based on re-shuffling of nodes rather than edges, which obviously allowed for keeping the same degree distribution.

The criteria for the significance of a particular network was motivated by several observations. If a strain is indeed transmitted from infected farms via contacts in the network, it would be expected that the study farms that are newly infected at t_2 were exposed to more infection than those that remained susceptible (i.e. $E_{NI} > E_{PS}$). In addition, NI farms at t_2 would be expected to be more exposed to infection in the observed than in the random networks. Similarly, the PS study farms at t_2 should be less exposed to infection in the observed than in the random networks.

In contrast with approach by García Álvarez et al., here type II links (i.e. those connecting study farms that purchased cattle from a common location outside the study population) were not considered, because the study population included the total dairy population. Plus the assumption that cattle was unlikely to get infected on premises different from registered dairy farms.

Outputs of the method by García Álvarez et al. [García Álvarez et al., 2011] are presented in the form of two-dimensional plot (mean exposure to infection of NI and PS herds as x- and y-axis, respectively). Simulated random networks usually form a cluster of points on the plot (around the mean value of mean exposures to infection of NI and PS herds). The decision to accept or reject the initial hypothesis depends on the position of the point representing the observed network.

If the contact network explains the distribution of cases (or a certain strain) we expect the point representing the observed network:

1. To be below the $y = x$ line (for the observed network, mean exposure to infection of NI is expected to be higher than the one of PS herds).

2. To be outside the cluster of points for randomly generated networks (the observed network is expected to be different from the simulated).

3. To be below the cluster (mean exposure to infection of NI herds is expected to be higher for the observed network than for the simulated).

4. To be to the right of the cluster (mean exposure to infection of PS herds is expected to be lower for the observed network than for the simulated).

Formally, the null hypothesis was that a contact network of interest is not different from random networks with the same properties (in this case, the same number of nodes and edges) in explaining distribution of the considered bacterial subpopulation (determined by the two-dimensional plot of mean exposures to infection of NI and PS herds). Thus, if the point for the observed contact network is outside the cloud of the random networks and satisfies all of the expected conditions, we can reject the null hypothesis and, thus, report significant association between the observed network and the distribution of the considered bacterial subpopulation.

4.3 Results

4.3.1 Characteristics of contact networks

Cattle movement networks

Two networks for 2009–2010 and 2010–2011 summarised movements of cattle between dairy herds for the specified time periods (between starting dates of the annual national surveillance in 2009–2011). The network measures (Table 4.2) were calculated using **igraph** R package [Csárdi and Nepusz, 2006]. Not all the nodes (defined as herds registered in both considered years) were active, i.e. moved animals to or from other dairy herds (allowing for several intermediate movements).

The degree distribution is presented in Figure 4.3 and indicates scale-free properties (power law degree distribution) of the observed networks.

The movement networks were relatively sparse but had low average path length within connected components.

Veterinary practice networks

The same measures are presented for the veterinary networks (Table 4.3). Degree distributions are presented in Figure 4.4.

Table 4.2: Characteristics of cattle movement networks for 2009–2010 and 2010–2011. Herds registered in both years were treated as nodes.

	2009–2010	2010–2011
Herds registered in either year	4284	4114
Nodes (herds registered in both years)	4065	3895
Active (connected) nodes	2579	2203
Edges	4674	3420
Giant weak component size	2341	1940
Giant strong component size	25	9
Average path length	4.65	4.11
Diameter	14	15
Clustering coefficient	0.038	0.036
Density	7.02×10^{-4}	7.05×10^{-4}

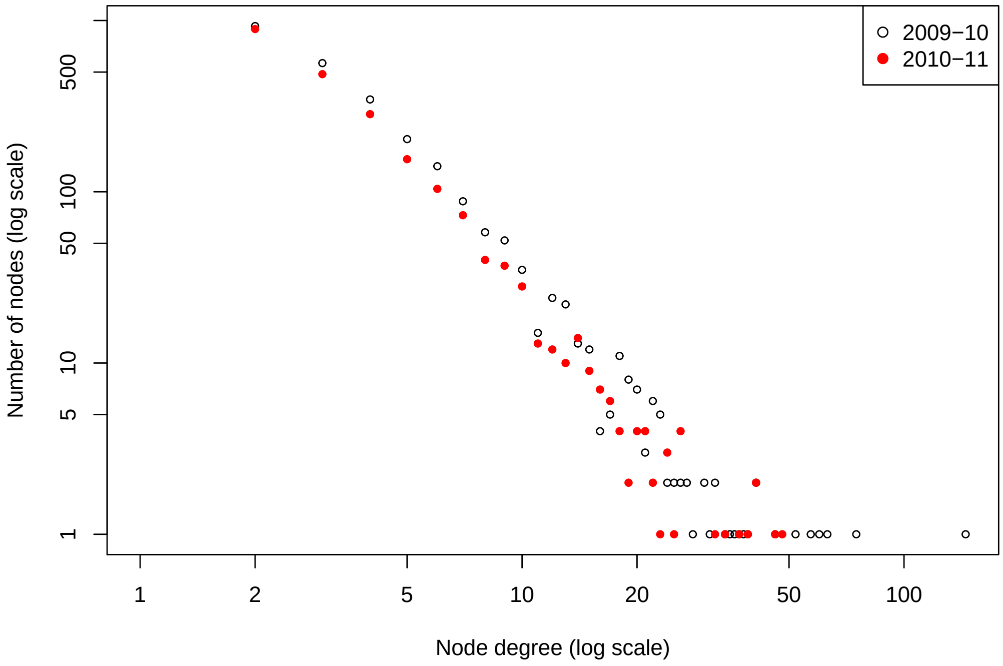

Figure 4.3: Degree distributions of the observed movement networks (log scale).

Table 4.3: Characteristics of veterinary networks for 2009–2010 and 2010–2011. All the nodes were active, i.e. had at least one edge.

	2009–2010	2010–2011
Nodes	4041	3859
Edges	304198	281354
Giant connected component size	2943	2164
Average path length	8.05	4.89
Diameter	17	12
Clustering coefficient	0.981	0.987
Density	3.72×10^{-2}	3.78×10^{-2}

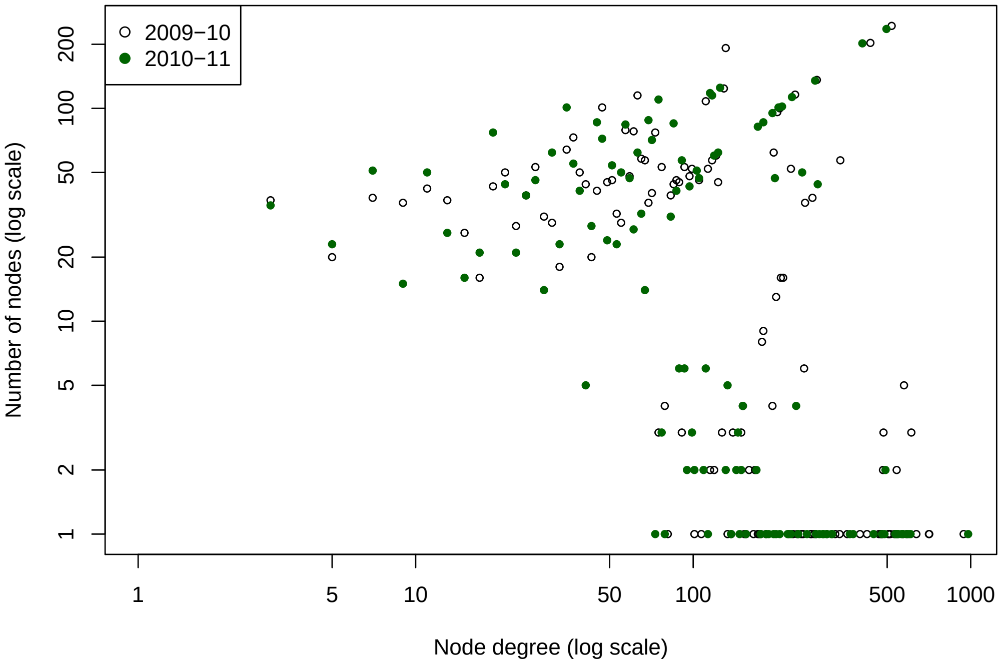

Figure 4.4: Degree distributions of the observed veterinary practice networks (log scale).

4.3.2 Distribution of particular sequence types among Danish dairy herds

Here, the addressed question was whether distributions of any of the individual STs present in infected dairy herds can be explained by either of the considered contact networks. Only frequent STs that were present in NIs for both years 2010 and 2011 (ST1, ST2, S23, ST103 and ST314) were used (see Figure 4.1).

Cattle movement networks

Even for frequent STs among NI herds there were little incoming cattle movements from other herds infected with the same considered ST (see Table 4.4).

Table 4.4: Number of risk potential contacts in movement networks for herds newly infected with predominant sequence types.

2010

ST	Contacts with infected herds	Number of NI herds	Mean exposure
1	1	16	0.0625
2	1	6	0.1666
23	1	10	0.1
103	0	6	0
314	0	5	0

2011

ST	Contacts with infected herds	Number of NI herds	Mean exposure
1	0	15	0
2	1	2	0.5
23	1	12	0.0833
103	3	14	0.2143
314	0	3	0

1000 random networks (with the same nodes and the same number of edges) were generated by reassigning edges to test if they result in different mean exposures for NI and PS herds.

Individual outputs for STs can be found in Figures A.6 and A.7. Graphs are less informative for more infrequent STs.

In Figure 4.5, the summary of the network analysis is presented. For each of the outputs of the method by García Álvarez et al. [García Álvarez et al., 2011], the difference between the observed “mean exposure to infection of NI/PS herds” and the average “mean exposure to infection of NI and PS herds” for simulated random networks was computed in standard deviations (SD) of simulated outputs. Thus, outputs for different STs could be compared against each other.

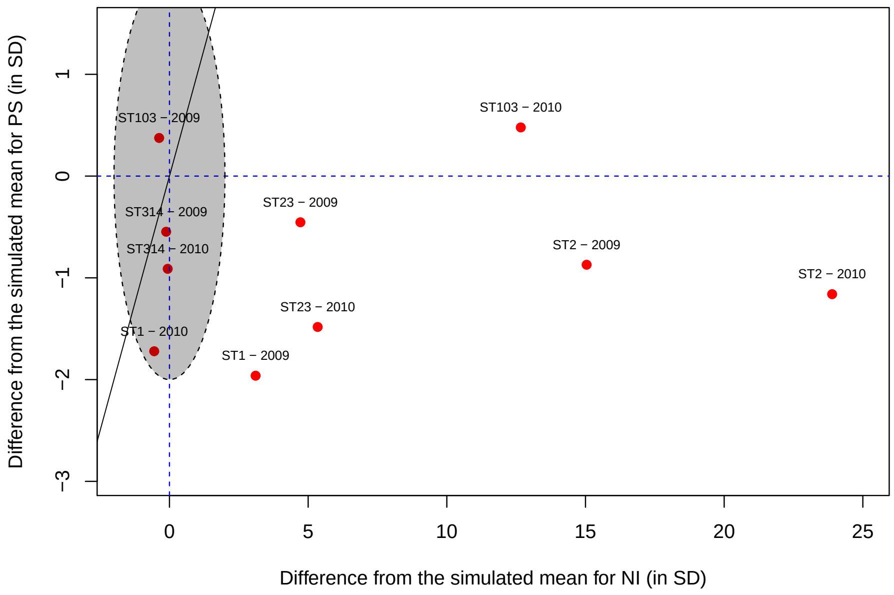

Figure 4.5: Summary of the sequence type specific network analysis for the movement networks: differences in standard deviations (SDs) for NI and PS mean exposure to infection of the observed and the mean of the simulated networks. Solid line indicates the equality of the mean exposures, dotted lines show their mean values for the simulated networks, and the ellipse is used to show the position of the simulated cloud of random networks.

Veterinary practice networks

For the veterinary practice networks, there were more risk potential links that could lead to infection of NI herds (see Table 4.5).

Individual graphs with mean exposures to infection for major STs in the networks of veterinary practices can be found in Figures A.8 and A.9.

The summary of outputs for considered STs is presented in Figure 4.6.

4.3.3 Distribution of clonal complexes among Danish dairy herds

The same analysis was performed on CC level: strains defined by clonal complexes were used to assess the role of contact networks in their distributions. In Figure 4.7, numbers of NI herds for CCs are presented.

There were no newly infected herds with strains from CC19 in 2009 and from CC7 in 2010, so they were not included in further analysis.

Table 4.5: Number of risk potential contacts in veterinary networks for herds newly infected with predominant sequence types.

2010

ST	Contacts with infected herds	Number of NI herds	Mean exposure
1	19	16	1.1875
2	1	6	0.1666
23	5	10	0.5
103	8	6	1.3333
314	0	5	0

2011

ST	Contacts with infected herds	Number of NI herds	Mean exposure
1	8	15	0
2	1	2	0.5
23	11	12	0.9166
103	10	14	0.7143
314	3	3	1.0

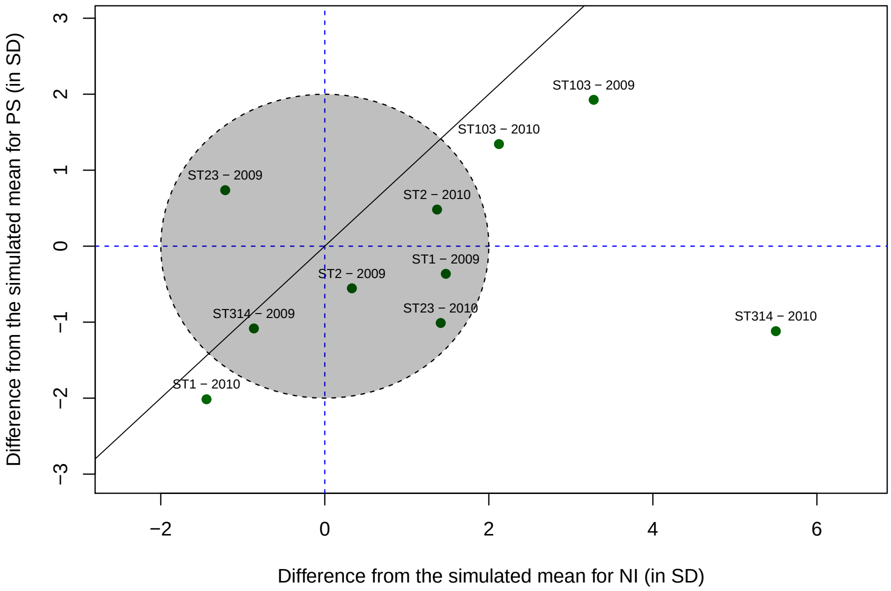

Figure 4.6: Summary of the sequence type specific network analysis for the veterinary practices networks: differences in standard deviations (SDs) for NI and PS mean exposure to infection of the observed and the mean of the simulated networks.

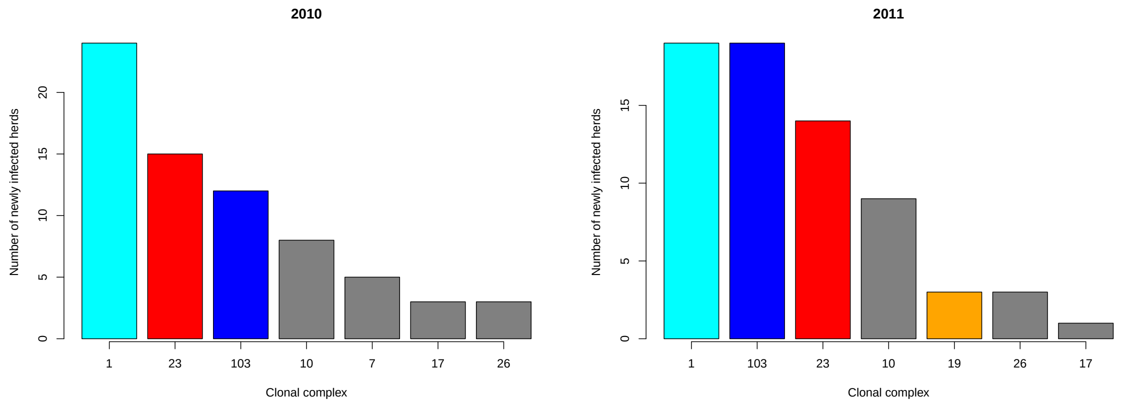

Figure 4.7: Number of newly infected herds with strains from different clonal complexes in 2010 and 2011.

Cattle movement networks

The number of incoming risk potential links due to movements for CC level NI herds can be found in Table 4.6). Despite the softer criteria of group definitions (CC is broader than ST), the number of contacts with previously infected herds was still low: only four CCs (1, 10, 23, 103) had them.

The summary for risk-potential analysis on CC level for the cattle movement networks is presented in Figure 4.8. Individual graphs can be found in Figures A.10 and A.11.

Veterinary practice networks

For the veterinary practice networks, (just as for ST level analysis) there were more risk potential links that could lead to infection of NI herds (see Table 4.7).

The summary for risk-potential analysis on CC level for the veterinary networks is presented in Figure 4.9. Individual graphs can be found in Figures A.12 and A.13.

4.4 Discussion

Two types of contact networks were assessed in this chapter: cattle movement networks summarised all the risk-potential linkages via cattle relocations from one herd to another during the yearly periods, while veterinary networks represented sharing of veterinary practices (and potentially other resources that can be attributed to the risk of infection). These two types of networks were different: directed weighted and quite sparse, and undirected unweighted and relatively dense, respectively.

Table 4.6: Number of risk potential contacts in movement networks for herds newly infected with predominant clonal complexes.

2010

CC	Contacts with infected herds	Number of NI herds	Mean exposure
1	3	24	0.125
7	0	5	0
10	1	8	0.125
17	0	3	0
19	0	0	0
23	1	15	0.0667
26	0	3	0
103	0	12	0

2011

CC	Contacts with infected herds	Number of NI herds	Mean exposure
1	1	19	0.0526
7	0	0	0
10	0	9	0
17	0	1	0
19	0	3	0
23	1	14	0.0714
26	0	3	0
103	4	19	0.2105

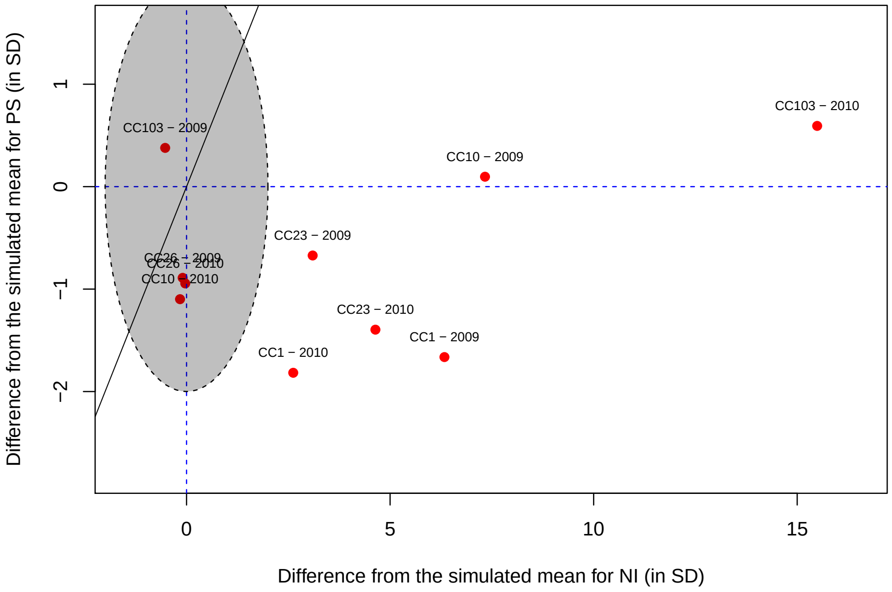

Figure 4.8: Summary of the clonal complex level network analysis for the cattle movement networks: differences in standard deviations (SDs) for NI and PS mean exposure to infection of the observed and the mean of the simulated networks.

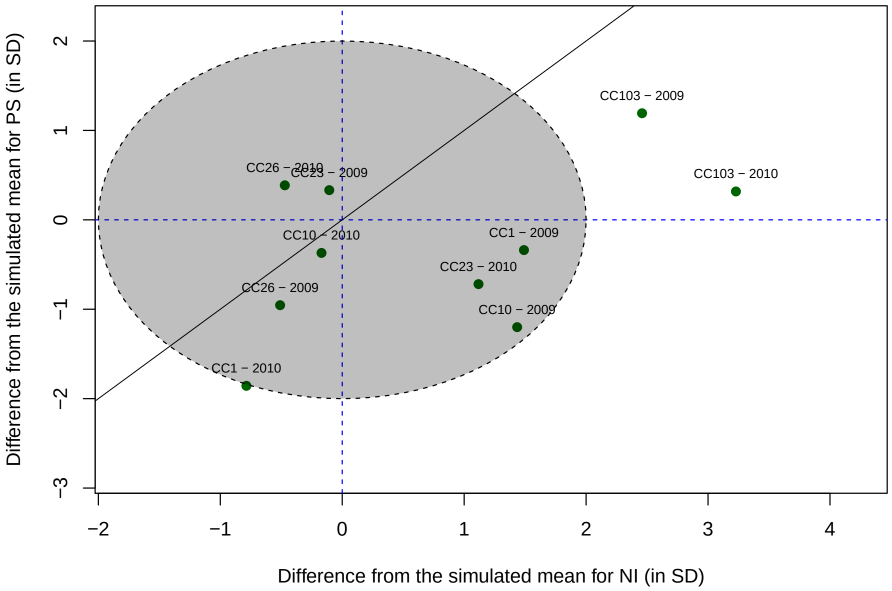

Figure 4.9: Summary of the clonal complex level network analysis for the veterinary practices networks: differences in standard deviations (SDs) for NI and PS mean exposure to infection of the observed and the mean of the simulated networks.

Table 4.7: Number of risk potential contacts in veterinary networks for herds newly infected with predominant clonal complexes.

2010

CC	Contacts with infected herds	Number of NI herds	Mean exposure
1	37	24	1.5417
7	0	5	0
10	4	8	0.5
17	0	3	0
19	0	0	0
23	13	15	0.8667
26	0	3	0
103	15	12	1.25

2011

CC	Contacts with infected herds	Number of NI herds	Mean exposure
1	18	19	0.9474
7	0	0	0
10	2	9	0.2222
17	0	1	0
19	1	3	0.3333
23	15	14	1.0714
26	0	3	0
103	22	19	1.1579

4.4.1 Network characteristics

The constructed networks of cattle movements for 2009–2010 and 2010–2011 were not dense. In contrast, the veterinary networks were quite dense but had no scale-free properties.

As a consequence, more NI herds had were connected with previously infected herds in the veterinary networks (see Tables 4.4 and 4.5). However, this was taken into account for the simulated networks (they had the same number of edges) when risk-potential links were assessed.

Limited number of potentially infective contacts due to cattle movements might be the indication that legislative restrictions to sell cattle infected with *S.agalactiae* are effective in decreasing the dissemination of the pathogen.

4.4.2 *Streptococcus agalactiae* distribution and cattle movements

ST level network analysis was restricted to major STs: less frequent strains result in zero mean exposure to infection of NI herds in the observed network (there are no potentially infective links). Distributions of ST2 and ST23 showed significant association with cattle movements (consistent for both periods, 2009–10 and 2010–11) ST1 in 2009–10 and ST103 in 2010–11 also showed some evidence that their distributions could be explained by cattle movements.

On CC level, CC1 and CC23 were significantly associated with cattle movements, which is in agreement with ST level analysis. Results for CC10 and CC103 were not consistent for the two periods. Other strains defined by CCs showed that their distributions were equally dependent on the observed cattle movement networks and the randomly simulated, suggesting that their transmission mechanisms could not be associated by cattle movements.

4.4.3 *Streptococcus agalactiae* distribution and veterinary networks

For the veterinary networks, only ST103 showed significant results that were consistent for two years. Also, the distribution of ST314 in 2010–11 could be associated with the veterinary affiliation network.

Similarly, the considered clonal complexes showed no significant associations between their distributions and the observed veterinary networks, apart from CC103.

4.4.4 Comparison with spatial clustering results

Consistent and significant spatial clustering of ST103 reported in previous chapter could be attributed to specific transmission patterns. The network analysis showed that the veterinary networks could explain the transmission of ST103 throughout the study period, but the movement networks were associated with ST103 distributions only for the second time period 2010-11. This might suggest that ST103 was introduced to danish dairy cattle population just recently and spreads primarily through veterinary networks. This means that its transmission mechanisms have more complex nature as it was previously expected, and they should be investigated further to establish the main driver of the transmission (on-farm visitors, farming equipment, etc).

4.4.5 Possible extensions

Original method [García Álvarez et al., 2011] was extended here by using a different sampling technique to generate random networks. Instead of changing edges, we changed nodes and, thus, kept the same degree distribution, which is an important characteristic of complex animal movement networks.

Other potential improvements of the method are discussed further.

Relaxation of assumptions

It was assumed that the chance that a cow gets infected with *S. agalactiae* outside a dairy herd (e.g. when on a market or communal pasture) is negligible. This is a very strong assumption and might be relaxed in further studies. In this case, links should be added between herds that received cattle that could have got infected on some premises different from registered dairy farms (analogue of type II links from [García Álvarez et al., 2011]).

Weighted edges

Number of cattle moved can be used in calculation of mean exposure, e.g. E_{ij} can become the number of cattle that arrived from farms infected with the same strain. This would provide a better estimate of infection risk potential, but at the same time would complicate the generation of random networks (to simulate the random weighted graphs is not straightforward).

More precise dates

Dates of samplings for first annual PCR tests in each of the study years determined two time periods for aggregation of movement data (and also veterinary data). An alternative approach could be to take into account date of PCR sampling of each herd. In this case we include edge between herds A and B if there was a movement of cattle during the time period specific for this pair of herds (annual sample date for A in previous year and annual sample date for B in this year).

Chapter 5

Estimation of possible transmission trees: application to the Darlington cluster within the 2001 FMD epidemic in the UK

5.1 Introduction

In this chapter, I observe the existing methods to reconstruct transmission trees based on epidemiological data, and then introduce an alternative approach that can extend one of the previous methods to account for temporal constraints between transmission links within a transmission tree.

5.1.1 Transmission trees: background

A transmission tree is a directed tree[1] that represents directions of disease spread between individuals during an outbreak. It is one of the key concepts in epidemiology, which can be used to gain greater insights into mechanisms of pathogen transmission [Spada et al., 2004], estimate epidemiological parameters [Haydon et al., 2003] and inform intervention strategies. One of the major obstacles blocking more extensive use of epidemic histories in epidemiological analysis is the difficulty of reconstructing transmission trees in the absence of contact tracing data. Nevertheless, other types of data can be used to estimate transmission routes of a pathogen.

[1] Assuming that there is only one index case within the study population. Otherwise, a set of transmission trees, one for each index case, should be used instead.

Typically, an outbreak of a certain disease affecting farm animals provides us with data on locations of farms, reporting and culling dates, and, occasionally, genetic samples of the pathogen. Estimations of possible transmission trees based on different types of data and assumptions can be inconsistent with each other. Combination of data types used could potentially tackle this issue.

5.1.2 Previous work on reconstruction of transmission trees

Reconstruction of transmission trees is straight-forward when reliable contact tracing data are known. However, in most cases, estimation of transmission routes is based on other available data sources.

Historically, detailed epidemiological data can be used to deduce 'who infected whom?' pathways. While this can sometimes be done with great success, it is almost inevitably laborious and resource hungry, and extremely difficult to do on a mass scale. In contrast, epidemiological models have also been used to understand contact processes that generate epidemics but these are usually used to fit broad, population level parameters and are not intended to exactly reconstruct particular transmission trees.

In one effort that partially bridged the gap between these two paradigms, Haydon et al. proposed an heuristic algorithm for construction of possible epidemic trees [Haydon et al., 2003] to assess R_0 and control policies during the 2001 foot-and-mouth disease (FMD) epidemic in the UK. The algorithm generated a putative source of infection for every infected farm from contact-tracing data or chosen from a list of 'candidates' based on reporting dates and spatial proximity (when contact data were unavailable). Several trees were then used to estimate case-reproductive values and to model alternative control scenarios. The estimated values were consistent with those estimated previously by fitting dynamical models [Ferguson et al., 2001a].

Since 2001, the rapid development of mass scale next generation sequencing has provided the opportunity to use mutations in the pathogen itself as a means of tracking an epidemic from farm to farm. Cottam et al. [Cottam et al., 2008] observed a relatively small cluster of infected farms in Durham county during the same 2001 FMD epidemic. The authors did not consider proximity of the farms but used genetic samples of the virus to restrict the possible tree structures. Cottam et al. enumerated all the possible transmission trees that were consistent with the collected genetic data and then assessed the likelihood of these trees. The addition of the epidemiological information to the genetic data allowed for a substantial reduction in the number of transmission trees supported by the data, with only four trees accounting for 95% of the sum of the individual tree likelihoods. Also, the authors examined the next most likely sources of infection for each farm, and the most likely tree was confirmed to be robust as it appeared that for most cases an alternative candidate source was at least 80 times less likely.

The most common structures to deal with genetic data are phylogenetic trees that represent estimated evolutionary history of the pathogen. However, phylogenetic trees cannot be directly used as transmission trees given that the direction of pathogen evolution does not necessarily coincide with the direction of the pathogen's spread between individuals [Pybus and Rambaut, 2009].

Jombart et al. [Jombart et al., 2011] proposed a graph approach which is conceptually different from estimation of phylogenetic trees and is promising as it aims to reconstruct transmission trees without identifying unobserved ancestral genotypes. Genetic sequences and collection dates were used to find ancestors directly from the sampled isolates (rather than attempting to reconstruct unobserved and hypothetical ancestral genotypes). Each observed isolate has one and only one ancestor that always precede their descendants in time. Among all possible ancestries of a given isolate, some are more likely than others, and this likelihood can be inferred from the genetic divergence between isolates. Having taken into account these observations, the authors proposed the SeqTrack algorithm [Jombart et al., 2011] that assigns weights to each possible ancestry and then finds the spanning tree that maximises the overall weight.

It is likely that using several data types will provide more robust trees than any individual data type alone. To this end, Ypma et al. [Ypma et al., 2012] proposed that for integration of several types of data one can use a likelihood function that combines individual likelihoods based on each type of data. This idea was used to implement a Bayesian framework to reconstruct transmission trees for an epidemic of avian influenza in poultry farms in The Netherlands in 2003. The authors combined in one likelihood function temporal, spatial and genetic data (assuming that all three types of data are independent):

$$L = L_{temporal} \cdot L_{spatial} \cdot L_{genetic}$$

Markov Chain Monte Carlo (MCMC) simulations were performed to sample from the space of all transmission trees and parameters of the likelihood function (allowing the simultaneous estimation of both). The likelihood of a tree was defined as a product of the likelihoods of its links and the result was presented as a weighted average over the set of possible transmission trees. The combined approach estimates the transmission tree with higher confidence and resolution than analyses based on genetic and epidemiological data alone. Furthermore, the approach can handle missing genetic data — if a farm was not sampled, the ancestor's sequence is used. However, the assumption that the three types of data are independent should be relaxed in further studies as epidemiological and genetic data are evidently correlated. Also, the authors mentioned that more sophisticated evolutionary model (that depends on time and infection events) would be a desirable improvement.

Several Bayesian frameworks [Morelli et al., 2012], [Jombart et al., 2014] have been implemented recently to combine epidemiological and genetic data to reconstruct most likely transmission patterns and infection dates. Morelli et al. relaxed the assumption of independence between the information sources and defined the likelihood of a transmission tree as a joint likelihood that depends on both spatiotemporal and genetic data. The results included posterior probabilities of pairwise transmissions and inferred infection dates. Several simulated and real datasets were used against this approach, including the data from [Cottam et al., 2008] on the Darlington cluster within the 2001 UK FMD epidemic. More coherent use of genetic data and more advanced scheme to obtain the posterior distribution of unknown parameters (transmission tree and infection dates) improved the previous results by Cottam et al., suggesting higher role of a hub in the FMD transmission (farm K became infectious very early in the outbreak and was estimated to have seeded the infection to other premises) and the likely incompleteness of the dataset (the presence of unobserved premises that could be part of the transmission chains) indicated by the unrealistically long latency periods.

In [Jombart et al., 2014], a similar Bayesian approach with the joint likelihood was implemented and helped to identify likely transmission events, infection dates, unobserved cases and separate introductions of infection. One of the main differences was that the distribution of the generation time was used to define the epidemiological likelihood of a transmission tree (in contrast with [Morelli et al., 2012], where empirically characterised distributions of infectiousness were used).

5.1.3 Epidemiological context: the 2001 foot-and-mouth disease epidemic in the UK

In this chapter, I consider the spread of FMD — one of the most contagious and economically important animal diseases worldwide [Bachrach, 1968, Haydon et al., 2004]. It has low associated mortality but can decrease livestock productivity and therefore represents a serious threat to farming. Countries with circulating FMD in their livestock populations are subject to considerable international trade restrictions on animals and animal products.

The causative agent of FMD is a single stranded RNA virus that has a genome 8500 bases long [Domingo et al., 2002] is subject to rapid evolution due to its high mutation rate [Domingo et al., 2003]. The rapid change in the virus makes the use of genetic information a marker of disease transmission feasible even over short timeframes.

In 2001 an FMD epidemic occurred in the UK. The first cases were confirmed in an abattoir in Essex in February, followed by an epidemic that lasted for seven months with 50 newly infected premises per day at its peak in late March [Gibbens et al., 2001]. Sequentially applied culling policies seriously affected populations of cloven-hoofed animals (including sheep, cattle, pigs and others): more than 6.5 million animals were culled on 9900 premises (2030 of them were infected) [Haydon et al., 2004, Kao, 2002], with an estimated cost of £3–5 billion to the national economy.

An unprecedentedly large and detailed dataset collected for the 2001 FMD epidemic in the UK enabled epidemiologists to create number of epidemiological models and investigate their applicability. Despite the significant improvements in the understanding of the underlying epidemiological processes, there remain many issues in developing more complex methods of analysing disease outbreaks and predicting its consequences.

5.1.4 Independence of individual links of transmission tree

In [Cottam et al., 2008] possible transmission links between farms were considered to be independent of each other. The likelihood of a tree was defined as a product of the likelihoods λ_{ij} of its links. This independence assumption is quite common, e.g. in [Wallinga and Teunis, 2004] Wallinga et al. considered pairs of cases rather than the entire infection network to obtain likelihood-based estimates of effective reproduction number R during SARS epidemics in Hong Kong, Vietnam, Singapore and Canada.

In reality, the likelihood of a link depends on previous events. Let us consider a transmission chain for three farms: $A \to B \to C$. The likelihood of infection from A to B (λ_{AB}) is computed by marginalising over all possible days of infections regardless the timing of $B \to C$ transmission and vice versa. Therefore, the product of λ_{AB} and λ_{BC} will be the probability of all infection dates combinations. In other words, cases when B was infected after C are taken into account, although these are impossible cases and they should be excluded when computing the probability of the overall chain $A \to B \to C$. Ultimately, if there is a substantial overlap in link probabilities over time, conditional probabilities should be used to marginalise over the particulars of $A \to B$ when computing the probability of $A \to B \to C$.

To come up with an approach of the transmission tree estimation that takes into account mutual dependence of individual links, one should consider timings of infection transmissions along with their order. Thus, an epidemic scenario (a transmission tree with associated dates of infection) is considered here.

In the next section I discuss the assumptions made by Cottam et al. in [Cottam et al., 2008] and propose an extension to their approach for estimation of possible transmission pathways.

5.2 Materials and methods

5.2.1 Data

In this study, the data from the well-studied Darlington cluster within the 2001 FMD epidemic were used. This group of infected premises was relatively isolated from other infections and was believed to have a monophyletic origin, which made it suitable for transmission tree analysis [Cottam et al., 2008, Morelli et al., 2012]. The cluster contained 15 farms that were infected in Durham county in April–June 2001 after the national movement ban (implemented in February 2001).

Following the work by Morelli et al. [Morelli et al., 2012], farms A and N were removed from the dataset as they were confirmed to be a distinct branch of transmission from the rest of the cluster. However, it was decided to keep farm B as it was within the spatial cluster of other farms. Premise K was assumed to be the only index case. The first day of the outbreak in the study population was assumed to be the most likely day of infection of common source of A and K: examination date (24 February 2001) minus lesion age (4 days) minus mean incubation period (5 days), i.e. 15 February 2001.

The estimated date of infection for each farm was defined as the examination date, minus the age of the oldest lesion, minus the mean incubation period (5 days). The latest possible date of infection was the examination date minus 2 days, to allow for a minimum incubation period. These two dates (see Table 5.1) were used to define probabilities of infection events over time.

Table 5.1: Epidemiological data for the 13 farms used in this analysis with the most likely estimated date of infection and the latest possible infection date. The dates are given since the first day of the outbreak (15 February 2001).

Farm	Examination	Cull date	Lesions	Estimated infection	Latest
K	46	49	1	40	44
B	47	48	2	40	45
L	55	55	2	48	53
O	59	60	2	52	57
C	65	65	3	57	63
E	68	69	7	56	66
F	68	69	4	59	66
P	75	76	4	66	73
D	86	86	1	80	84
G	89	89	4	80	87
M	92	96	5	82	90
I	103	103	3	95	101
J	109	110	2	102	107

For each of the 13 farms the available data consisted of geographical location and three dates: the day when a farm was reported to be infectious, the day when all the animals on a farm were culled and the most likely infection date (the time at which clinical disease began on a farm estimated from lesion dating performed by the experts onsite).

5.2.2 Likelihood estimates of transmission trees

Starting with replicating the maximum-likelihood approach from [Cottam et al., 2008] to estimate transmission trees based on reporting dates only, later in this chapter it will be discussed how this approach can be extended to account for timing constraints between the potential links.

In [Cottam et al., 2008] genetic and temporal data were used sequentially to identify transmission pathways of FMD virus. Firstly, the set of all the possible transmission trees was substantially reduced to 1728 trees that were consistent[2] with the genetic data. Then, Cottam et al. used a maximum-likelihood approach to assess the rest of the probable transmission trees. It was assumed that farms only have a single source of infection and are possible sources of infection immediately after the incubation period and up to (and including) the day the last animal on the farm was culled. The likelihood of a transmission tree was assumed to be the product of the likelihoods of its links, which were calculated using two auxiliary functions: L_k is the probability that the first infected individual on a given farm incubates the virus for k days prior to becoming infectious (i.e. distribution of incubation periods), and $I_i(t)$ is the probability that farm i was first infected at day t.

The probability $L(k)$ (see Figure 5.1) that the first infected individual on a farm incubates virus for k days prior to becoming infectious was assumed to follow a discrete form of the gamma distribution with scale and shape parameters of 3 and 1.67, respectively. This was in agreement with the previous estimates [Gibbens and Wilesmith, 2002].

Function $I_i(t)$ described the probability that farm i was infected at time t (see Figure 5.2). It had a form of the beta distribution: scale parameter was 10 and shape parameter was chosen to make the mode the same as the most likely date of infection.

Functions $L(k)$ and $I_i(t)$ can be used to obtain $F_i(t)$ — the probability that farm i is a source of infection on day t:

$$F_i(t) = \begin{cases} \sum\limits_{\tau=0}^{t} \left(I_i(\tau) \cdot \left(\sum\limits_{k=1}^{t-\tau} L(k) \right) \right) & t \leq C_i \\ 0 & t > C_i \end{cases} \tag{5.1}$$

[2] The authors built the maximum parsimony tree using TCS software package that inferred history of nucleotide changes between isolates from the farms. This made it possible to enumerate all the feasible transmission trees that were in agreement with the genetic data.

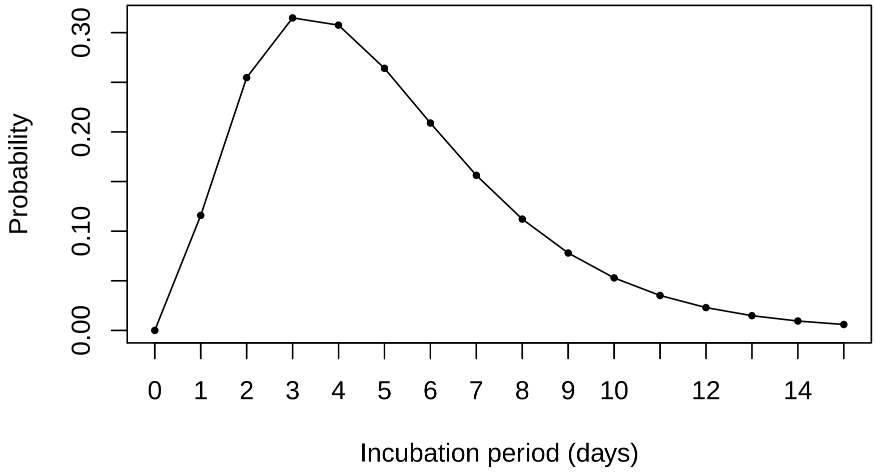

Figure 5.1: The distribution of incubation periods used in the analysis. The mean incubation period was assumed to be 5 days.

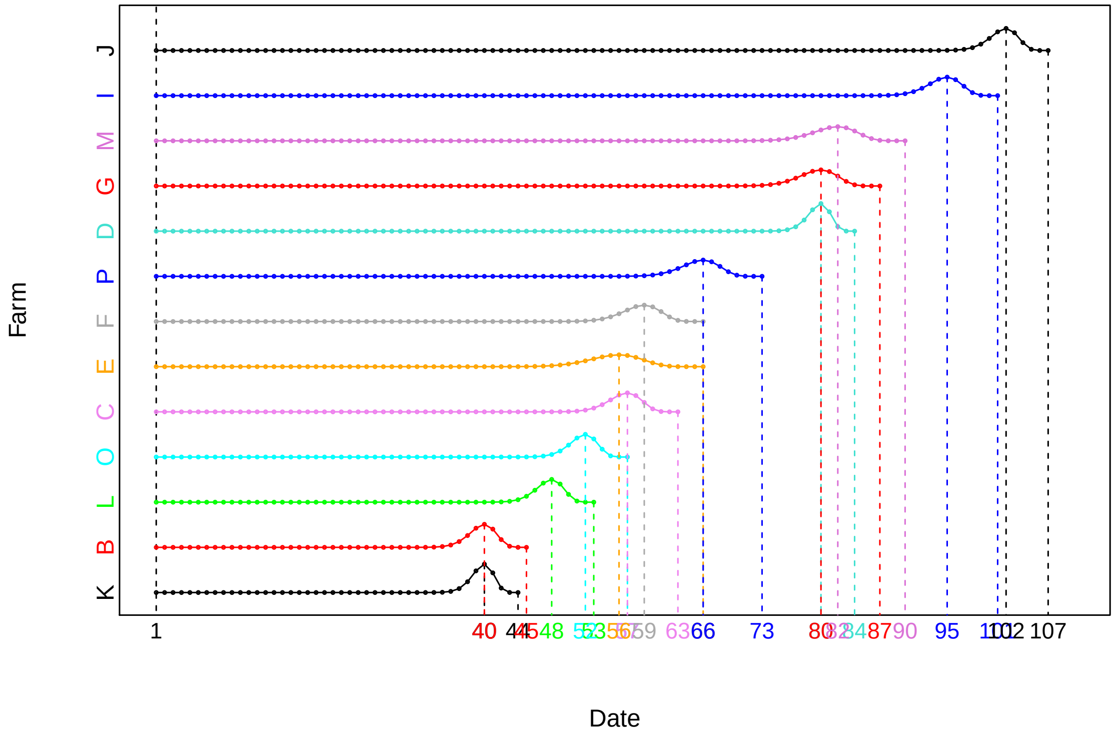

Figure 5.2: Probability profiles of infection over time for the study farms (calculated as in [Cottam et al., 2008]). For every farm, the most likely days of infection and the culling dates are indicated using dashed lines.

where C_i is the day at which farm i was culled. $F_i(t)$ (see Figure 5.3) increases towards one (which means that any farm will eventually become infectious), but drops down to zero after the day when all the animals on the farm are slaughtered.

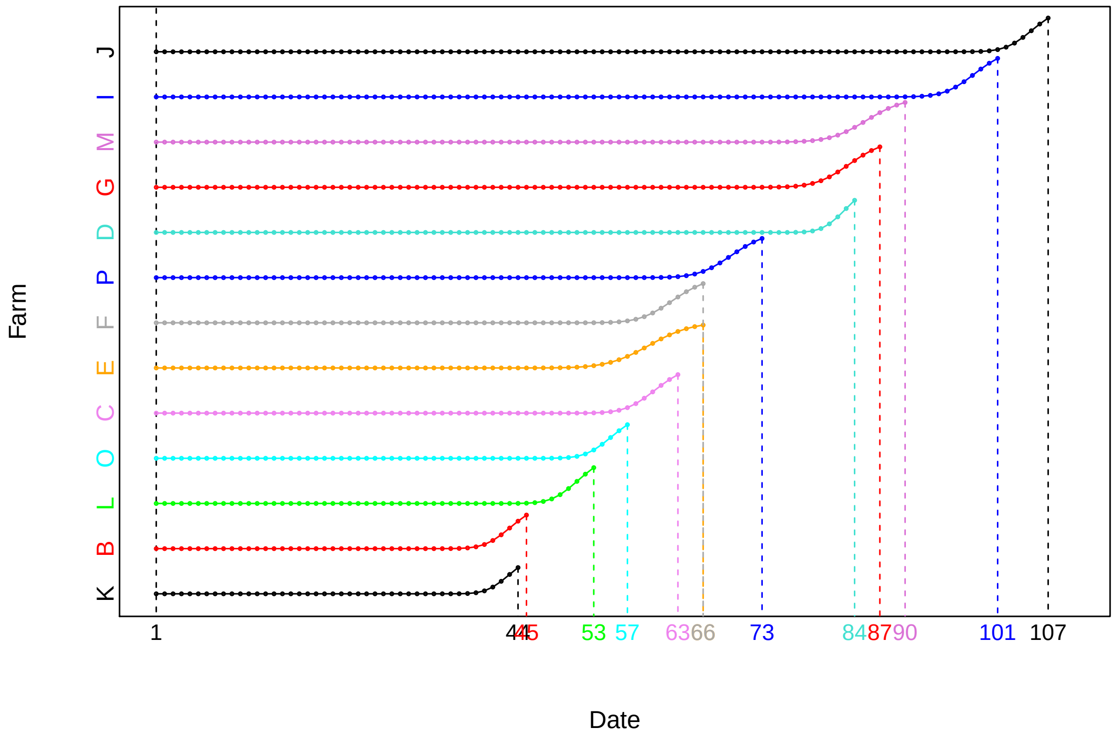

Figure 5.3: Probability profiles of the study farms to be infectious on a given day (calculated as in [Cottam et al., 2008]). Dashed lines indicate the culling dates.

The probability of infection between certain farms i and j on day t is proportional to:

$$P_{i \to j}(t) = F_i(t) \cdot I_j(t). \tag{5.2}$$

And finally, the likelihood that farm i infected farm j on some day is given by:

$$\lambda_{ij} = \frac{\sum_t P_{i \to j}(t)}{\sum_{k \neq j} \sum_t P_{k \to j}(t)}, \tag{5.3}$$

where the denominator is used to normalise λ_{ij}.

In Cottam et al., reconstruction of transmission trees was based on two-dimensional matrix λ_{ij}. It is summed up in Figure 5.4 as likelihoods of pairwise transmissions. Note that genetic data were used in [Cottam et al., 2008] to limit the space of possible trees in the first place. Therefore, the most likely tree that was identified is not the most likely tree structure that could have been obtained from temporal data alone.

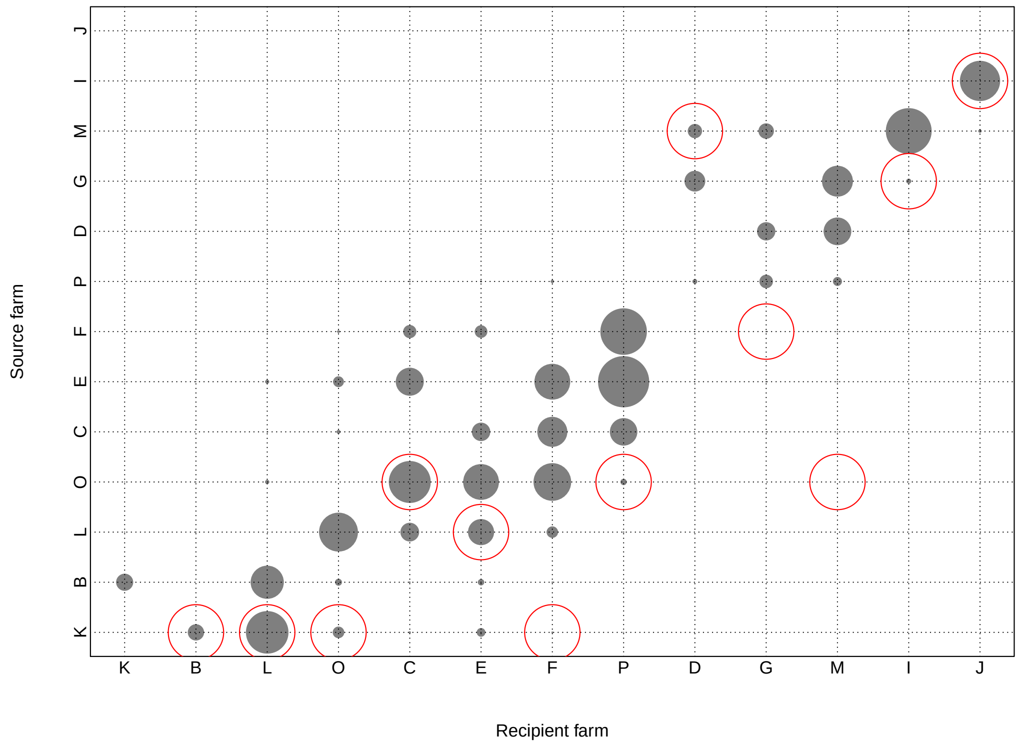

Figure 5.4: Probabilities of pairwise transmissions by Cottam et al. method are represented as the size (area) of gray circles. Red circles indicate the most likely tree structure reported in [Cottam et al., 2008] amongst the trees that were consistent with genetic data (hence, not all the possible trees).

Cottam et al. did not infer infection dates, and did not use them for the calculation of the tree likelihood function. To get the overall likelihood of a given tree, one has to multiply probabilities of individual links (assuming that the individual links are independent). Therefore, implicitly, all the possible combinations of infection dates are considered. Although it might be a reasonable approximation, combinations of infection dates that are not feasible (when a farm is infected after it infects other farms) are also taken into account in [Cottam et al., 2008]. In order to account for the temporal restrictions and exclude impossible scenarios, here we consider timings of infection events and define the likelihood of a tree as the sum of all the likelihoods of scenarios with fixed tree structure and the dates of infection that are in agreement with each other.

5.2.3 Likelihood of epidemic scenario

Here, it is described how the approach by Cottam et al. can be modified to account for timing constraints between potential transmission links. The concept of transmission trees will be extended by taking into account dates of actual infection events. Therefore, infection dates are additional random variables to estimate, which increase the dimensionality of the parameter space.

Let us consider a discrete time epidemiological model that uses a three dimensional matrix M that holds likelihoods of infection for every pair of farms on each day of a simulated epidemic. Every element of the transmission matrix M (m_{ijt}) is proportional to $P_{i\to j}(t)$ (the probability of i infecting j on day t). The likelihood of an epidemic scenario can be obtained by multiplying the probabilities of corresponding likelihoods of transmissions of disease. However, to account for inter-dependance of transmission events, the matrix needs to be updated after every infection.

Recalculation after infection event

If farm i was infected at day d_{inf}, then:

$$I_i(t) = \begin{cases} 1 & t = d_{inf} \\ 0 & t \neq d_{inf} \end{cases}, \tag{5.4}$$

which implies that:

$$F_i(t) = \sum_{\tau=0}^{t} \left(I_i(\tau) \cdot \sum_{k=1}^{t-\tau} L(k) \right) = \sum_{k=1}^{t-d_{inf}} L(k). \tag{5.5}$$

This will affect all the $P_{i\to j}(t)$ that need to be recalculated for $t > d_{inf}$.

In other words, extra information on days of infection that becomes available during simulations is used to correct probabilities of infection for still susceptible farms.

Normalisation of likelihoods in transmission matrix

In order to treat elements of the transmission matrix as probabilities, they need to sum up to one. In [Cottam et al., 2008] where probabilities of transmission do not change over time, elements of the matrix M are normalised so that every farm i will be eventually infected:

$$\sum_j \sum_t m_{ijt} = 1. \tag{5.6}$$

This is already embedded in the calculation of λ_{ij} as it has normalising denominator:

$$\lambda_{ij} = \sum_t m_{ijt} = \sum_t \frac{P_{i\to j}(t)}{\sum_{k\neq j} \sum_t P_{k\to j}(t)}. \tag{5.7}$$

If probabilities of transmission now depend on the current date and the state of epidemic, different normalising procedures are needed to update the transmission matrix after another infection event.

Let us consider a running forward simulation of disease spread between several farms. The current day is $t = d_{cur}$, I is a set of already infected but not yet culled farms, S is a set of still susceptible farms. Before the next infection event all the probabilities that "some farm from I will infect some farm from S" should sum up to one. More formally:

$$\sum_{i \in I} \sum_{j \in S} \sum_{t > d_{cur}} m_{ijt} = 1. \tag{5.8}$$

Thus, until I and S change, elements of Mshould be normalised:

$$m_{ijt} = \frac{P_{i \to j}(t)}{\sum\limits_{k \in I} \sum\limits_{l \in S} \sum\limits_{t > d_{cur}} P_{k \to l}(t)} \tag{5.9}$$

Hence, the next infection event and its timing can be stochastically drawn according to the elements of the matrix. Then transmission matrix is recalculated and renormalised before the next step of simulation.

5.2.4 Likelihood distribution of epidemic scenarios

In order to obtain the likelihood of a particular transmission tree, one should sum up all the possible transmission scenarios, which might be computationally difficult. Alternatively, one might estimate the likelihood of a transmission tree given a sample from the overall distribution of transmission scenarios.

Forward simulations can be used to get the infection dates and sources of infection, which in turn will allow to calculate the likelihood of an epidemic scenarios. But it takes to sum up all the relevant scenarios to obtain the likelihood of a particular transmission tree. Thus, even for a fixed tree structure to generate the full probability density function over all possible timings would require an exhaustive calculations and is therefore computationally intractable (assuming that there are E links in a tree, the timing of every single epidemic event can change within k days, then there are about k^E different scenarios[3]).

Markov Chain Monte Carlo (MCMC) methods can be used to sample possible dates of infection events and obtain a distribution of the likelihoods of epidemic scenarios associated

[3] Impossible timing combinations should be excluded, which will lead to a bit lower estimates. But when possible timings of infections between different farms do not overlap a lot, the order of magnitude will stay the same.

with a fixed transmission tree without considering every possible combination of timings of infection events.

In particular, because it is possible to compute all of the necessary conditional distributions, Gibbs sampling (a common MCMC technique) can be used to sample infection scenarios variable in both tree structure and infection dates.

Sampling of epidemic scenarios using Gibbs sampler

Gibbs sampling is a MCMC algorithm for obtaining a sequence of samples from a specified joint probability distribution of several random variables, when direct sampling is difficult. This is achieved by generating a new sample by changing one random variable at each step by drawing it according to a specified distribution.

Gibbs sampling allows for dealing with high-dimensional parameter space. In our case, possible sources and infection dates are the random variables. Also, the dates when farms started being infectious were taken into account and thus also became random variables.

At every step of Gibbs sampling, all the transmission events but one are fixed. For the considered link the three values are sampled: source farm, date of infection and date of becoming infectious.

For each possible source of infection, the date of infection (γ) can vary from the day when the source farm became infectious (the lower limit) until the upper limit that is the minimum of:

- Cull dates for possible source and the recipient.
- All the dates of next infections caused by the recipient.

The date of becoming infectious (δ) can vary from γ to the upper limit.

For any possible combination of γ and δ the likelihood will be:

$$\mathcal{L}(i, \delta, \gamma) = \sum_{t=\gamma}^{\delta} I_i(t) \mathcal{L}(\delta - \gamma), \tag{5.10}$$

because $F_i(t)$ will be 1 for $t \in [d_{inf}, d_{cull}]$, and 0 otherwise.

Since all other links and dates are fixed, it is guaranteed that all other farms became infectious. Therefore, those that became infectious before the recipient of the considered link was culled are selected. For each of them, the upper and lower limits, and likelihoods for every valid combination of γ and δ are defined. Then, a triplet (source, γ, δ) is drawn using Monte

Carlo method and according to the defined likelihoods. Therefore, the new sampled scenario is obtained. The described process is continued until enough number of samples is reached.

Having taken the likelihood approach for estimation of transmission trees by Cottam et al. [Cottam et al., 2008] as a starting point, two steps were made to extend this method.

Firstly, the overall likelihood of infection between two farms was stratified by considering all the possible dates of this infection. This allowed us to write down transmission matrix and consider timings of infection events in a transmission tree.

The second step was to take into account dependence of sequential infection events. It was done by adjusting the transmission matrix in order to calculate the likelihood of a certain epidemic scenario. This enabled us to obtain a posterior distribution of likelihoods of epidemic scenarios with varying transmission tree structures and infection dates.

Gibbs sampling was implemented in Java. 16 simulation runs were performed for three different starting points. Each of 48 sets contained 10000 samples, and each first 1000 of them were removed as a burn-in period.

5.3 Results

5.3.1 Likelihood distribution of epidemic scenarios

Distribution of the log-likelihoods of sampled epidemic scenarios is presented in Figure 5.5. Even though individual runs could be different from each other, their distributions had similar shapes and means. The distributions have a long tail due to a large number of epidemic scenarios that have low likelihood (unlikely transmission trees and/or dates).

5.3.2 Likelihoods of transmission links

For each observed transmission link in the obtained samples, all the sampled scenarios that contained this link were summarised to comprise the posterior probabilities of individual links (Figure 5.6). The pairwise likelihood distribution of links was different from that defined by Cottam et al. (Figure 5.4), which reflects the differences in two approaches (i.e. accounting for temporal restrictions between individual transmission links).

5.3.3 Distribution of possible dates of infection

Distribution of possible dates of infection for a particular farm can be multi-modal, given several possible source farms. For example, for farm M the distribution was bi-modal (see Figure 5.7).

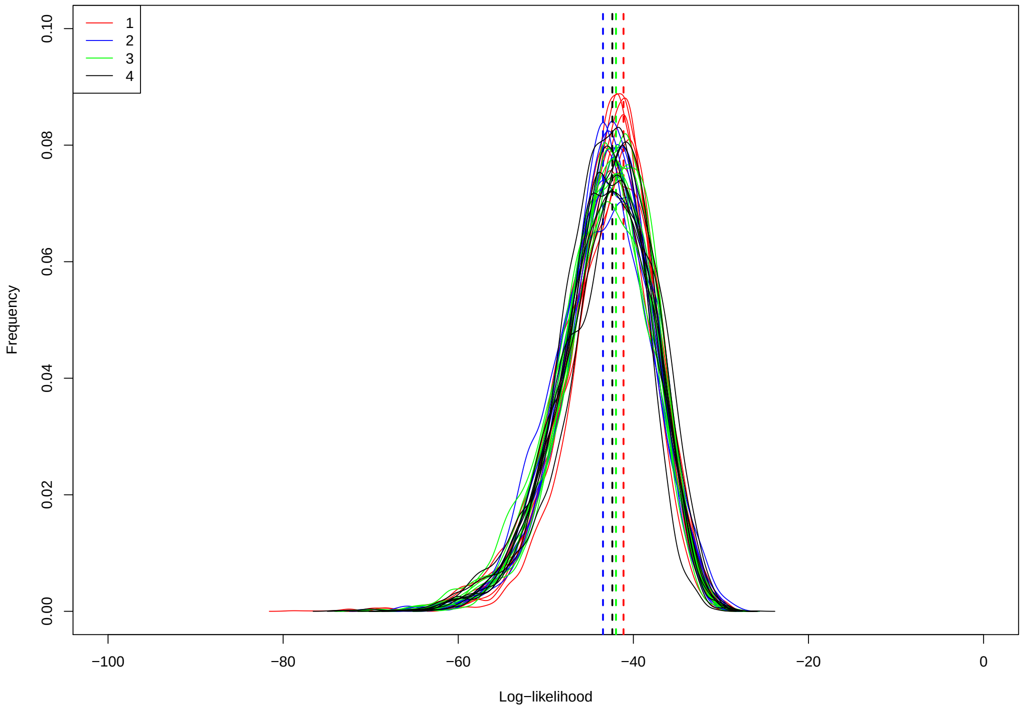

Figure 5.5: Results of the Gibbs sampling: distribution of posterior log-likelihoods of the sampled epidemic scenarios. For each of three starting points (represented by different colours), 16 threads of the program generated 10000 samples each, 1000 samples were discarded as a burn-in period. Dashed lines represent means of three overall distributions (combining results of respective 16 threads).

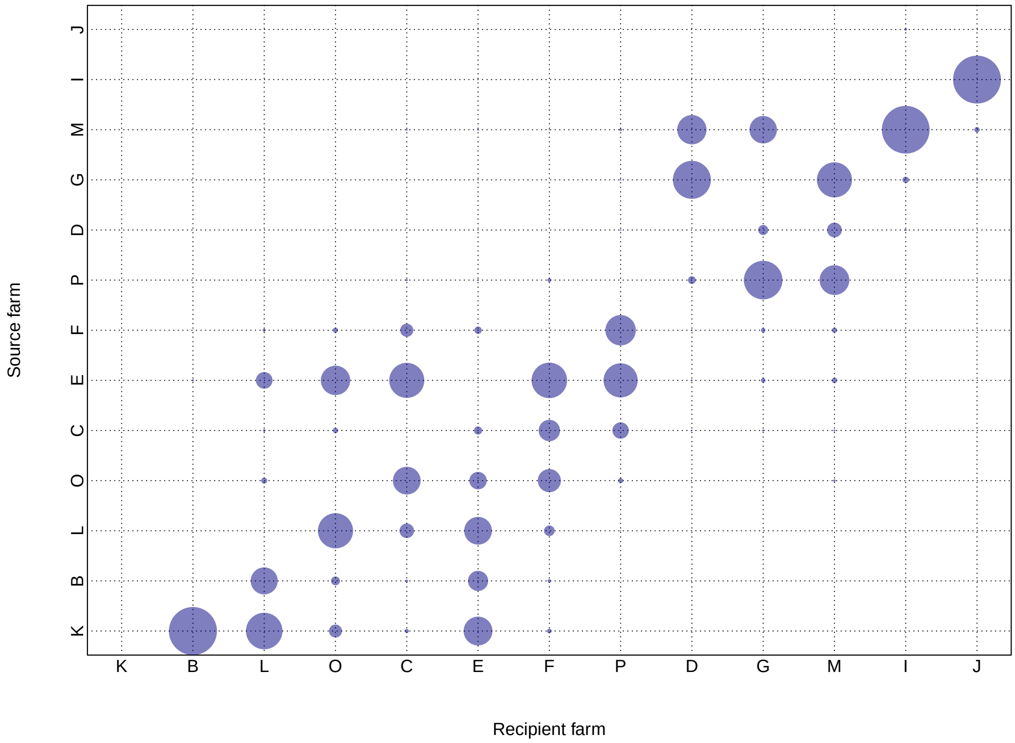

Figure 5.6: The most likely transmission links based on sampled epidemic scenarios. The size of circles (area) is proportional to the posterior probabilities of corresponding transmission links (i.e. the radius is proportional to the square root of the likelihood).

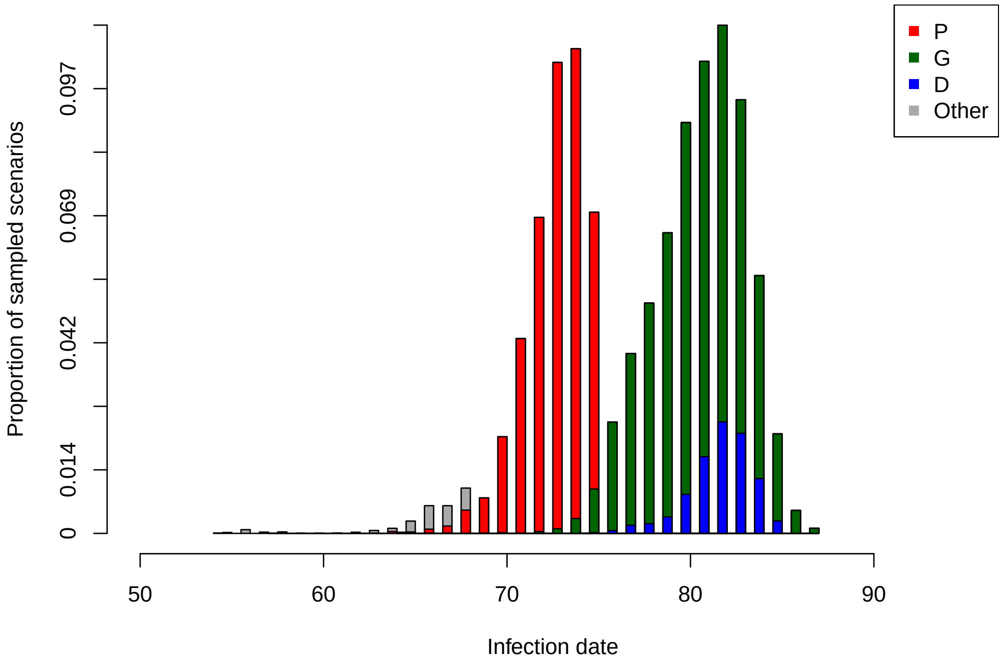

Figure 5.7: Distribution of possible infection dates for farm M. Coloured bars represent proportions of sampled scenarios for different infection sources.

5.4 Discussion

In this chapter, a sampling approach to determine most likely transmission trees was proposed, which allowed to account for temporal restrictions between individual transmission links.

The search space was multi-dimensional (source for each but one farm, plus the same number of infection dates), which made it extremely complicated to write down the likelihood function that would take into account all the temporal dependencies and restrictions. However, Gibbs sampling was applicable in this case to draw a posterior distribution of epidemic scenarios and, ultimately, investigate this distribution to understand the nature of transmission trees that are in agreement with temporal data.

The main difference from the previously published likelihood-based method of transmission tree comparison [Cottam et al., 2008] is that the dependence of links in a transmission chain was taken into account. Which means that to compute the likelihood of a tree one cannot just multiply likelihoods of its individual links. The necessary recalculation procedure for transmission matrix was described, which allowed to calculate the likelihood of epidemic scenario.

The results of Gibbs sampling show that the resulting likelihoods of individual links are different from those computed by the method of Cottam et al. Which highlights the importance of taking into account this interplay of timings of individual transmission events.

The aim here was not to identify the best epidemic scenario which would represent the most likely direction and timing of infection spread. Timings of infection events were considered only for precise computation of likelihoods (to take into account dependence of transmission links). Furthermore, identification of the best epidemic scenario, considering that it is computationally problematic (it cannot be guaranteed that the most likely epidemic scenario was not missed unless all of them were considered), would not lead to a better understanding of the mechanism of infection transmission. Thus, the whole posterior distribution drawn using Gibbs sampling is of interest.

The results of this analysis cannot be directly compared with earlier work (as the whole posterior distribution is more informative as the most likely tree). Also, in other methods genetic data were used, which affected the final outcome. The purpose of this analysis was to show the importance of considering the whole tree structure rather than identifying the most likely tree.

5.4.1 Limiting the search space

The number of all possible transmission trees even for a relatively small number of farms is enormous. However, most of these trees are not feasible. Mainly, due to the time constraints, that can be estimated using collected temporal data (i.e. reporting dates). Other types of data can also be used to limit the search space of possible transmission trees. In [Cottam et al., 2008] the search space was limited at the first step of the algorithm that enumerated only the trees that were consistent with genetic data. This study focused on approaches based on temporal data only, and genetic or spatial data were not used to limit the number of possible transmission trees. Nevertheless, the integration of genetic and temporal data identifies useful criteria for restricting the number of likely transmission trees, and is therefore an invaluable aid to the understanding of the underlying processes. But additional data should be used with caution, because potential errors might lead to exclusion of feasible transmission trees from further analysis. Thus, this can only be done if the data are trustworthy and relevant to the transmission mechanisms of the considered disease.

Incorporating additional data

The approach discussed in this chapter can be extended to allow for using additional data.

In the simplest case, to take into account information that farms are not equally susceptible to infections from each other, one can add to the likelihood function a multiplier K_{ij}:

$$P_{i \to j}(t) = F_i(t) \cdot I_j(t) \cdot K_{ij}. \tag{5.11}$$

For example, K_{ij} can be the distance kernel function that shows at which rate an infectious farm i could have infected a susceptible farm j according to the distance between them. The addition of spatial component, as found in prior models of FMD transmission, would be a natural extension of the presented approach.

Information about number and species of livestock presented on farms can also be incorporated (e.g. parameters for the functions representing probabilities of infection depending on the present animals were estimated in [Tildesley and Keeling, 2009]).

Genetic data possess valuable information about the observed pathogen and can be used to inform assessment of transmission trees. However, more complicated approaches will be needed to update the likelihood function [Ypma et al., 2012, Morelli et al., 2012, Jombart et al., 2014].

For ongoing outbreaks, understanding of the transmission history can be vital in targeting further control. Methods that report only one (the most likely under certain assumptions) transmission tree are limited in advising further targeting of the efforts to control the spread of the pathogen.

During the real outbreaks, some of the transmission links might be confirmed by other sources of data (e.g. closely related genetic samples, contact tracing, etc.) as well as the timing of these links. This information can be taken into account by assessing only those of sampled scenarios that are in agreement with these data.

5.4.2 Conclusions

The presented approach to compare transmission trees is a logical extension of a method previously published by Cottam et al. [Cottam et al., 2008]. It was shown that using conditional probabilities that account for temporal dependence of individual transmission links causes differences in the outcomes of the analysis. This, in turn, allows for more detailed interpretation of posterior distributions of possible epidemic scenarios, which ultimately is useful during outbreak investigations.

Temporal data alone can provide important insight in the transmission histories. However, use of spatial and genetic data can provide extra evidence for situations when several transmission trees are equally supported by reporting dates.

Chapter 6

The effect of rapid contact tracing using movement recording systems for controlling disease outbreaks

6.1 Introduction

Animal movements between agricultural holdings play a vital role in the transmission of many infectious pathogens at regional and national scales. In addition to local spread, long-distance movements complicate the nature of transmission, requiring sometimes relatively complex control measures in order to stop the spread of infection without overly restricting commercial trade.

One of the most notable examples is that sheep movements (via markets) prior to the national movement ban played an important role in the initial spread of FMD in 2001 [Gibbens et al., 2001]. By the time when the index case was identified, the epidemic had gone out of control and further control policies included intensive culling of animals in several regions.

Contact tracing is considered to be one of the key components in controlling infectious diseases. It is effective if infected animals are detected before clinical signs are obvious (i.e. the farm can be reported earlier than without contact tracing procedures). In turn, movement records obtained from animal movement databases play crucial role in locating potentially infected premises.

Tracking of all the farm animals is what everyone would like to have. In reality, this is rarely implemented. In sheep industry, historically, movements are poorly recorded (potentially, because of the relatively smaller price of individual animals, e.g. compared to cattle). This leads to delays in contact tracing when it takes up to several days to identify each sheeps movement history.

There are two major animal movement tracing systems in the UK that record sheep movements: AMLS (Animal Movements license System), SAMS (Scottish Animal Movement System).

AMLS records movements of sheep, goats, pigs and deer within England and Wales and cross-border movements to and from Scotland. Each data record consists of: date, batch size, departure and destination. For movements via markets, the two legs of the movement are captured completely independent, thus there are two records for each movement in the database. This can be problematic to trace individual animals moved from one farm to another via a market (in case the batches are split or merged at the market).

SAMS that is Scottish analogue of AMLS captures each market movement as a single record. The database holds a separate record for each pair of departure and destination farms with an indication of the market involved (if applicable, i.e. not direct farm-to-farm movement). Thus, even if the group of animals was split on a market, it is possible to say that the animals moved from the departure farm to the market are the same animals that moved from the market to the destination farm.

SAMS data provide more information due to the finer batch sizes. However, animal tracing procedures at markets are imperfect, with delays in recording of data and notable errors [Kao et al., 2008].

Most of programmes of FMD control include national or local movement bans because it is important to localise infected premises and disallow long-distance transmissions. Imposed movement restrictions make it easier to control FMD that spreads only locally, but movements that happened during the silent spread of the disease should be recovered and used to identify distant hotspots before they go out of control.

Local spread of FMD is usually controlled by culling [Gibbens et al., 2001] or vaccination of infected animals [Hutber et al., 2011]. However, if the outbreak goes out of control these measures can be applied to potentially infected farms without actual confirmation of infection status (as in the 2001 UK epidemic [Gibbens et al., 2001]). This reduces the efficiency of control programmes.

Examination of animals potentially infected with FMD involves observations by experienced veterinarians: if they observe characteristic lesions, the disease is diagnosed. However, FMD affects different species differently, e.g. clinical signs in sheep are not usually as evident as in cattle or pigs [Kitching and Hughes, 2002]. Thus, examination of farms is performed using ELISA tests [Ma et al., 2011].

The advantages of animal movement tracing for the control of FMD were previously assessed in [Mardones et al., 2013], where a spatial stochastic individual-animal-based model was used against data on Californian farm animals (cattle, sheep, pigs and goats) to compare paper-based (slow) and electronic animal (fast) tracing. For the electronic animal tracing, Mardones et al. assumed that all IPs linked with the diagnosed herd become diagnosed the day after the initially diagnosed herd. Number of traced herds was not limited by manpower constraints. All the herds that were not diagnosed through tracing become diagnosed after two days of the first clinical case in the herd. After being diagnosed, IP no longer sends or receives animals and will be depopulated in 1–2 days. Paper-based tracing was another scenario considered in [Mardones et al., 2013]. The efficiency of this kind of tracing was determined by the number days of animal shipments that could be traced in a single day. It starts with the index case and goes backward to determine all shipments (trace-outs) during the past 28 days. The same procedure is repeated for every diagnosed IP.

The simulations performed by Mardones et al. suggested that an electronic tracing system would reduce the median number of infected premises (IPs) by 8–81%, depending on size of the index case herd compared with the results expected from identifying IPs based on clinical signs alone.

After the 2001 UK epidemic, a standstill policy was implemented, which restricted farmers from moving newly bought animals off their premises for several days after the purchase. This aimed to reduce the probability of disease dissemination: the standstill period was chosen to be long enough for most of the infected animals to show clinical signs, and thus be detected. But the standstill policy restricts farmers from frequent trading and is sometimes violated, imposing an ongoing debate on whether implementation of traceable movement recording systems will allow for relaxation of the standstill policy (http://www.nfus.org.uk/news/view-from-the-top/presidents-blog-15-october-2013).

6.2 Materials and methods

6.2.1 Objective

Here, efficiency of rapid contact tracing in relation to disease control is assessed. The final epidemic size was measured for the simulated outbreaks as a characteristic of epidemic severity and efficiency of control measures.

If there is a difference in the final epidemic size for various contact tracing delays, then it will be possible to determine the threshold delay, which might be used as a recommended standard for implementation of control measures. If, on the other hand, there is no effect of faster or slower contact tracing, this will imply that other control measures should be considered to limit the epidemic burden.

6.2.2 Data

Two datasets were used as sources for the inputs: Agricultural Census data from 2010 and SAMS movements in 2011 to derive input data for the model:

1. Farms data (on Scottish farms: CPH[1], location, numbers of cattle and sheep according to Agricultural Census measures in December 2010).
2. Markets data (locations of 32 markets operating in Scotland).
3. Movements (sheep movements between Scottish farms in 2011).
4. Initial infections (list of farms that could be index cases in the probable FMD outbreak was obtained by extracting sources of sheep movements on each day of the study period).

Several adjustments were made to the data, including: exclusion of movements outwith Scotland, addition of farms missing in the census data but present in movement data. The final dataset consisted of 25219 farms and 103357 movements (77996 through markets) between them in 2011. Distribution of number of movements by month within the year 2011 is shown below (see Figure 6.1).

In order to consider only scenarios where presence and quality of movement recording systems can actually affect the control of epidemic, only situations when the disease spread started after the initially infected sheep moved through a market were modelled.

6.2.3 Transmission model

The disease transmission was modelled using an individual based (at the level of holdings) stochastic simulation model that was inspired by previous work [Green et al., 2006]. Infections were possible via direct farm-to-farm movements, via movements through markets (in addition to point-to-point transmission, uninfected sheep can pick up infection when being on market with infectious sheep) and via local spread. Each farm during the simulations belongs to one of the compartments:

- S — premises with only susceptible animals.
- H — restricted farms, i.e. farms that have exposed animals on them, but these animals are subject to isolation for 13 days after they came to the farm. There is no risk for transmission (through movements or locally) to other farms. Owners of these farms are not allowed to move animals off the farm within next 13 days.

[1] CPH number is a unique farm identifier in the UK, it consists of three numbers that correspond to unique numbers for county, parish and holding, respectively.

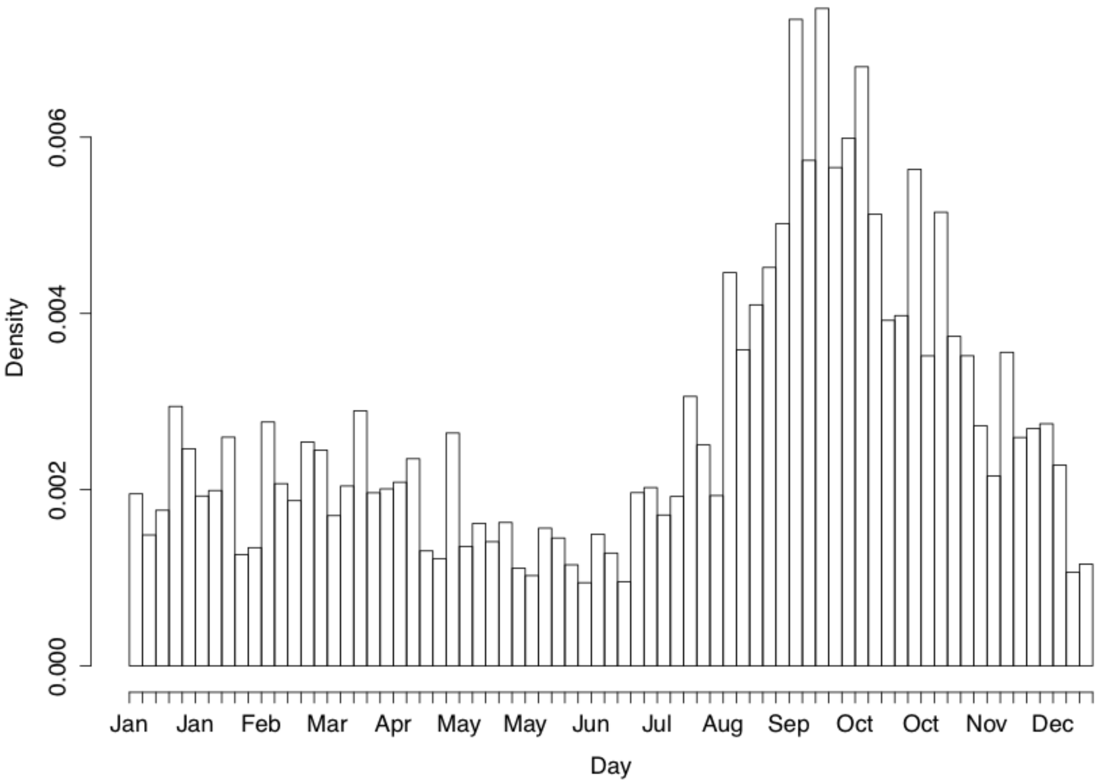

Figure 6.1: Intensity of sheep movements within Scotland in 2011 shown as a density plot indicates a prominent yearly pattern.

- E — farms that can infect other farms by off-movements (of exposed animals) only, not by local spread. A latent (incubation) period of five days was used here.
- I — farms with infectious animals (after incubation period), are possible source of infection.
- R — farms that are either culled or under control, thus do not cause risk of further spread of infection.

The compartmental model is visually summarised in Figure 6.2.

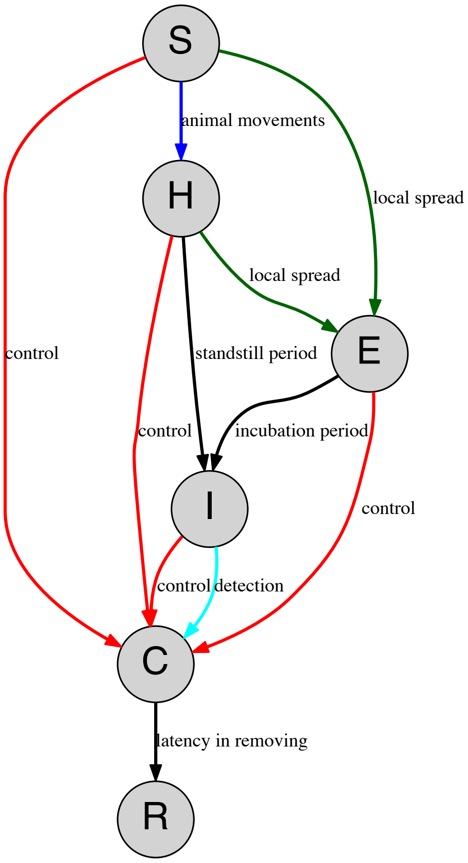

Figure 6.2: The transition diagram of the model. Susceptible farms (S) can get infected via local spread (E) or animal movements (H), and then become infectious (I) after a certain period of time. Susceptible or infected farms can be removed (culled or taken under control) by going via the (C) compartment to (R). Farms with clinical signs in animals (I) can be detected during or contact tracing. Farms that were confirmed to be infected are placed to the confirmed compartment (C) and will be moved removed (R) within a short period of time. Only farms from (I) and (H) (after having passed the incubation period) can infect others.

Most of the previous models of FMD spread used daily time-steps. And although difference of several hours in control procedures can potentially have a significant impact, it was decided to follow previous authors and consider only delays in days.

While disease spread via direct movements was implemented as described in [Green et al., 2006] (the probability of farm infection caused by a movement of a batch of sheep of size b was set to $1 - (1 - \mu)^b$, where $\mu = 0.02$), several changes were made while computing the probability of a farm to be infected by on-movements from markets when one considered the impact of precise knowledge of movements from IPs.

It was assumed that restricted farms (in H state) could not get infected locally (if a farm was infected by movement, it must have infected animals, thus local infection will take more time for virus to develop, so it can be assumed that effectively the farm cannot be infected locally). Also, in the model, restricted farms can infect other farms locally after the incubation period has passed.

When the information about the source farm for a batch of sheep was unavailable, the probability that a batch of animals containing at least one infected sheep can transmit an infection when entering a farm was set to $p_{trans} = 1 - (1 - \mu)^b$, where b is the batch size and μ is the probability of a sheep to become infected and is assumed to be the same for all sheep sold on the same day (since AMLS does not record individual batch movements through market).

If the movement recording system allows to track route of every single batch (e.g. SAMS), the probability of transmission can be corrected to account for higher chance of infection via batches coming from infected farms, and, accordingly, lower chance of infection via those coming from susceptible farms. This was implemented by setting the new probability of transmission to $p_{trans} = 1 - (1 - \mu_F)^i(1 - \mu')^s$, where i is the number of sheep moved from infected farms, s is the number of sheep moved from susceptible farms, and μ_F is the probability of infection through known movement from infected farm for one sheep (which is higher than μ). Thus, the increase of infection probability via sheep coming from an IP must be counterbalanced by a decreased probability of infection from susceptible farms (μ') that can be explained by a baseline risk of infection due to contact with infectious sheep from other batches at the market. To ensure consistency with the previous model [Green et al., 2006], μ_F was set to be the same as for direct farm-to-farm movements (0.02), μ was the same as for off-market movements regarding the source farm (0.004), and μ' was fitted so that the overall force of infection via markets for each market should be the same as previously (in other words, the same as the average overall number of IPs generated via one market by the passage of sheep from IPs found in previous work [Green et al., 2006]).

For the local spread, the same constant rate of generation of new cases was used as in [Green et al., 2006] ($\beta = 0.065$ per day per infectious holding). On each simulation day, a number of infectious contacts were sampled from a Poisson distribution (without replacement) independently for each infectious farm. Then infectious contacts are chosen from all susceptible farms within 10 km, weighted according to the distance d by $p \sim e^{-ad}$, where $a = 0.5$ km^{-1}. However, the likelihoods of infectious contacts can be weighted differently, e.g. one might use previously estimated distance kernels or even take into account species-specific parameters as in [Tildesley et al., 2008].

6.2.4 Control strategies

Control programmes are often complicated and include various parts from pre-emptive culling to vaccination. In 2001 in the UK, control policies changed with the course of the epidemic and were also different from region to region.

Previously, detection of dangerous contacts (DCs) in simulation models was stochastic and based on the actual infection [Tildesley et al., 2008]. The assumption was that the probability of tracing matched the probability of infection in terms of distance. In [Mardones et al., 2013], after the primary case was diagnosed, IPs that were not diagnosed through tracing became diagnosed 2 days after the first clinical case in this herd. In [Boklund et al., 2013], simulated control measures included imposition of protection and surveillance zones (within 3 and 10 km from an IP) and contact tracing.

Here, control measures were simulated using two mechanisms:

1. Detection of infectious premises after a certain number of days with clinical signs.
2. Contact tracing of on- and off-movements to target serological examinations at farms that could have received infected animals from other farms.

In this model, it was assumed that after seven days of clinical signs farm would be reported to be infectious by the owner, which then in turn implies diagnostic and removal.

Implementation of contact tracing

Contact tracing is simulated using the following assumption: once a farm is confirmed to be an IP, all the animals on it will be culled within 24 hours. The next task (both, in modelling and in reality) is to identify dangerous contacts (DC) and examine farms that were at risk of infection.

3km and CP culls were used during the 2001 epidemic in the UK [Keeling et al., 2001, Kao, 2002, Haydon et al., 2004], but are unlikely to be immediately applied in the future FMD outbreaks in Scotland (according to the "Foot and mouth disease control strategy for Great Britain", available at `https://www.gov.uk/government/publications/foot-and-mouth-disease-control-strategy-for-great-britain`).

Thus, it was assumed that there would be no pre-emptive culling, only diagnosed farms would be removed (i.e. culled or put under restrictions that will allow for no further spread of the infection). Actual implementation of the culling procedures is beyond the scope of this paper, and the number of culled animals was not assessed. Confirmed IPs are transferred to the (R) compartment and cannot infect other farms.

Figure 6.3 summarises all the possible cases of dangerous contacts with a farm that was confirmed to be infectious (IP).

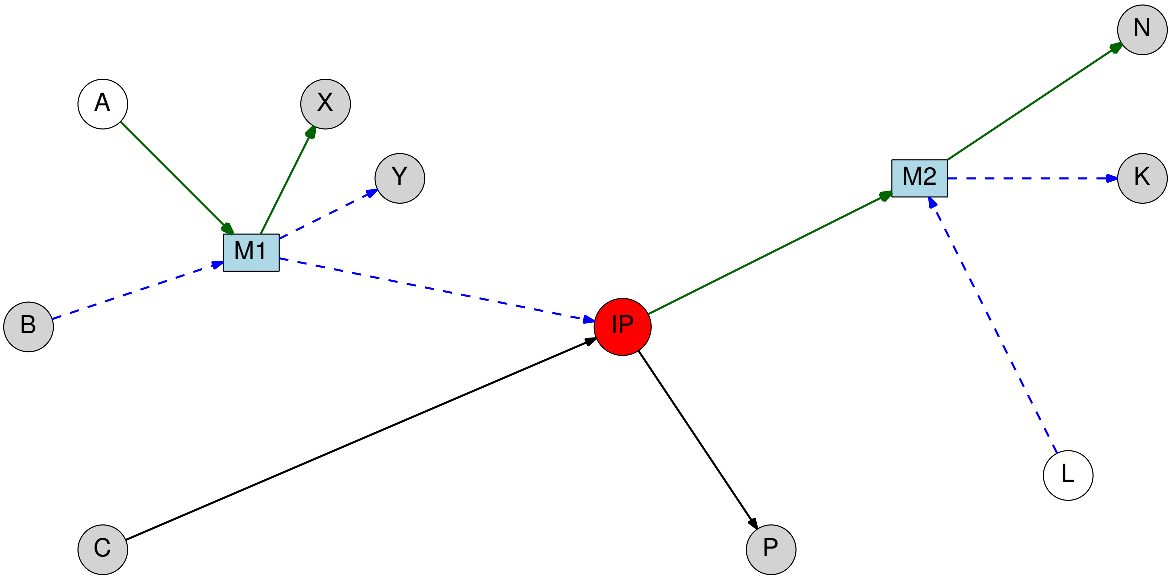

Figure 6.3: Schematic explanation of farms that are considered to be dangerous contacts (DCs) with infected premises (IPs): gray circles are DCs, while white ones are not. Green solid and blue dotted edges are used to show different batches of sheep moved on and off market (e.g. batch from A was moved to X, while batches from B were moved to Y and IP).

To obtain farms that traded sheep with an IP, the contact tracing window was set to 21 days, which is in agreement with current policies in the UK.

During simulations, for all the movements on and off the IP that happened within 21 days preceding the identification:

1. Direct contacts (via farm-to-farm movements) are put into the investigation queue so that they will be examined after 1 day (to simulate latency in contact tracing).

2. For movements through markets, after identifying dangerous markets (i.e. those that received or sent animals from/to the IP) for the days of contact with the IP, all the recipients of animals on those days are put into the investigation queue to be examined after 4 days.

Further in the text, these delays of 1 and 4 days will be referred as tracing delays (or latencies in tracing) for direct and market movements, respectively. It is assumed that they are higher than zero reflecting the time needed to obtain movement records (which will vary for direct and via market animal movements).

In case of using more advanced movement tracking systems (e.g. EID systems), the delay of obtaining contacts via markets will decrease and become close to the one for direct movements. Different values for the delay in movement tracing for markets will be explored later, but the delay for direct movements was constant.

The contact tracing procedures are performed iteratively, starting at the farms that were diagnosed by clinical signs first. Once the examined farm (from the investigation queue) is confirmed to be infectious, it becomes an IP and the same procedures are performed for its contacts. If the farm is not infectious, further contact tracing is not triggered.

6.2.5 Modelling strategy

In order to choose index case(s) for the simulated outbreak, one should think about the most feasible and realistic scenarios of disease introduction, setting initial conditions by choosing index farm(s) arbitrarily is not always sufficient.

Given the focus of this analysis, it was decided to seed the simulations at farms that have sold sheep, allowing for the dissemination of the pathogen.

The number of initially infected farms was set to 5, and choose them from the list of all the Scottish farms that sell sheep within 3 days after the start of simulations.

Index cases are put into I (infectious) compartment, and the disease spreads unnoticeably for 20 days (after the start of simulations), which is in agreement with 2001 UK outbreak [Gibbens et al., 2001]. Then, national movement ban is applied, so that the infection can only spread locally. Together with movement restrictions, other control efforts (examination and culling of suspicious farms) are also put into place.

The simulation finishes when there are no infected (E, H or I) farms in the population, and the investigation queue is empty. Simulations were set to start at different time points throughout the year (first and fifteenth days of each month). The parameters used in the model are summarised in Table 6.1.

Absence of EID tagging system or high read failure can indirectly affect contact tracing procedures, which will result in increased time needed to identify farms that received/sent animals from/to confirmed infected premises. Therefore, different parameter regimes for varying movement tracing delays were considered, and outputs of the simulations were investigated with a focus on the final epidemic size: the total number of infected farms (both: locally and through movements), which is the same as the number of removed (including five index cases).

Table 6.1: Summary of parameters used in the model.

Parameter	Value (days)
Number of index cases	5
Latent (incubation) period	5
Control measures delay	7
Movement ban delay	20
Movement restrictions	13
Contact tracing depth	21
Delay for direct movements	1
Delay for market movements	4
Detection time	7

6.3 Results

Figure 6.4 presents the number of infected farms (through local spread and movements) against starting points of simulations for different tracing delays. Latent period was set to 6 days, every infectious farm was detected after 4 days of clinical signs and index cases were fixed between repeated simulations.

The epidemic size follows the same annual pattern as in previous studies [Green et al., 2006, Orton et al., 2012]. Which is due to the fact that mainly the intensity of sheep movements determines the final size of the affected population. Although it was expected to observe an increase of the number of infected farms for the increasing movement tracing delay, it was not present. This can be explained by the presence of standstill policy: the movement restriction period is long enough for clinical signs to appear in infected animals, and hence to make the disease visually detectable.

Then the aim was to assess how detection time and tracing delays are connected. Thus, the regime when standstill period (13 days) was less than incubation plus detection was taken into consideration. Simulation outputs for latent period and detection time of 7 days, and tracing delays of 1 and 10 days are presented in Figure 6.5.

Outputs for unrealistic set of parameters

Exploration of parameter space with realistic values (those that were not far away from previously estimated) did not yield in simulation results that showed dependence on the tracing delays. Therefore, to show that such parameter sets exist, unrealistic values for parameters were chosen.

Local transmission was disabled, probability of infection through movements was raised to 100%, and index cases were fixed for each starting point. Therefore, the results were the same for any run of the simulations. The incubation period was 7 days, detection time was

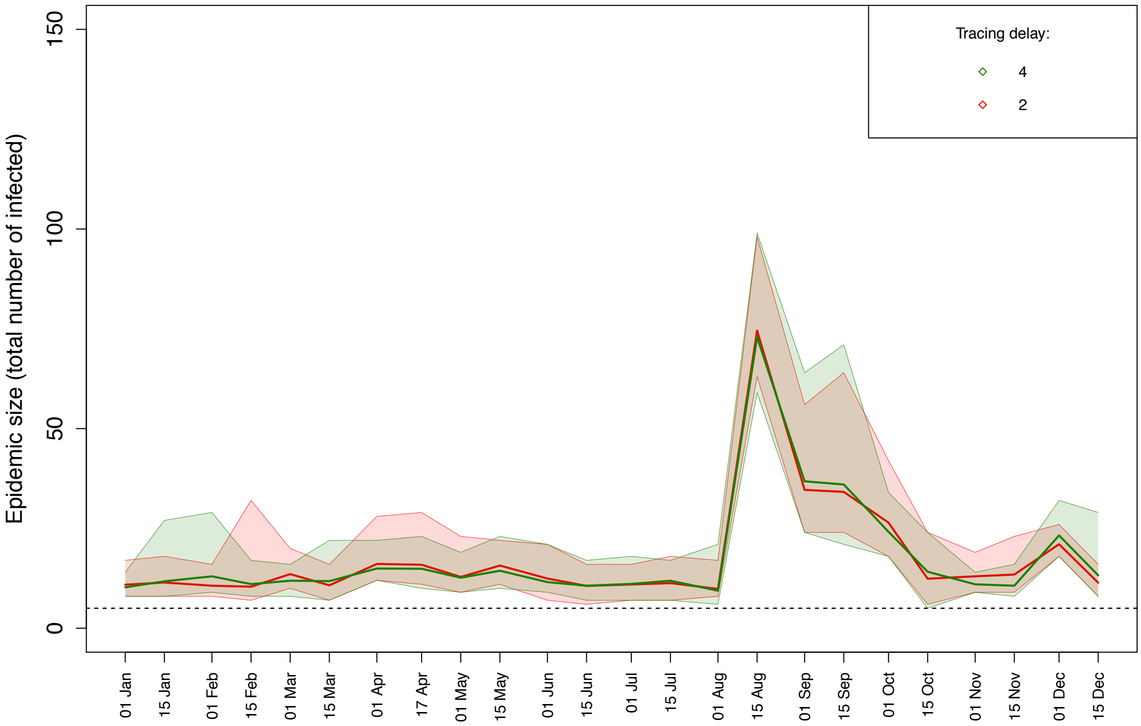

Figure 6.4: Effect of movement tracing delay on epidemic size for detection delay of 4 days and tracing delays of 2 or 4 days (red and green colours used, respectively). The graph shows mean number of farms (along with 95% envelopes) that were infected during simulations that started at different time points throughout the year. There is no significant effect of movement tracing delay on epidemic size.

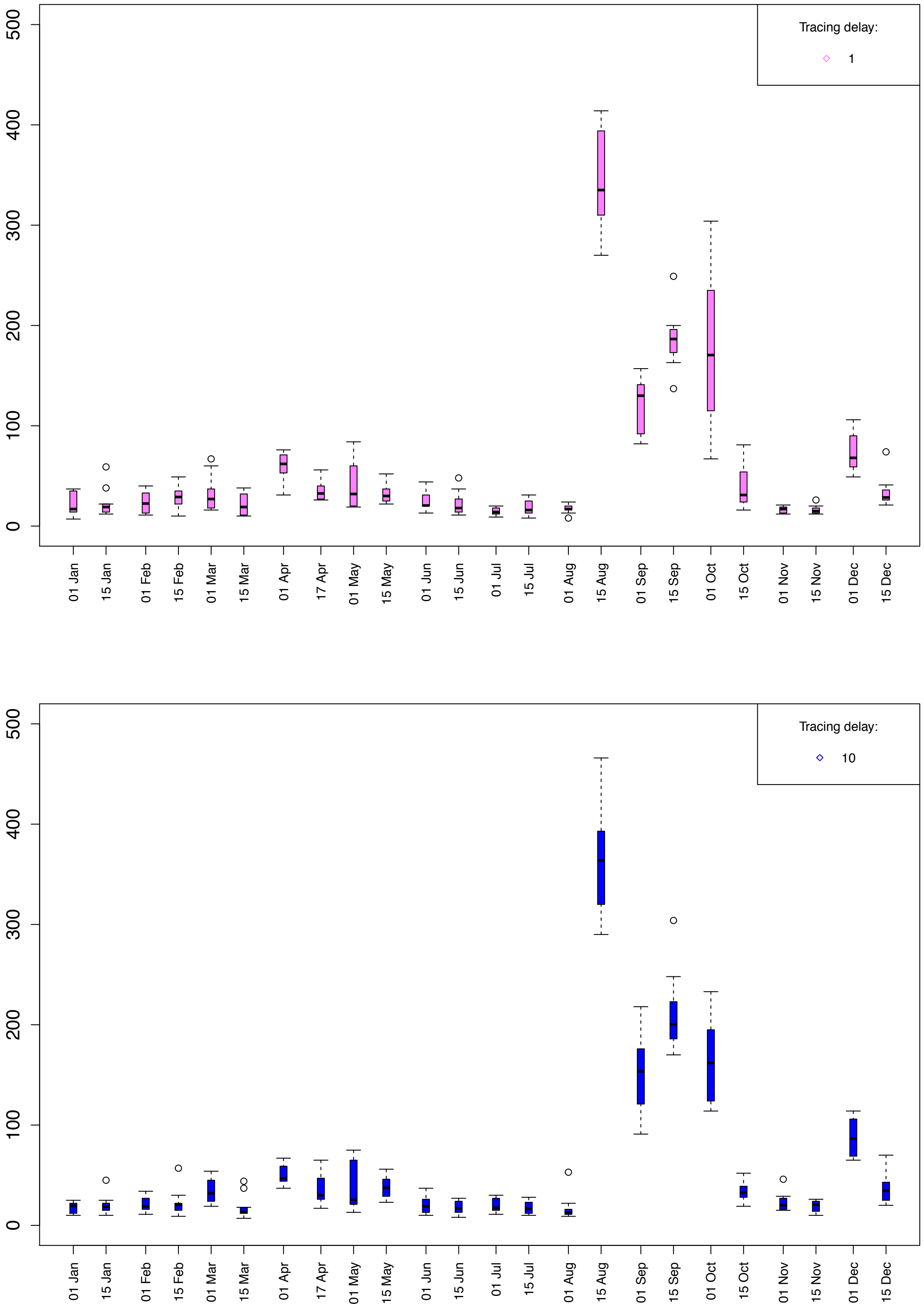

Figure 6.5: Results of simulations for the situation when latent period (7 days) and detection time (7) summed up are higher than standstill period (13 days). There is no significant difference in final epidemic size for tracing delays of 1 and 10 days.

7 days, control measures delay was 7 days. The results for this unrealistic set of parameters are presented in Figure 6.6.

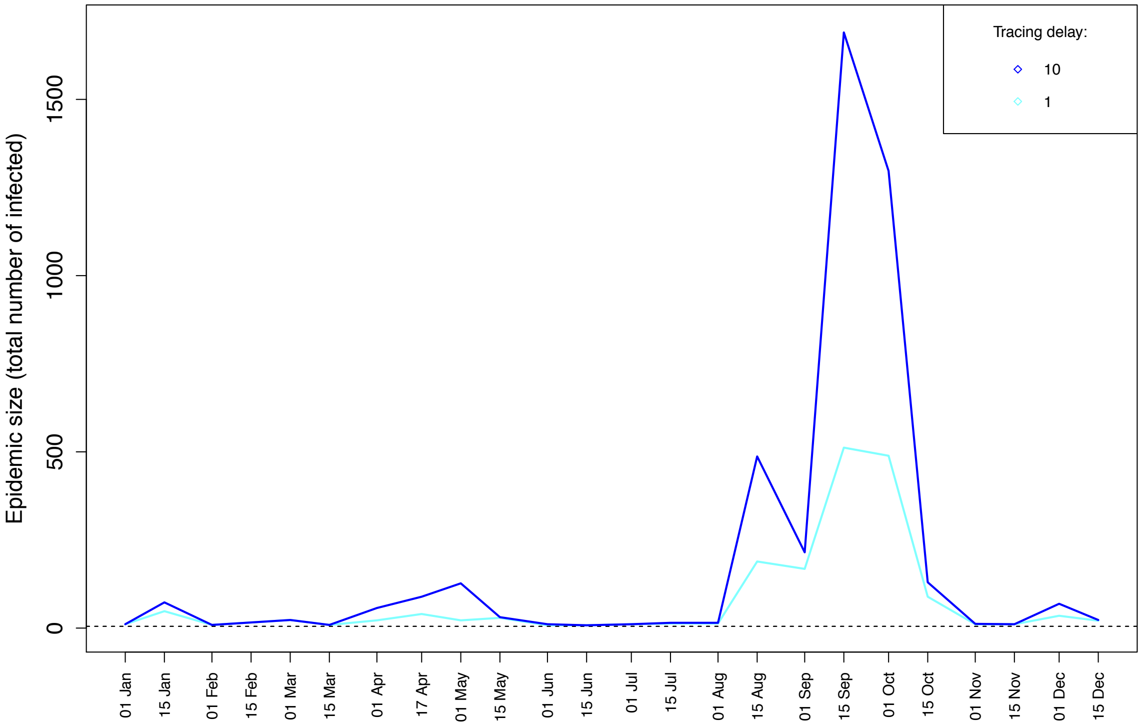

Figure 6.6: Epidemic size for simulations with unrealistic parameters (no spatial spread, 100% chance of infection via movements, detection after 7 days). The raise of the number of affected premises (for certain starting dates) caused by the increase of tracing delay for movements through markets (1 and 10 days, shown in cyan and blue, respectively). This can be explained by the presence of farms with on- and off-movements that are close in time and compose infection chains that can be broken by the contact tracing procedures with sufficiently small delays.

These results confirm that, for some parameters, tracing delays can influence the course of the epidemic. However, for realistic (for FMD) parameters the effect of faster contact tracing gets lost due to the implementation of standstill policies.

6.4 Discussion

The purpose of contact tracing is to find and examine infectious farms before they are notified by the owner (and thus facilitate the control of the disease). Thus, detection time and delays in tracing (direct/markets) counter-balance each other regarding the efficiency of contact tracing. When delay in contact tracing is much higher than the detection time, the purpose of additional investigations is absent (farms will be diagnosed earlier anyway).

In [Mardones et al., 2013], the tracing procedures were implemented differently: firstly, there were no standstill policies that restrict farmers to sell animals shortly after purchasing them elsewhere; secondly, all IPs linked with the diagnosed herd through a network of traced shipments become diagnosed the day after the initially diagnosed herd. In the model presented in this thesis, examinations and tracing are performed iteratively (first order contacts are examined, which triggers further actions only if the infection has been found).

The whole point of restricting farmers from selling is to increase probability of identifying the disease on farm, before the infected part of the herd is divided and moved to a number of farms. Thus, the choice of timing (13 days for sheep in Scotland) that allows the infection to develop and become obvious. Mathematically, this can be expressed as: $t_{standstill} > t_{incubation} + t_{detection}$.

The impact of tracing speed on epidemic size can be mitigated by the noise added by the stochasticity of local spread. Although the effect of rapid tracing is here, the resulting epidemic sizes may vary to a larger extent.

Also, the effect is easier to observe in the simulations that started during the period with higher level of animal movements (August–October). This is due to availability of more transmission chains that lead to higher epidemic impact and can be broken during rapid contact tracing.

Other species that would be involved in the spread of FMD were not taken into account. The assumption was made that transmission between sheep and other species (and within other species) will not change the modelling results, as transmission in other species is likely to result in a movement ban (clinical signs would be rapidly seen in either cattle or pigs). Implementation of control policies was also assumed to be perfect, in reality it can be implemented with errors but these jitter/noise effects were neglected.

The purpose of this study was to estimate the effect of latency in tracing. The effect of possible vaccinations was not assessed here, because it has never been used in the UK to control FMD.

6.4.1 Data issues

In the initial dataset, there were premises that were in different counties from those specified by their CPH numbers. Also, a certain amount of movements did not fall into two categories of considered movements: direct movement from farm to farm, movement through market (e.g. movement from market to another market and then to farm). It was assumed that elimination of the data entries that were inconsistent with the models description did not affect the results.

6.5 Conclusion

The final epidemic size was reported to be dependent on the intensity of movements that happened in the initial (silent) period of the outbreak. However, the effect of faster contact tracing was not evident due to high variance of the final epidemic size. The results of the simulations suggest that for FMD-like parameters, earlier detection of farms by movement tracing does not appear to enhance control of post-silent spread.

This might seem as an unexpected result, given that on the level of farm earlier detection of the disease usually results in isolation of infected animals from others, and therefore decreases the epidemic burden for farm. Contrastingly, on the national level this was not observed: epidemic size was not affected by the changes in contact tracing delays. Primarily, this was due to the fact that by the time of examination all the traced DCs have already progressed to infectious stages and had enough time to infect large numbers of potential sources for farms in almost all areas, which in turn implied a farm is not infected by another farm that was detected by movement tracing, there is a number of alternative sources that will infect the initial farm in any case. Here, it should be noted that this "demographic surplus" is dependent on the transmission parameters of the modelled disease and is not just an effect of high farm density in certain areas.

It was demonstrated that for the unrealistic set of parameters increase of the tracing delay is positively correlated with the final epidemic size.

On the other hand, for extremely high levels of local transmission, any affected region will become saturated with local cases of FMD. This is an indication that there are two types of parameter sets that capture disease transmission: those that are sensitive to delays in movement tracing, and those that are not.

Further studies should be focused on broader explorations of parameter sets that are sensitive to variable contact tracing delays. Knowledge of these parameter sets will lead to a better understanding of whether the implementation of particular control strategies is worthwhile or not.

Additional considerations of logistics could however, increase the benefit of early detection — should the initial number of cases post-silent spread rise more slowly as a result of rapid detection, this may provide the opportunity for control teams to "get ahead" of the epidemic, an effect not considered in this study, but that could be added with additional data.

Chapter 7

General discussion

7.1 Data in epidemiology

The central component of any epidemiological research is data: quality, resolution and amount of available data define the range of scientific questions that can be addressed.

Density of veterinary data is increasing at a vast pace. This includes detailed information on the contact structure amongst livestock units, spatial locations as well as densely sampled disease datasets which increasingly include information about the relatedness of the sampled pathogen itself.

The increased availability of these data is only gradually being accompanied by methodological advances to fully explore it.

In this thesis, I have considered different angles to these methodological questions, considering two exemplars of disease systems for which dense data were available: FMD in the UK and *Streptococcus agalactiae* in Denmark.

7.1.1 The role of centralised databases

Movement recording systems provide invaluable help to the understanding of the dynamics of livestock populations. They have undergone significant developments in the recent decades: from paper-based versions to large-scale databases holding information on all animal relocations within a whole country. In countries, where there are no such databases or they are used only on the regional level, it is difficult to control major outbreaks that might happen on the national scale.

Centralised databases that hold genetic data are also extremely useful nowadays when molecular data are generated in vast amounts worldwide. For example, the public MLST database [Jolley and Maiden, 2010] (available at `http://pubmlst.org/sagalactiae/`) that was used in chapter 2 helped to relate the population composition of *S. agalactiae* from the Danish dairy cattle to the data collected from other places.

7.1.2 Data cleaning

Usually collected data can not be used directly for the epidemiological analysis: it should be verified and cleaned first. In certain cases, some data should be eliminated for consistency of the dataset, even if it will decrease the statistical power of the analysis. For example, in chapter 2, it was decided to exclude passive surveillance data on *S. agalactiae*, because it was collected at different dates throughout year and obviously did not guarantee a necessary coverage of dairy herds.

Large datasets need good systematic approaches of handling data issues. Translating from qualitative assertions and expert opinions to quantitative approaches is therefore important and nontrivial.

Case definition plays a central role in forming the principal dataset for epidemiological analyses and is normally based on one standard method. However, in this thesis, several possibilities were considered to define a herd that is infected with *S.agalactiae*. The study period overlapped with legislative transition and changing of standard surveillance method (switch from bacteriological culture to PCR in Denmark in 2010).

In chapters 4 and 6, adjustments were made to the animal movement data by exclusion of those datelines that do not impose risk of disease transmission. This is critical for movement data as the final datasets are large and inclusion of extra data can slow down the network analysis methods. Also, the approaches used must be relevant to specific diseases rather than just generic.

7.1.3 Contribution of genetic data

Genetic data provide another source of information about a particular pathogen. However, genome sequences or their fragments should be treated differently for different pathogens. For example, for fast-evolving pathogens like FMD virus, genetic data can provide additional information regarding the relationship between isolates [Cottam, 2007]. On the other hand, bacteria usually have slow mutation rate and can not fix mutations at the rate that would allow for the necessary discriminatory power to use it as a marker for transmission [Spratt, 1999, Spratt and Maiden, 1999]. Furthermore, such techniques as MLST [Enright et al., 2001,Jones et al., 2003] only consider small parts of the genome (conservative genes that do not change fast), which makes it inapplicable for phylogenetic analysis.

Genetic data are not perfect — recombination and reassortment can cause interpretation difficulties [Spratt and Maiden, 1999,Feil and Spratt, 2001], thus the results of genetic analyses should be handled with caution.

Usefulness of strain data

However, strain typing data provide invaluable information on bacterial population composition, which can be used to investigate features of transmission specific to different subpopulations [García Álvarez et al., 2011].

In this thesis, bacterial subpopulations of *S. agalactiae* were considered on two different levels: sequence type (ST) and clonal complex (CC, that was defined as a set of genetically closely related STs).

Use of CC instead of ST level analysis can refine the relationship regarding spatial clustering or contact patterns (if transmission properties are indeed shared between closely related STs). It also increases the statistical power as more herds can be included in the same group for analysis (e.g. CC1 is bigger than ST1).

Aggregation of STs into CCs for strain-specific analysis will be meaningful only if the member STs share similar transmission features, otherwise distinctive patterns of member STs will interfere with each other and it would be harder to identify consistent patterns for the overall CC. Therefore, by amalgamating one can lose strain-specific differences (e.g. how it happened in chapter 3, when CC103 showed less spatial clustering than ST103).

7.1.4 Analysis of strain-specific transmission patterns

The strain-specific features of transmission of *Streptococcus agalactiae* between Danish dairy cattle herds were assessed in this thesis.

Spatial clustering analysis revealed differences between some of the strains of *S. agalactiae*: ST103 was significantly clustered among other strains throughout the whole study period, unlike other predominant strains ST1 and ST23. The fact that ST103 was associated with cattle as its primary host (while ST1 and ST23 were common in humans) supports the hypothesis of significant role of humans in spreading the pathogen. Results of CC-level analysis suggested that specific transmission properties are likely to be linked with STs rather than broader subpopulations defined by CCs (several STs grouped together).

Network analysis performed in chapter 4 helped to identify that cattle movements between dairy farms could be associated with the spread of major STs. Notably, cattle movements could not explain the distribution of ST103 in 2010, unlike in 2011. Accompanied by the fact that ST103 was the only strain, which distribution was linked with sharing of veterinary practices, it is reasonable to suggest that ST103 was just recently established in Danish dairy population andm thus the switch in potential transmission routes.

Further analysis is required to determine drivers of strain-specific transmission mechanisms. This might include studies of the *S. agalactiae* populations isolated from human carriers (mainly, on-farm visitors and farm workers) and their association with the bovine isolates from the same area.

Further investigations of *Streptococcus agalactiae* transmission might consider joint methods for looking at spatial and network clustering at the same time. Future analyses should not be limited to computational analysis of herd-level data. Clinical experiments can also play vital role and validate the hypothesis of the role of humans in the spread of the pathogen.

7.1.5 Estimation of transmission trees

Initial likelihood-based approaches [Cottam et al., 2008] to reconstruct transmission trees have considered only the timing of individual events rather than attempting to fit the entire tree simultaneously. This process is technically challenging as one has to account for temporal constraints between subsequent infection events, and therefore it was proposed to use a Gibbs sampling approach which is demonstrably efficient.

The most likely trees identified in the previous approach and in this thesis depict several differences between them, suggesting that the evaluation of the tree as a whole can be important.

Alternative bayesian approaches [Morelli et al., 2012, Jombart et al., 2014] have advantages of incorporating genetic data. And the approach presented in this thesis might be extended to account for possible transmission patterns that are in agreement with observed genetic samples.

7.1.6 The role of earlier detection of highly infectious pathogens by movement tracing

In chapter 6, mathematical modelling revealed that for FMD-like parameters, earlier detection of farms by movement tracing does not appear to enhance control of post-silent spread.

However, other sets of parameters showed that there is a correlation between the delays in contact tracing and the final epidemic size. Therefore, investigation of what sets of parameters are sensitive to contact tracing delays possesses a significant interest as it will allow to design more relevant control policies.

7.2 Future perspectives

7.2.1 Whole genome sequencing

Typing techniques like MLST are based only on several conservative fragments genomes, whereas whole genome sequencing (WGS) can capture even slightest changes in pathogen genomes on a relatively small time scale.

WGS data hold a vast amount of information that can be used not only to identify different strains (like MLST or serotyping), but also to pursue questions on pathogen evolution, virulence and antimicrobial resistance (AMR).

WGS is becoming more affordable nowadays. However, MLST techniques are well-established for most of bacterial pathogens and structures of their populations can be characterised with sufficient level of detail.

In this thesis, the questions of evolutionary changes in *S. agalactiae* were not addressed. The timing of sample collection (annual periodicity) did not allow for this. However, further studies that will involve intensive sampling of individual cows might reveal some fine-grain properties of within-herd and between herd transmission.

7.2.2 Methodological improvements

The recent trends in availability of data for epidemiological research can be summarised into several points:

1. Modern animal surveillance systems are capable of collecting large amount of high-resolution data on a daily basis.

2. National and regional control programs ensure the sufficient coverage of collected data.

3. Amount of available genetic data has also increased recently, given the advances in next-generation sequencing. And WGS analysis can reveal epidemiological features at multiple scales.

All of the above increases the level of requirements for methods used in analysis and modelling of infectious pathogens. They should be more efficient in working with large datasets, robust in their performance for different disease systems, but at the same time capable to account for distinctive features of a considered pathogen.

This thesis is yet another attempt to reconsider existing methods regarding their applicability for a range of epidemiological questions related to spatial and network aspects of livestock disease transmission.

7.3 Conclusion

In this thesis, methodological aspect of dealing with spatial and network data were considered and discussed. Given the increasing importance of novel, more efficient methods in epidemiological analyses, the purpose of this thesis was to make another step forward in the direction of understanding and improving existing methods.

Appendix A

Additional graphs and tables

Table A.1: Summary of corrections for carry-over: confirmed false positive herds are in gray. The table summarises information for potential FP herds (column “FP herd”) and respective potential sources of contamination (“Src herd”). Ct values (“PCR” for potential FP, and “PCR-1” for source) were assessed: we expected “PCR-1” to be lower than PCR for true FP herds (i.e. “PCR diff” > 0). Also MLST data were used: we expected the same ST for true FPs (“Same ST”).

Year	FP herd	Src herd	PCR	PCR-1	ST09	ST10	ST11	ST09-1	ST10-1	ST11-1	PCR diff	PCR consist.	Same ST
2009	8	9	28.2	26.2	1	1	1	1	-	-	2	YES	YES
2009	9	284	26.2	17.3	1	-	-	-	0	1	8.9	YES	NO
2009	12	11	40	29.9	88	-	-	1	1	-	10.1	YES	NO
2009	19	20	35.9	26.1	23	-	-	23	23	23	9.8	YES	YES
2009	23	22	40	22.2	1	-	-	1	1	1	17.8	YES	YES
2009	64	62	26.8	22.5	23	23	-	314	314	314	4.3	YES	NO
2009	79	80	40	30.5	1	-	-	1	1	1	9.5	YES	YES
2009	88	86	19.9	20.1	103	-	103	103	-	103	-0.2	NO	YES
2009	89	90	40	31.2	103	-	-	103	-	-	8.8	YES	YES
2009	90	88	31.2	19.9	103	-	-	103	-	103	11.3	YES	YES
2009	92	93	40	28.4	1	-	-	1	1	0	11.6	YES	YES
2009	95	94	34.8	25.9	314	-	-	314	-	314	8.9	YES	YES
2009	99	144	24.2	22	103	103	-	1	1	1	2.2	YES	NO
2009	103	101	31.3	20.9	10	-	-	10	-	103	10.4	YES	YES
2009	112	113	30.7	31.5	103	-	103	23	23	-	-0.8	NO	NO
2009	113	114	31.5	31.5	23	23	-	625	1	-	0	NO	NO
2009	129	75	30.9	21.9	103	-	-	103	103	103	9	YES	YES
2009	130	332	30.9	37.5	1	1	-	0	-	-	-6.6	NO	NO
2009	133	132	20.8	31.3	103	1	-	8	8	314	-10.5	NO	NO
2009	134	135	32.5	35.9	103	-	-	103	-	-	-3.4	NO	YES
2009	135	133	35.9	20.8	103	-	-	103	1	-	15.1	YES	YES
2009	139	138	29.3	30.2	23	23	-	88	0	-	-0.89	NO	NO
2009	144	100	22	35.1	1	1	1	103	-	-	-13.1	NO	NO
2009	153	192	28	32.9	588	103	103	103	-	103	-4.9	NO	N
2009	162	159	31.3	29.3	1	-	-	41	41	41	2	YES	NO
2009	167	170	33.7	27.8	103	-	-	103	-	-	5.9	YES	YES
2009	169	233	18.3	27.7	103	103	0	-	103	103	-9.4	NO	NO
2009	170	169	27.8	18.3	103	-	-	103	103	0	9.5	YES	YES
2009	173	177	25.4	25.6	1	1	103	1	0	1	-0.2	NO	YES
2009	176	175	40	39.9	23	23	23	1	-	1	0.1	YES	NO
2009	177	174	25.6	37.3	1	0	1	1	1	-	-11.7	NO	YES
2009	186	184	28.3	19.6	103	-	-	103	103	0	8.7	YES	YES
2009	188	189	30.3	22.9	103	103	103	103	103	103	7.4	YES	YES
2009	189	190	22.9	37.8	103	103	103	103	-	-	-14.9	NO	YES
2009	192	180	32.9	21.2	103	-	103	103	103	296	11.7	YES	YES
2010	30	24	26.16	21.4	0	23	1	0	88	23	4.72	YES	NO
2010	35	247	40	28.4	23	8	-	-	23	23	11.6	YES	NO
2010	43	96	40	21.7	2	2	-	88	88	88	18.34	YES	NO
2010	99	144	24.2	22	103	103	-	1	1	1	2.2	YES	NO
2010	112	267	30.7	28.3	103	-	103	-	103	103	2.35	YES	NO
2010	114	112	40	38.9	625	1	-	103	-	103	1.1	YES	NO
2010	144	100	22	35.1	1	1	1	103	-	-	-13.1	NO	NO
2010	177	174	25.6	37.3	1	0	1	1	1	-	-11.7	NO	NO
2010	217	22	32.17	22.2	-	1	-	1	1	1	9.97	YES	YES
2010	238	78	28.22	24.1	-	7	628	9	9	-	4.18	YES	NO
2010	333	30	36.19	26.2	-	1003	-	0	23	1	10.03	YES	NO
2011	197	34	37	36	-	1	1	23	23	23	1	YES	NO
2011	297	298	38	32	-	-	103	-	-	103	6	YES	YES
2011	298	180	32	21.2	-	-	103	103	103	296	10.8	YES	NO
2011	303	169	37	18.3	-	-	103	103	103	0	18.7	YES	NO
2011	304	233	32	27.7	-	-	103	-	103	103	4.3	YES	YES
2011	316	315	35	29	-	-	628	-	-	26	6	YES	NO
2011	323	124	31	23	-	-	103	103	103	103	8	YES	YES

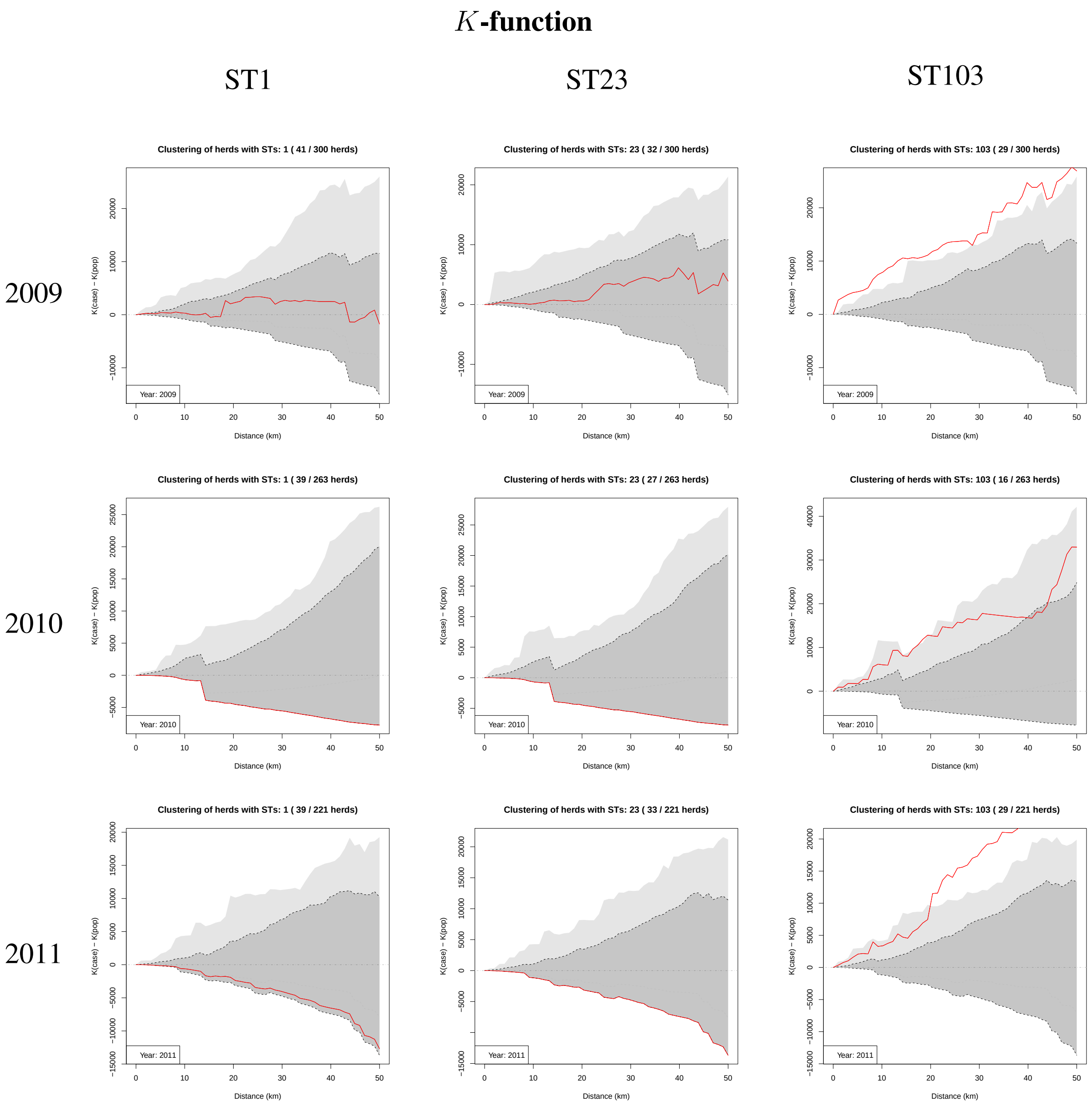

Figure A.1: Results of K-function test for predominant STs (ST1, ST23, ST103) in 2009, 2010 and 2011 using the conservative dataset (cases were defined by PCR only, no corrections for STs were made). Clustering is detected when the K-function estimate for the observed ST distribution (red line) is significantly higher than those for randomly generated ST distributions (shown in light gray, 95% confidence intervals are in dark gray).

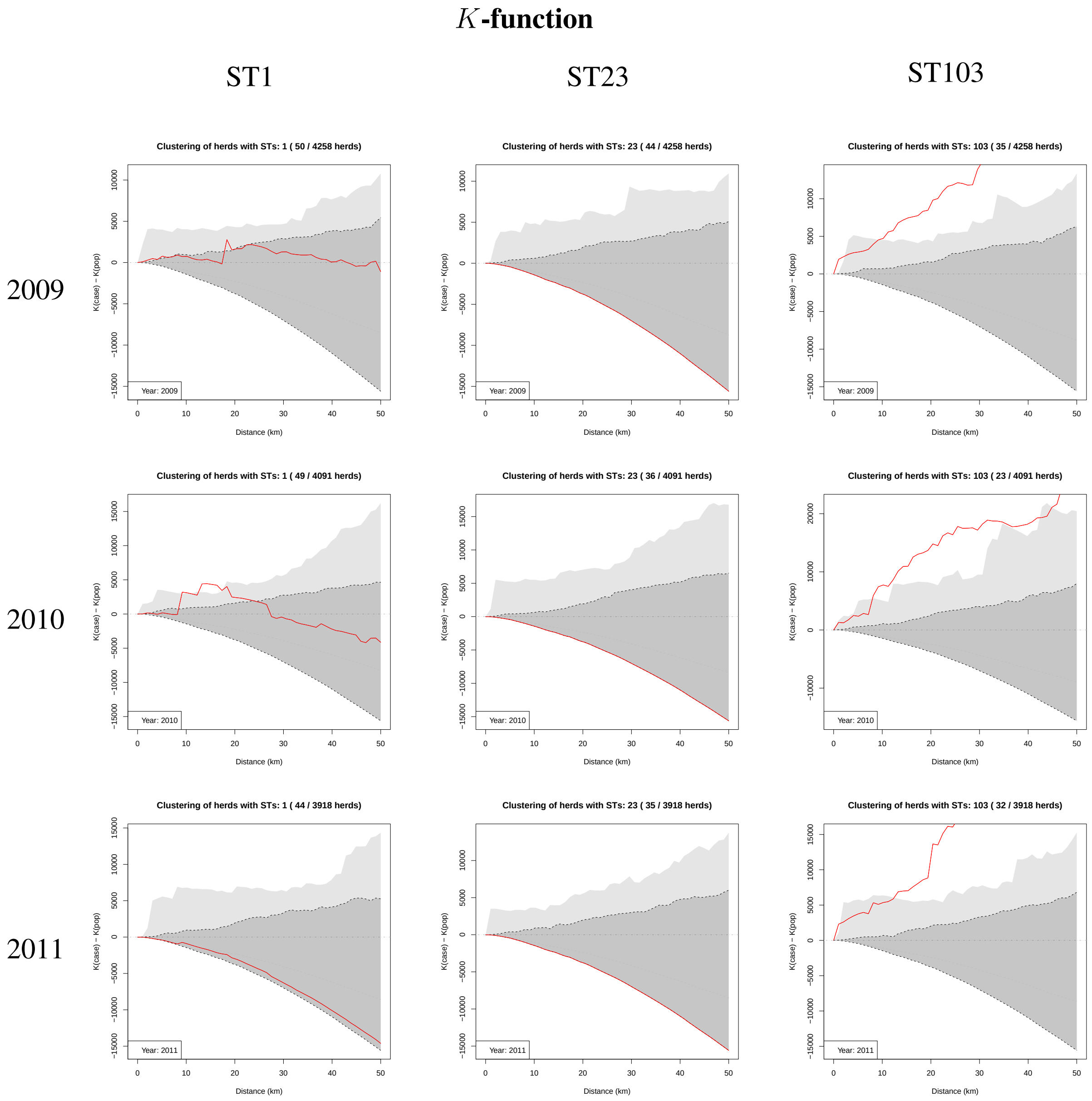

Figure A.2: Results of K-function test for predominant STs (ST1, ST23, ST103) in 2009, 2010 and 2011 using the total population at risk as denominator. Clustering is detected when the K-function estimate for the observed ST distribution (red line) is significantly higher than those for randomly generated ST distributions (shown in light gray, 95% confidence intervals are in dark gray).

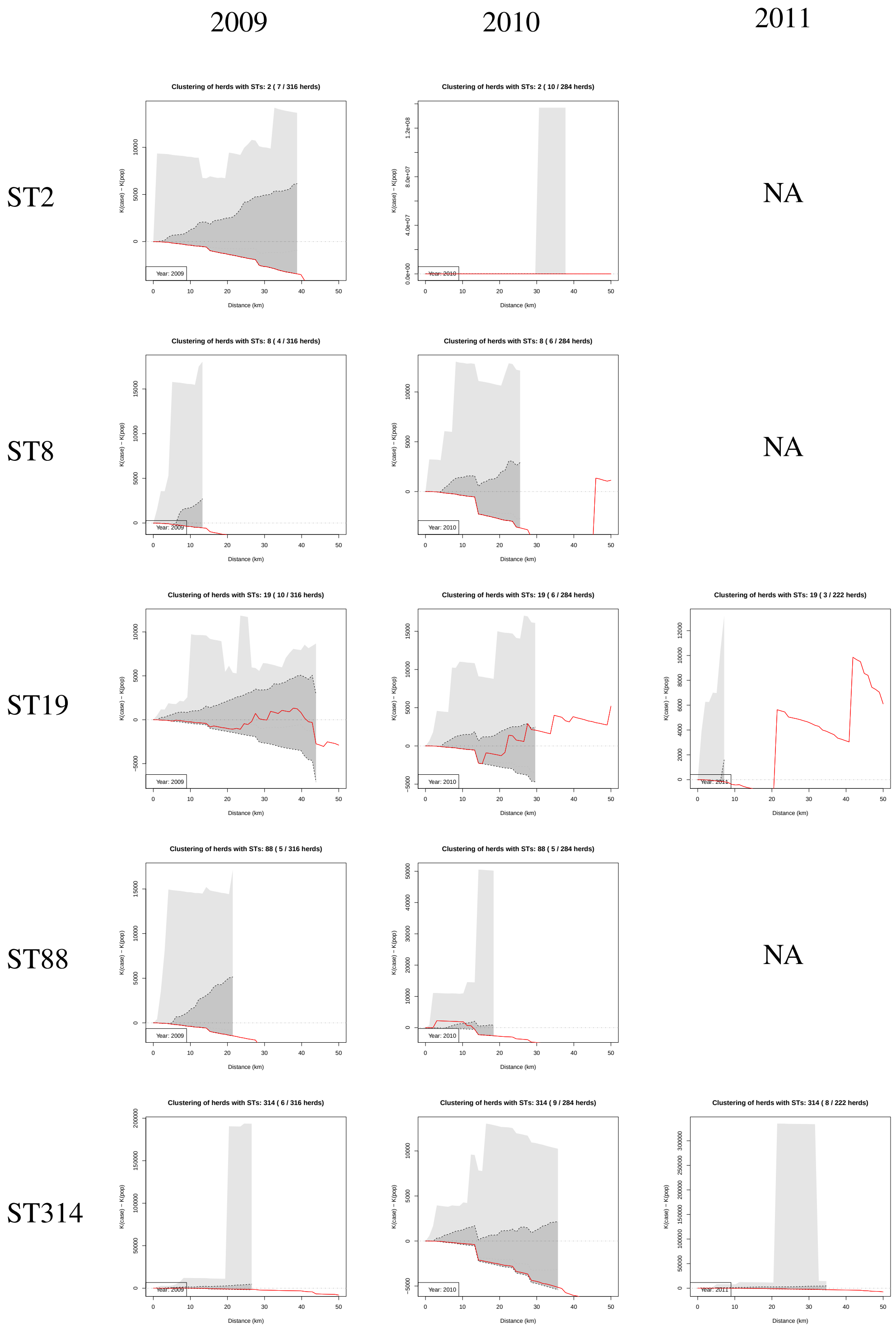

Figure A.3: Results of K-function method for less frequent sequence types. Clustering is detected when the estimation of K-function for the observed distribution of cases (red line) is significantly higher than for the randomly generated distributions (dark gray envelope). Some of the results are not present (NA) because the number of cases was insufficient to run the clustering algorithm.

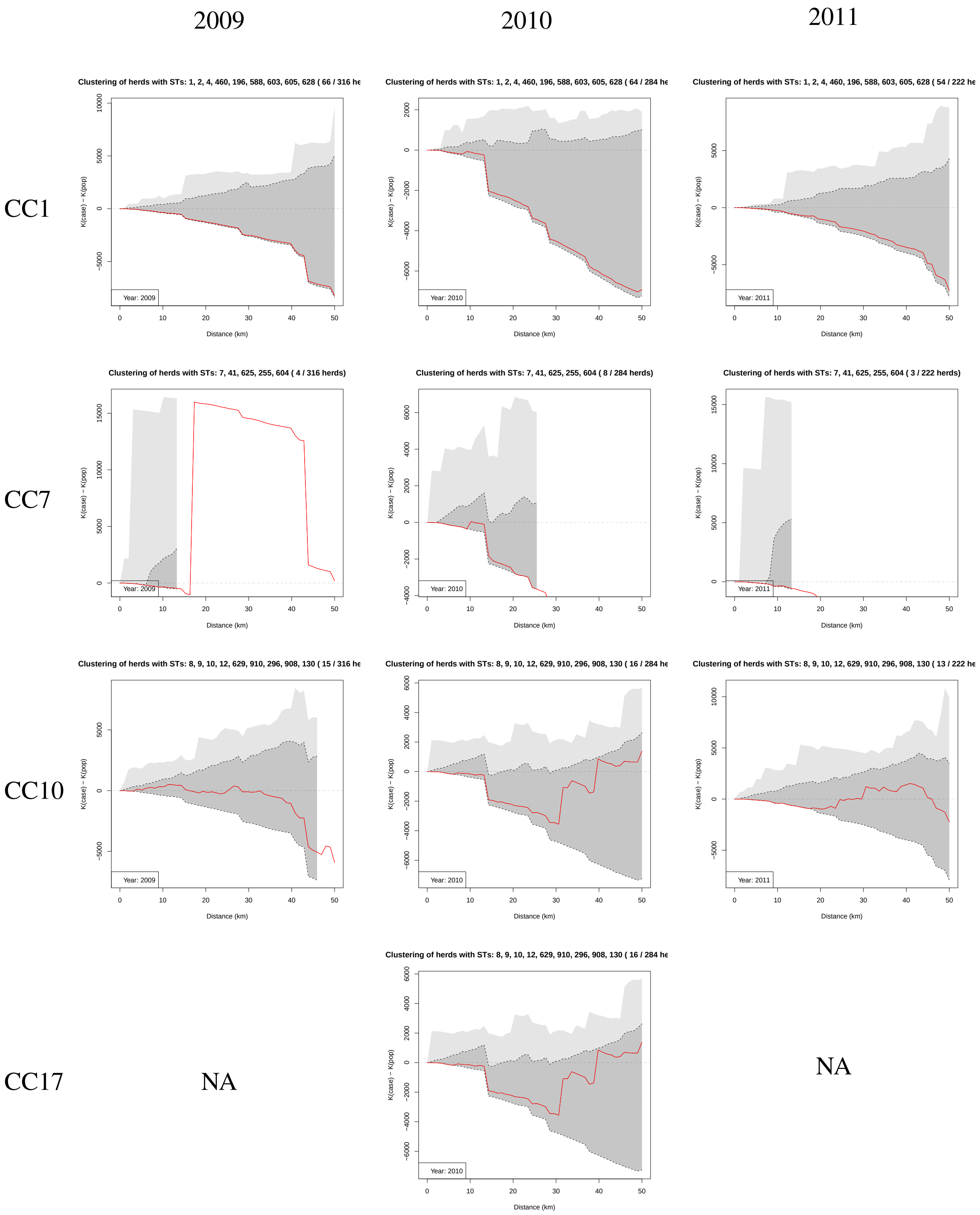

Figure A.4: Results of K-function method for clonal complexes. Clustering is detected when the estimation of K-function for the observed distribution of cases (red line) is significantly higher than for the randomly generated distributions (dark gray envelope). Some of the results are not present (NA) because the number of cases was insufficient to run the clustering algorithm.

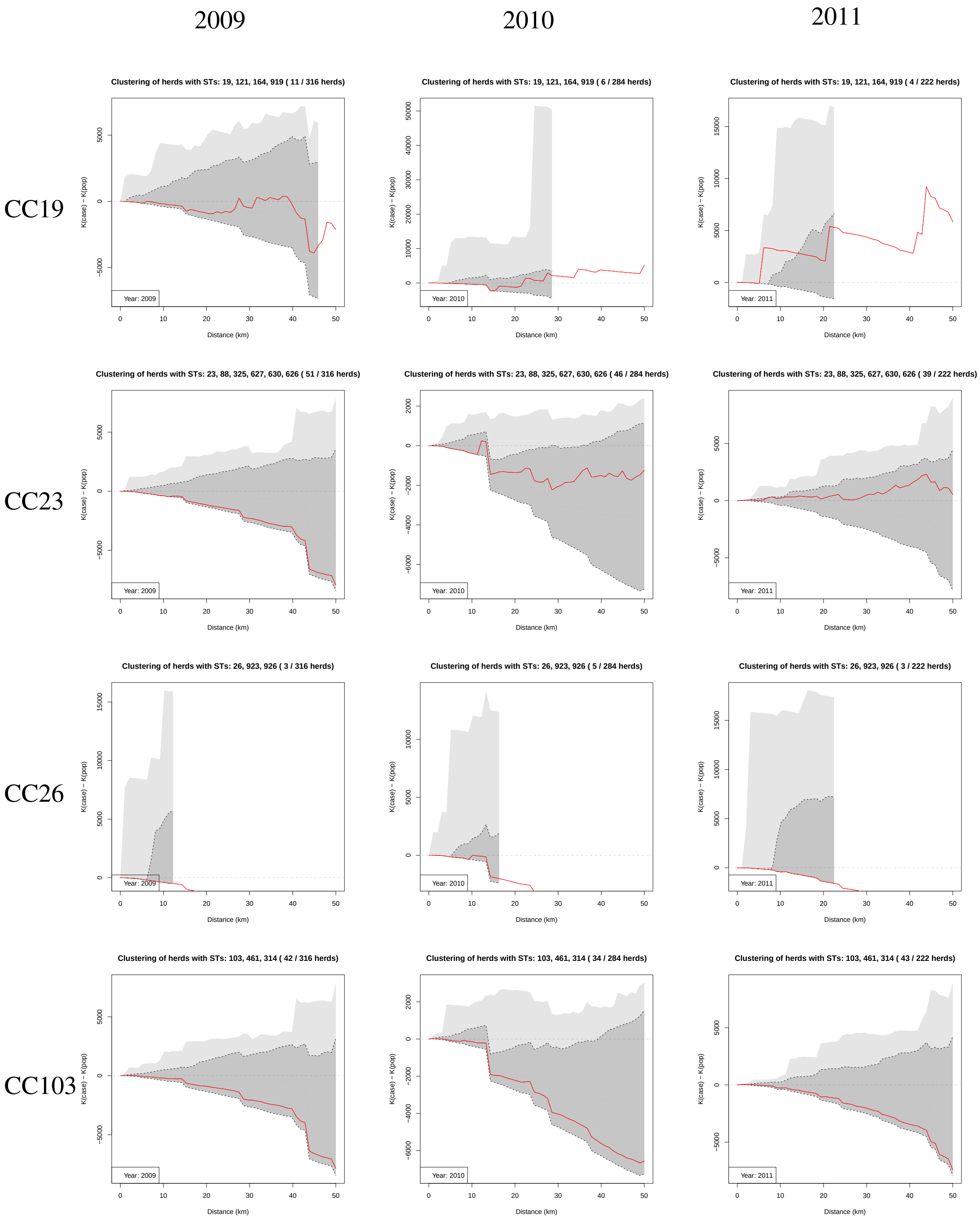

Figure A.5: Results of K-function method for clonal complexes. Clustering is detected when the estimation of K-function for the observed distribution of cases (red line) is significantly higher than for the randomly generated distributions (dark gray envelope).

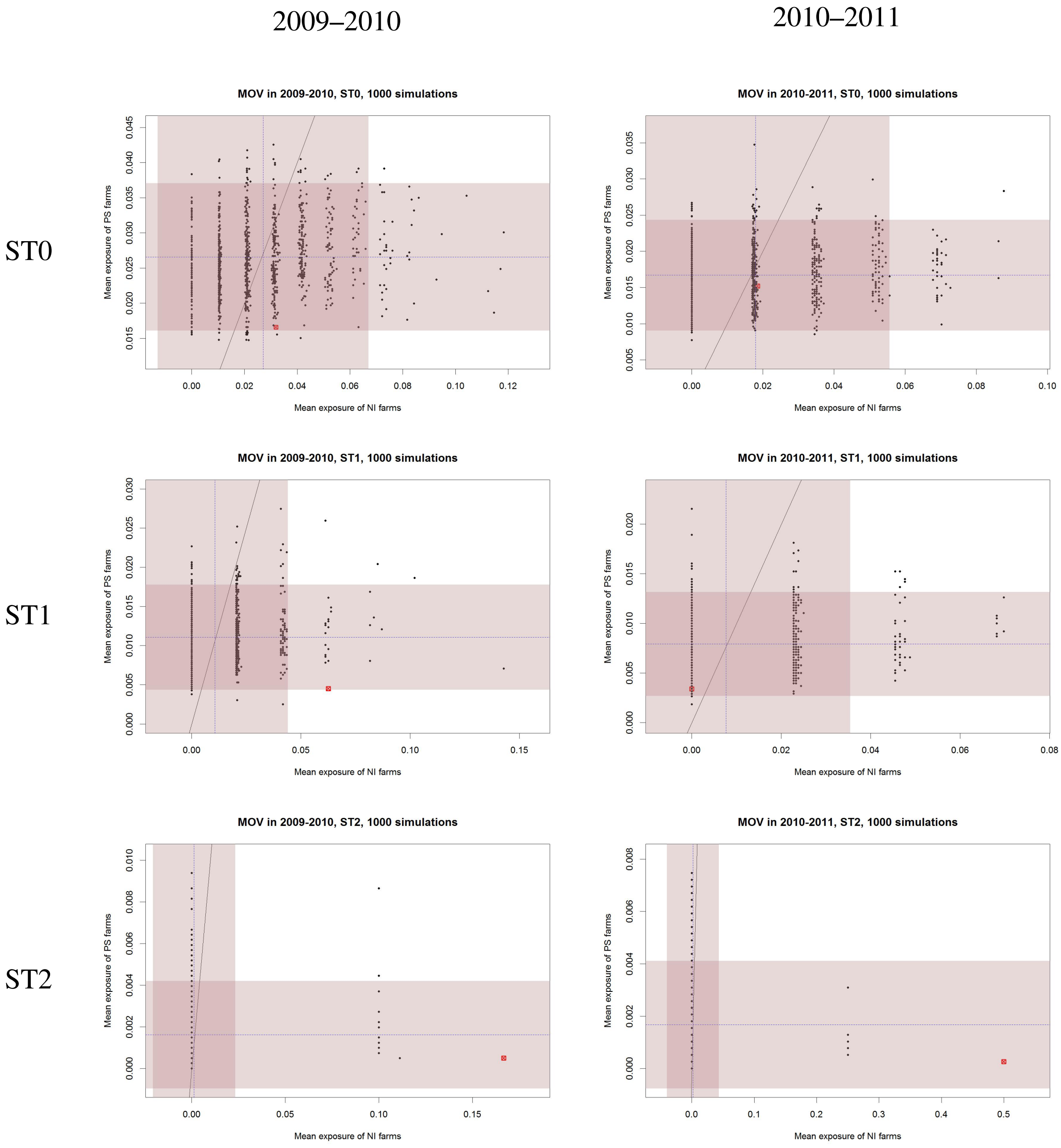

Figure A.6: ST specific mean exposure (E) of infection of newly infected (NI) and persistently susceptible (PS) herds for observed (red) movement networks and simulated random networks (black). Space below the diagonal line is where $E_{NI} > E_{PS}$. Areas in pale red are 95% confidence intervals for simulated E_{NI} and E_{PS}. Blue dashed lines indicate mean values of simulated E_{NI} and E_{PS}.

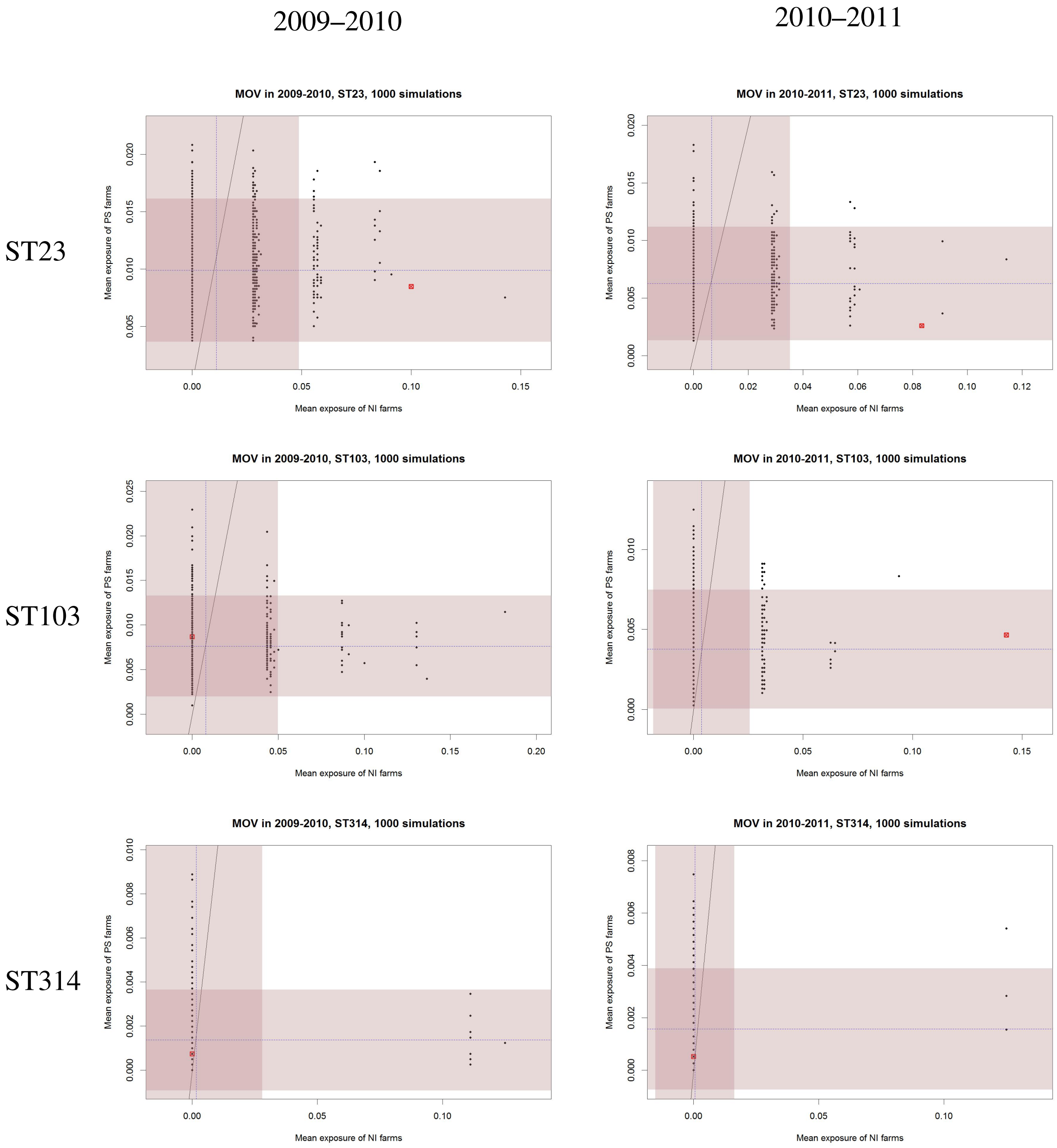

Figure A.7: ST specific mean exposure (E) of infection of newly infected (NI) and persistently susceptible (PS) herds for observed (red) movement networksand simulated random networks (black). Space below the diagonal line is where $E_{NI} > E_{PS}$. Areas in pale red are 95% confidence intervals for simulated E_{NI} and E_{PS}. Blue dashed lines indicate mean values of simulated E_{NI} and E_{PS}.

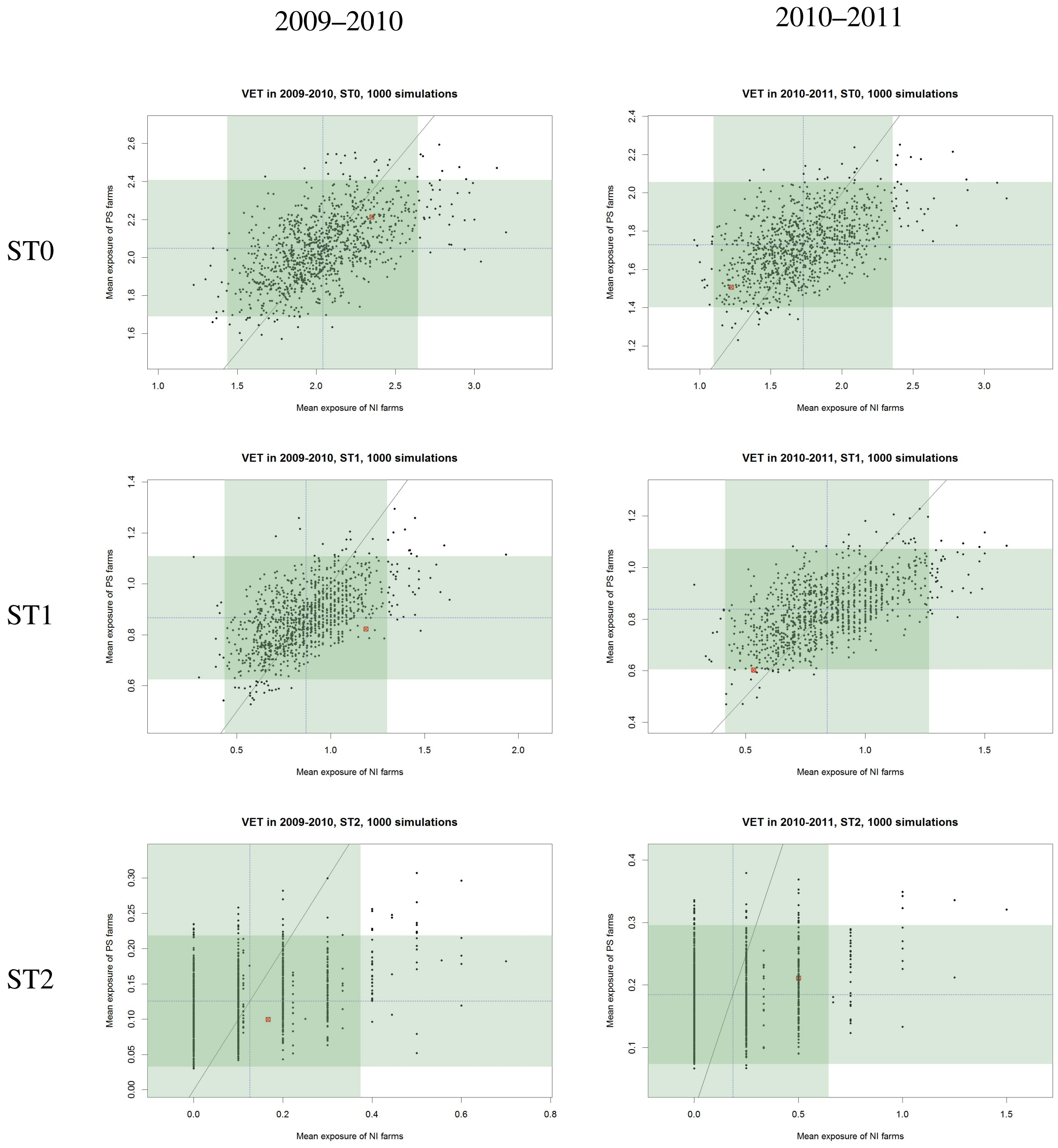

Figure A.8: ST specific mean exposure (E) of infection of newly infected (NI) and persistently susceptible (PS) herds for observed (red) veterinary networks and simulated random networks (black). Space below the diagonal line is where $E_{NI} > E_{PS}$. Areas in pale green are 95% confidence intervals for simulated E_{NI} and E_{PS}. Blue dashed lines indicate mean values of simulated E_{NI} and E_{PS}.

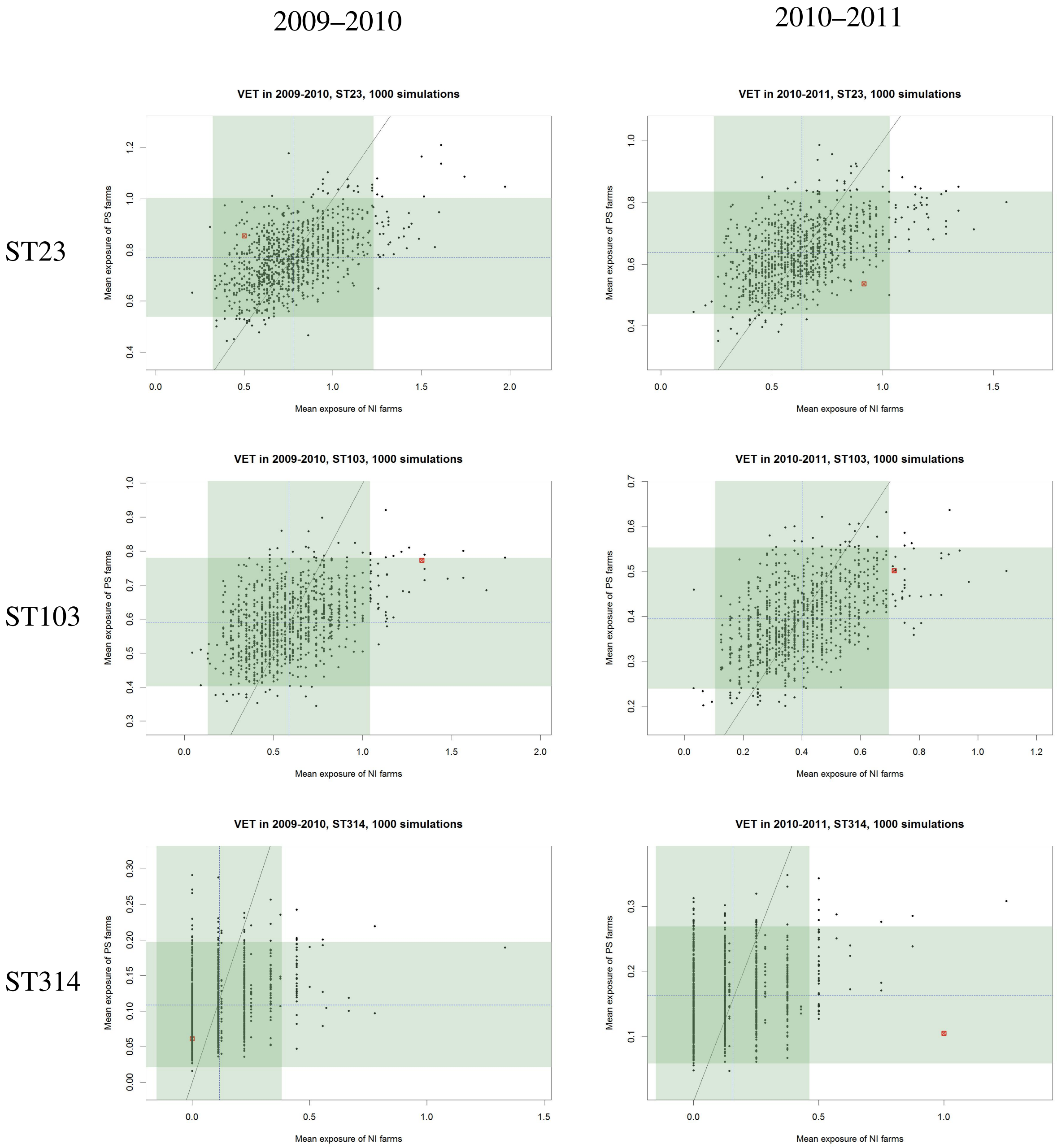

Figure A.9: ST specific mean exposure (E) of infection of newly infected (NI) and persistently susceptible (PS) herds for observed (red) veterinary networks and simulated random networks (black). Space below the diagonal line is where $E_{NI} > E_{PS}$. Areas in pale green are 95% confidence intervals for simulated E_{NI} and E_{PS}. Blue dashed lines indicate mean values of simulated E_{NI} and E_{PS}.

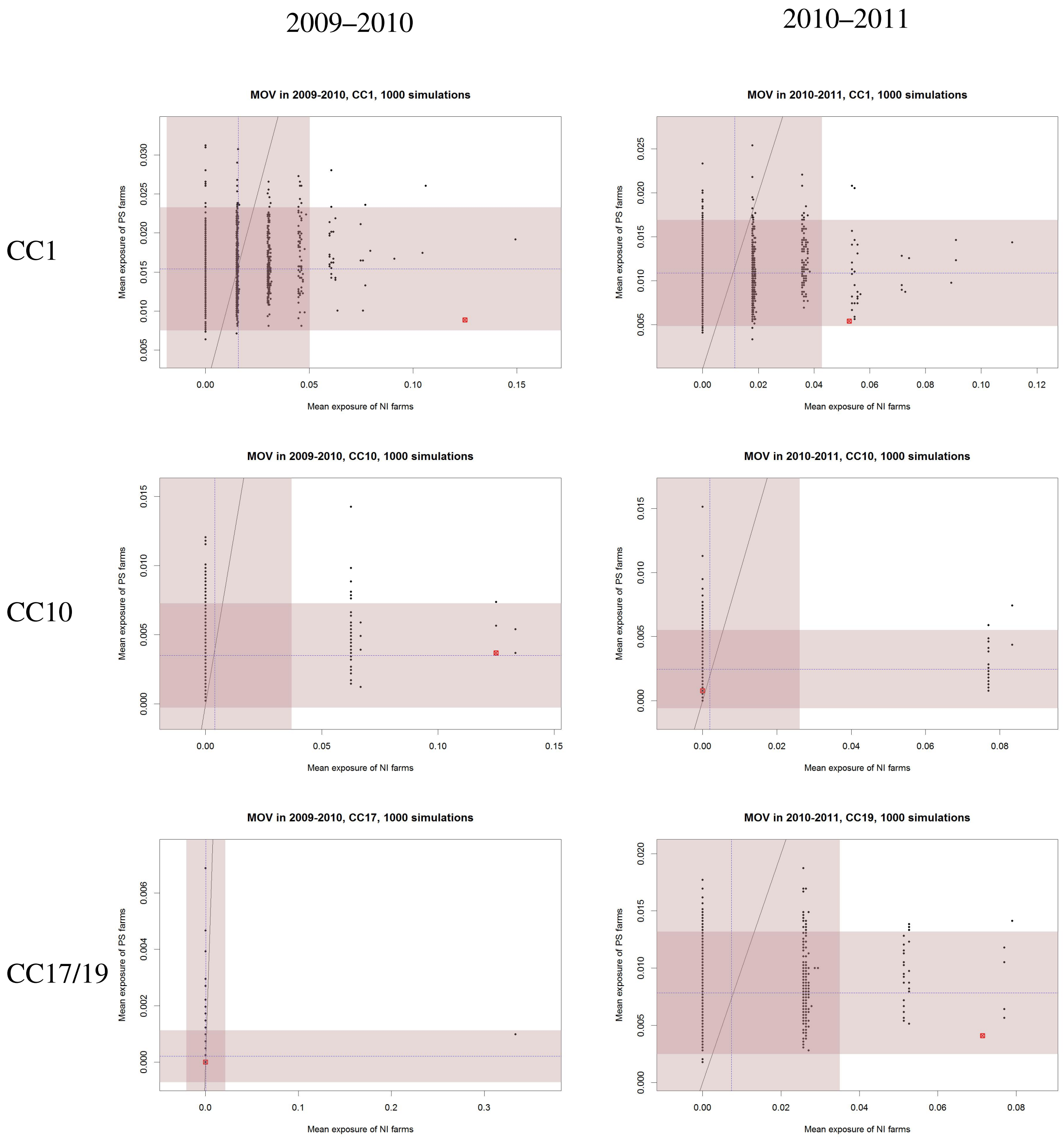

Figure A.10: CC specific mean exposure (E) of infection of newly infected (NI) and persistently susceptible (PS) herds for observed (red) movement networks and simulated random networks (black). Space below the diagonal line is where $E_{NI} > E_{PS}$. Areas in pale red are 95% confidence intervals for simulated E_{NI} and E_{PS}. Blue dashed lines indicate mean values of simulated E_{NI} and E_{PS}.

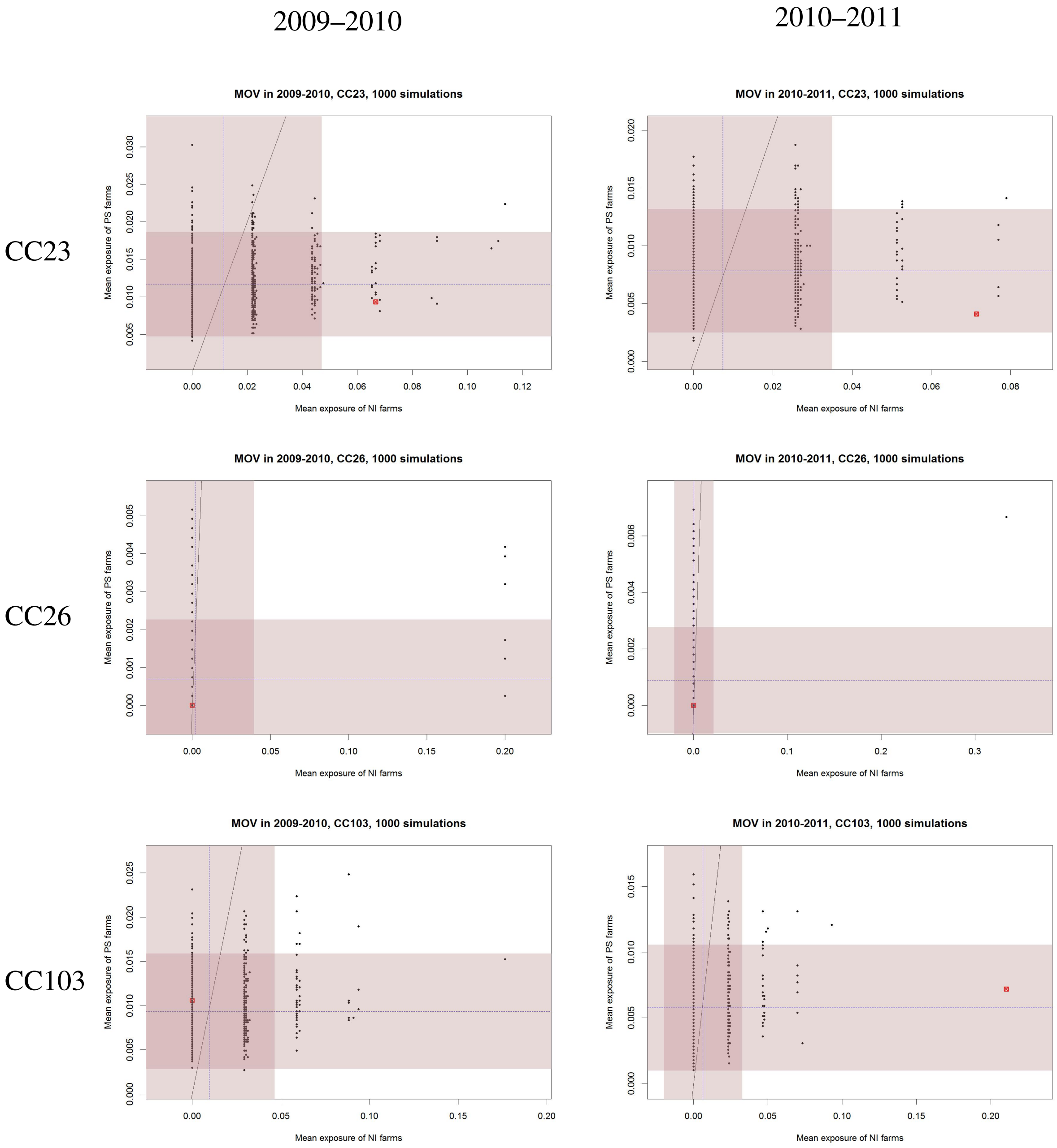

Figure A.11: CC specific mean exposure (E) of infection of newly infected (NI) and persistently susceptible (PS) herds for observed (red) movement networks and simulated random networks (black). Space below the diagonal line is where $E_{NI} > E_{PS}$. Areas in pale red are 95% confidence intervals for simulated E_{NI} and E_{PS}. Blue dashed lines indicate mean values of simulated E_{NI} and E_{PS}.

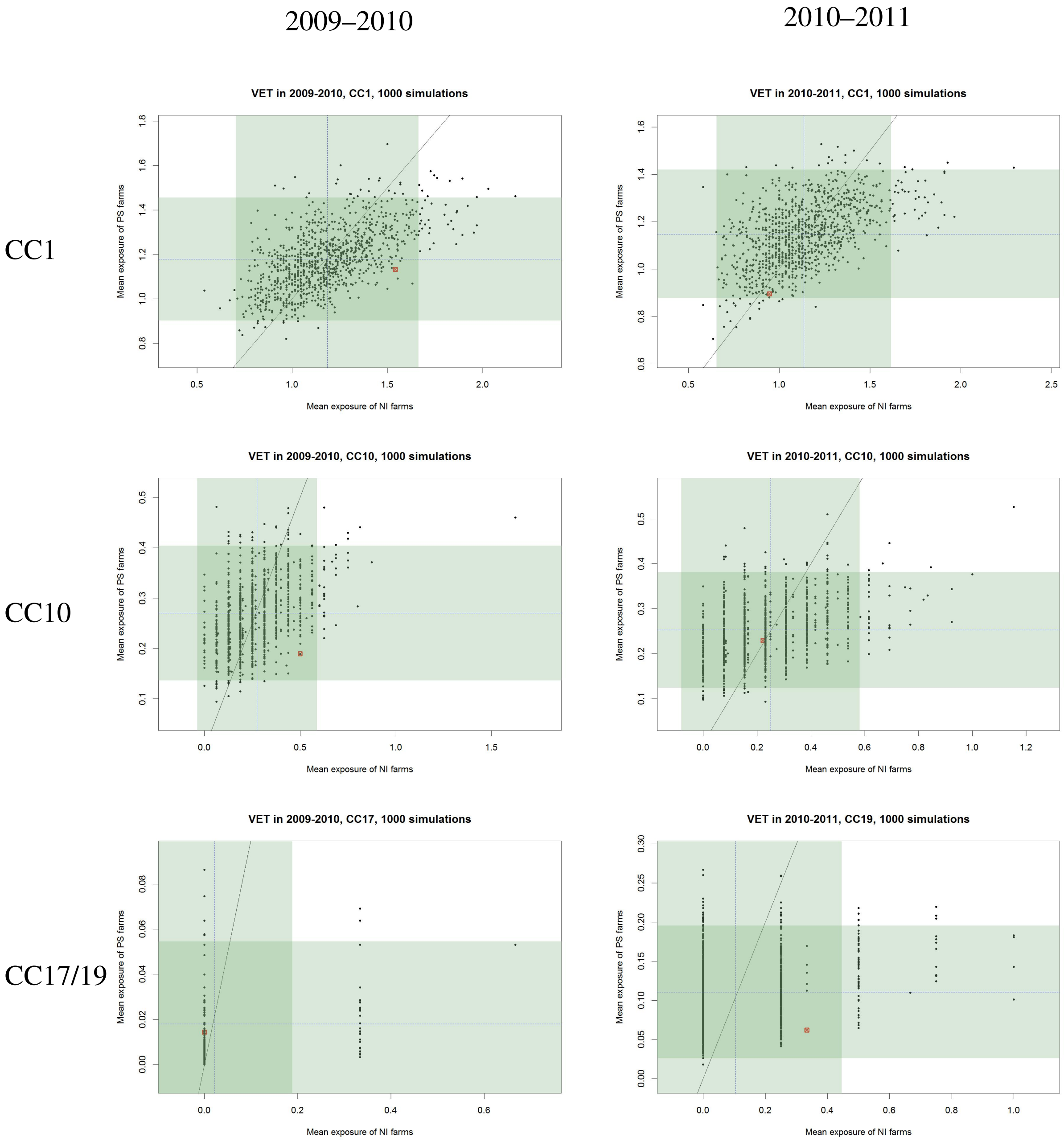

Figure A.12: CC specific mean exposure (E) of infection of newly infected (NI) and persistently susceptible (PS) herds for observed (red) veterinary networks and simulated random networks (black). Space below the diagonal line is where $E_{NI} > E_{PS}$. Areas in pale green are 95% confidence intervals for simulated E_{NI} and E_{PS}. Blue dashed lines indicate mean values of simulated E_{NI} and E_{PS}.

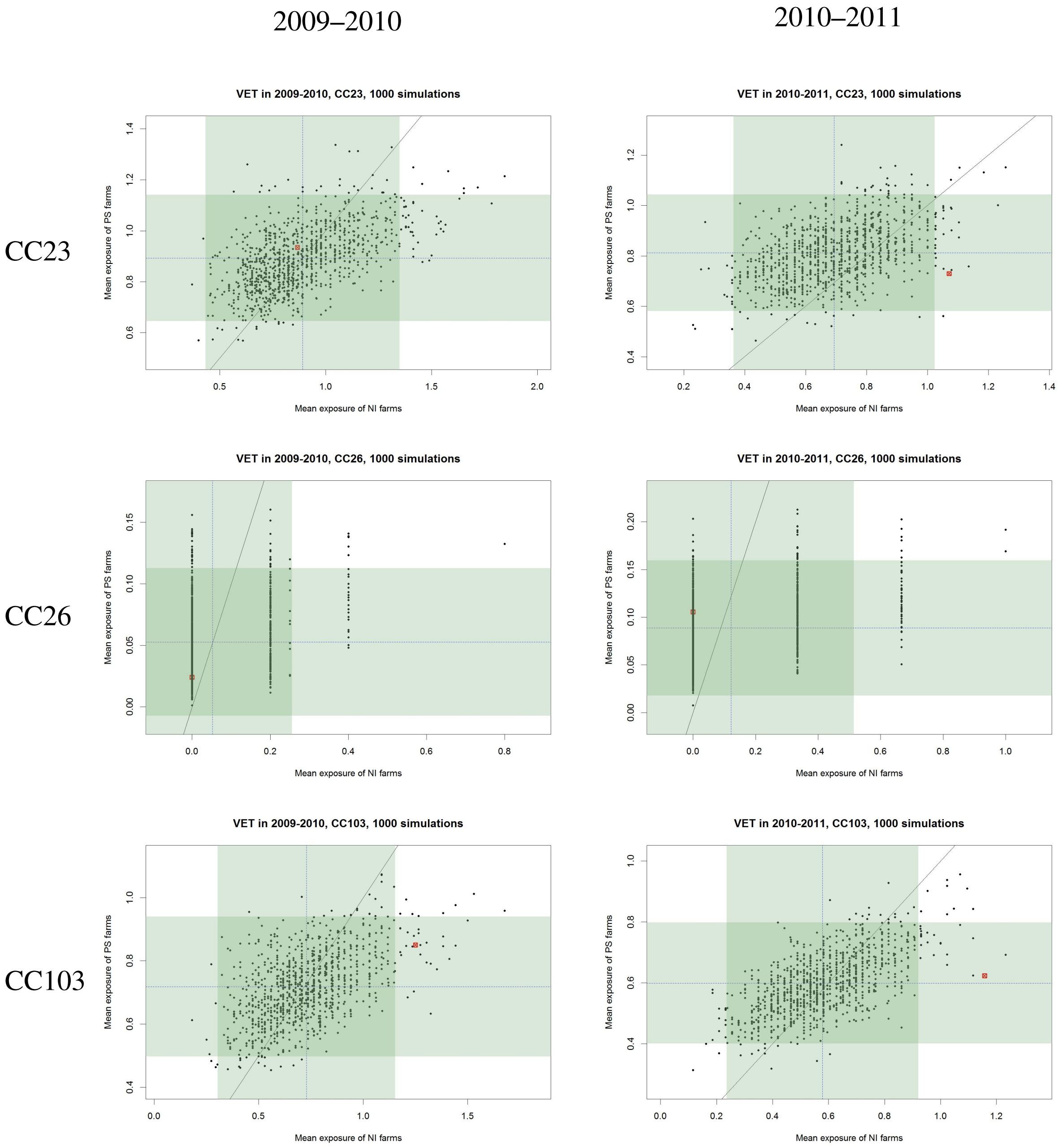

Figure A.13: CC specific mean exposure (E) of infection of newly infected (NI) and persistently susceptible (PS) herds for observed (red) veterinary networks and simulated random networks (black). Space below the diagonal line is where $E_{NI} > E_{PS}$. Areas in pale red are 95% confidence intervals for simulated E_{NI} and E_{PS}. Blue dashed lines indicate mean values of simulated E_{NI} and E_{PS}.

Bibliography

[Ahmed et al., 2010] Ahmed, S. S. U., Ersbøll, A. K. K., Biswas, P. K., and Christensen, J. P. (2010). The space-time clustering of highly pathogenic avian influenza (HPAI) H5N1 outbreaks in Bangladesh. *Epidemiology and Infection*, 138(6):843–852.

[Andersen et al., 2003] Andersen, H. J., Pedersen, L. H., Aarestrup, F. M., and Chriél, M. (2003). Evaluation of the surveillance program of *Streptococcus agalactiae* in Danish dairy herds. *Journal of Dairy Science*, 86(4):1233–1239.

[Bachrach, 1968] Bachrach, H. L. (1968). Foot-and-mouth disease. *Annual Review of Microbiology*, 22(V):201–244.

[Baddeley and Turner, 2005] Baddeley, A. and Turner, R. (2005). *spatstat*: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12:1–42.

[Baddeley et al., 2000] Baddeley, A. J., Møller, J., and Waagepetersen, R. (2000). Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica*, 54(3):329–350.

[Bisharat et al., 2004] Bisharat, N., Crook, D. W., Leigh, J., Harding, R. M., Ward, P. N., Coffey, T. J., Maiden, M. C., Peto, T., and Jones, N. (2004). Hyperinvasive neonatal Group B Streptococcus has arisen from a bovine ancestor. *Journal of Clinical Microbiology*, 42(5):2161–2167.

[Blancou, 2002] Blancou, J. (2002). History of the control foot and mouth disease. *Comparative Immunology, Microbiology and Infectious Diseases*, 25(5-6):283–296.

[Boers et al., 2012] Boers, S. A., van der Reijden, W. A., and Jansen, R. (2012). High-throughput multilocus sequence typing: bringing molecular typing to the next level. *PLoS ONE*, 7(7):e39630.

[Böhm et al., 2009] Böhm, M., Hutchings, M. R., and White, P. C. L. (2009). Contact networks in a wildlife-livestock host community: Identifying high-risk individuals in the transmission of bovine TB among badgers and cattle. *PLoS ONE*, 4(4):e5016.

[Bohnsack et al., 2004] Bohnsack, J. F., Whiting, A. A., Martinez, G., Jones, N., Adderson, E. E., Detrick, S., Blaschke-bonkowsky, A. J., Bisharat, N., and Gottschalk, M. (2004). Serotype III Streptococcus agalactiae from bovine milk and human neonatal infections. In *Emerging Infectious Diseases*, volume 10, pages 1412–1419.

[Boklund et al., 2013] Boklund, A., Halasa, T., Christiansen, L. E., and Enøe, C. (2013). Comparing control strategies against foot-and-mouth disease: Will vaccination be cost-effective in Denmark? *Preventive Veterinary Medicine*, 111(3-4):206–219.

[Carpenter, 2001] Carpenter, T. E. (2001). Methods to investigate spatial and temporal clustering in veterinary epidemiology. *Preventive Veterinary Medicine*, 48(4):303–320.

[Cottam, 2007] Cottam, E. M. (2007). *Micro-evolution of Foot-and-mouth disease virus*. PhD thesis.

[Cottam et al., 2008] Cottam, E. M., Thébaud, G., Wadsworth, J., Gloster, J., Mansley, L. M., Paton, D. J., King, D. P., and Haydon, D. T. (2008). Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proceedings. Biological sciences / The Royal Society*, 275(1637):887–895.

[Csárdi and Nepusz, 2006] Csárdi, G. and Nepusz, T. (2006). The *igraph* software package for complex network research. *InterJournal*, (Complex Systems):1695.

[Cuzick and Edwards, 1990] Cuzick, J. and Edwards, R. (1990). Spatial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(1):73–104.

[Dogan et al., 2005] Dogan, B., Schukken, Y. H., Santisteban, C., and Boor, K. J. (2005). Distribution of serotypes and antimicrobial resistance genes among *Streptococcus agalactiae* isolates from bovine and human hosts. *Journal of Clinical microbiology*, 43(12):5899.

[Domingo et al., 2002] Domingo, E., Baranowski, E., Escarmis, C., and Sobrino, F. (2002). Foot-and-mouth disease virus. *Comparative Immunology, Microbiology and Infectious Diseases*, 25(5-6):297–308.

[Domingo et al., 2003] Domingo, E., Escarmis, C., and Baranowski, E. (2003). Evolution of foot-and-mouth disease virus. *Virus Research*, 8(6):786–798.

[Dubé et al., 2009] Dubé, C., Ribble, C., Kelton, D., and McNab, B. (2009). A review of network analysis terminology and its application to foot-and-mouth disease modelling and policy development. *Transboundary and Emerging Diseases*, 56(3):73–85.

[Enright et al., 2001] Enright, M. C., Spratt, B. G., Kalia, a., Cross, J. H., and Bessen, D. E. (2001). Multilocus sequence typing of *Streptococcus pyogenes* and the relationships between *emm* type and clone. *Infection and Immunity*, 69(4):2416–2427.

[Ersbøll and Ersbøll, 2009] Ersbøll, A. K. and Ersbøll, B. K. (2009). Simulation of the K-function in the analysis of spatial clustering for non-randomly distributed locations–exemplified by bovine virus diarrhoea virus (BVDV) infection in Denmark. *Preventive Veterinary Medicine*, 91(1):64–71.

[FAO, 2013] FAO (2013). World Livestock 2013 — Changing disease landscapes. Technical report, Rome.

[Feil et al., 2004] Feil, E. J., Li, B. C., Aanensen, D. M., Hanage, W. P., and Spratt, B. G. (2004). eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *Journal of Bacteriology*, 186(5):1518–1530.

[Feil and Spratt, 2001] Feil, E. J. and Spratt, B. G. (2001). Recombination and the population structures of bacterial pathogens. *Annual Review of Microbiology*, 55:561–590.

[Fenton et al., 2009] Fenton, S. E., Clough, H. E., Diggle, P. J., Evans, S. J., Davison, H. C., Vink, W. D., and French, N. P. (2009). Spatial and spatio-temporal analysis of *Salmonella* infection in dairy herds in England and Wales. *Epidemiology and Infection*, 137(6):847–857.

[Ferguson et al., 2001a] Ferguson, N. M., Donnelly, C. A., and Anderson, R. M. (2001a). The foot-and-mouth epidemic in Great Britain: pattern of spread and impact of interventions. *Science (New York, N.Y.)*, 292(5519):1155–1160.

[Ferguson et al., 2001b] Ferguson, N. M., Donnelly, C. A., and Anderson, R. M. (2001b). Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain. *Nature*, 413(6855):542–548.

[García-Álvarez et al., 2011] García-Álvarez, L., Holden, M. T. G., Lindsay, H., Webb, C. R., Brown, D. F. J., Curran, M. D., Walpole, E., Brooks, K., Pickard, D. J., Teale, C., Parkhill, J., Bentley, S. D., Edwards, G. F., Girvan, E. K., Kearns, A. M., Pichon, B., Hill, R. L. R., Larsen, A. R., Skov, R. L., Peacock, S. J., Maskell, D. J., and Holmes, M. a. (2011). Meticillin-resistant *Staphylococcus aureus* with a novel mecA homologue in human and bovine populations in the UK and Denmark: A descriptive study. *The Lancet Infectious Diseases*, 11(8):595–603.

[García Álvarez et al., 2011] García Álvarez, L., Webb, C. R., Holmes, M. A., Álvarez, L. G., Webb, C. R., Holmes, M. a., García Álvarez, L., Webb, C. R., Holmes, M. a., Álvarez, L. G., Webb, C. R., and Holmes, M. a. (2011). A novel field-based approach to validate the use of network models for disease spread between dairy herds. *Epidemiology and Infection*, 139(12):1863–1874.

[Gerbier and Chadoeuf, 2000] Gerbier, G. and Chadoeuf, J. (2000). Appplication of bivariate point pattern analysis methods for describing spatial distribution of foot-and-mouth disease. *Proceedings of the 9th International Symposium on Veterinary Epidemiology and Economics*, 1:2–4.

[Gibbens et al., 2001] Gibbens, J. C., Sharpe, C. E., Wilesmith, J. W., Mansley, L. M., Michalopoulou, E., Ryan, J. B., and Hudson, M. (2001). Descriptive epidemiology of the 2001 foot-and-mouth disease epidemic in Great Britain: the first five months. *Veterinary Record*, 149(24):729–743.

[Gibbens and Wilesmith, 2002] Gibbens, J. C. and Wilesmith, J. W. (2002). Temporal and geographical distribution of cases of foot-and-mouth disease during the early weeks of the 2001 epidemic in Great Britain. *Veterinary Record*, 151(14):407–412.

[Glaser et al., 2002] Glaser, P., Rusniok, C., Buchrieser, C., Chevalier, F., Frangeul, L., Msadek, T., Zouine, M., Couvé, E., Lalioui, L., Poyart, C., Trieu-Cuot, P., and Kunst, F. (2002). Genome sequence of *Streptococcus agalactiae*, a pathogen causing invasive neonatal disease. *Molecular Microbiology*, 45:1499–1513.

[Green et al., 2006] Green, D. M., Kiss, I. Z., and Kao, R. R. (2006). Modelling the initial spread of foot-and-mouth disease through animal movements. *Proceedings. Biological sciences / The Royal Society*, 273(1602):2729–2735.

[Halasa et al., 2007] Halasa, T., Huijps, K., Ø sterås, O., and Hogeveen, H. (2007). Economic effects of bovine mastitis and mastitis management: a review. *The Veterinary Quarterly*, 29(1):18–31.

[Haydon et al., 2003] Haydon, D. T., Chase-Topping, M. E., Shaw, D. J., Matthews, L., Friar, J. K., Wilesmith, J. W., and Woolhouse, M. E. J. (2003). The construction and analysis of epidemic trees with reference to the 2001 UK foot-and-mouth outbreak. *Proceedings. Biological sciences / The Royal Society*, 270(1511):121–127.

[Haydon et al., 2004] Haydon, D. T., Kao, R. R., and Kitching, R. P. (2004). The UK foot-and-mouth disease outbreak - the aftermath. *Nature reviews: Microbiology*, 2(8):675–681.

[Hutber et al., 2011] Hutber, A. M., Kitching, R. P., Fishwick, J. C., and Bires, J. (2011). Foot-and-mouth disease: the question of implementing vaccinal control during an epidemic. *Veterinary Journal*, 188(1):18–23.

[Jamal et al., 2011] Jamal, S. M., Ferrari, G., Ahmed, S., Normann, P., and Belsham, G. J. (2011). Genetic diversity of foot-and-mouth disease virus serotype O in Pakistan and Afghanistan, 1997-2009. *Infection, Genetics and Evolution*, 11(6):1229–1238.

[James and Rushton, 2002] James, A. D. and Rushton, J. (2002). The economics of foot and mouth disease. *Revue scientifique et technique (International Office of Epizootics)*, 21(3):637–644.

[Jolley and Maiden, 2010] Jolley, K. A. and Maiden, M. C. J. (2010). BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*, 11(1):595.

[Jombart et al., 2014] Jombart, T., Cori, A., Didelot, X., Cauchemez, S., Fraser, C., and Ferguson, N. (2014). Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Computational Biology*, 10(1):e1003457.

[Jombart et al., 2011] Jombart, T., Eggo, R. M., Dodd, P. J., and Balloux, F. (2011). Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity*, 106(2):383–390.

[Jones et al., 2003] Jones, N., Bohnsack, J. F., Takahashi, S., Karen, A., Chan, M.-s. S., Kunst, F., Glaser, P., Rusniok, C., Crook, D. W. M., Rosalind, M., Bisharat, N., Spratt, B. G., Oliver, K. A., Harding, R. M., Chan, M.-s. S., Kunst, F., Glaser, P., Rusniok, C., Crook, D. W. M., Harding, R. M., Bisharat, N., and Spratt, B. G. (2003). Multilocus sequence typing system for group B streptococcus. *Journal of Clinical Microbiology*, 41:2530–2536.

[Kao, 2002] Kao, R. R. (2002). The role of mathematical modelling in the control of the 2001 FMD epidemic in the UK. *Trends in Microbiology*, 10(6):279–86.

[Kao, 2003] Kao, R. R. (2003). The impact of local heterogeneity on alternative control strategies for foot-and-mouth disease. *Proceedings. Biological sciences / The Royal Society*, 270(1533):2557–2564.

[Kao et al., 2006] Kao, R. R., Danon, L., Green, D. M., and Kiss, I. Z. (2006). Demographic structure and pathogen dynamics on the network of livestock movements in Great Britain. *Proceedings. Biological sciences / The Royal Society*, 273(1597):1999–2007.

[Kao et al., 2007] Kao, R. R., Green, D. M., Johnson, J., and Kiss, I. Z. (2007). Disease dynamics over very different time-scales: foot-and-mouth disease and scrapie on the network of livestock movements in the UK. *Journal of the Royal Society, Interface / the Royal Society*, 4:907–916.

[Kao et al., 2008] Kao, R. R., O'Reilly, K., Bronsvoort, B., Handel, I., Zadoks, R. N., Willoughby, K., Milne, C., and Gunn, G. (2008). Assessing the additional risk of FMD transmission from using field lairages at Scottish livestock markets. Technical report, Report to Animal Health and Welfare Division of the Scottish Government's Rural and Environment Research and Analysis Directorate.

[Katholm, 2010] Katholm, J. (2010). *Streptococcus agalactiae* — an increasing problem in Scandinavia. In *Proceedings The Nordic Dairy Association's Committee for Milk Quality, Mastitis symposium, Rebild, Denmark*, volume 9, pages 1–12.

[Keefe, 1997] Keefe, G. P. (1997). *Streptococcus agalactiae* mastitis: a review. *The Canadian Veterinary Journal. La Revue Vétérinaire Canadienne*, 38(7):429–437.

[Keeling et al., 2001] Keeling, M. J., Woolhouse, M. E. J., Shaw, D. J., Matthews, L., Chase-Topping, M. E., Haydon, D. T., Cornell, S. J., Kappey, J., Wilesmith, J. W., and Grenfell, B. T. (2001). Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. *Science (New York, N.Y.)*, 294(5543):813–817.

[Kiss et al., 2006] Kiss, I. Z., Green, D. M., and Kao, R. R. (2006). The network of sheep movements within Great Britain: Network properties and their implications for infectious disease spread. *Journal of the Royal Society, Interface / the Royal Society*, 3(10):669–677.

[Kitching and Hughes, 2002] Kitching, R. P. and Hughes, G. J. (2002). Clinical variation in foot and mouth disease: sheep and goats. *Revue scientifique et technique (International Office of Epizootics)*, 21(3):505–512.

[Knight-Jones and Rushton, 2013] Knight-Jones, T. J. D. and Rushton, J. (2013). The economic impacts of foot and mouth disease - what are they, how big are they and where do they occur? *Preventive Veterinary Medicine*, 112(3-4):162–173.

[Koskinen et al., 2009] Koskinen, M. T., Holopainen, J., Pyörälä, S., Bredbacka, P., Pitkälä, A., Barkema, H. W., Bexiga, R., Roberson, J., Sø lverød, L., Piccinini, R., Kelton, D., Lehmusto, H., Niskala, S., and Salmikivi, L. (2009). Analytical specificity and sensitivity of a real-time polymerase chain reaction assay for identification of bovine mastitis pathogens. *Journal of Dairy Science*, 92(3):952–959.

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

[Leforban and Gerbier, 2002] Leforban, Y. and Gerbier, G. (2002). Review of the status of foot and mouth disease and approach to control/eradication in Europe and Central Asia. *Revue Scientifique et Technique (International Office of Epizootics)*, 21(3):477–492.

[Ma et al., 2011] Ma, L.-N., Zhang, J., Chen, H.-T., Zhou, J.-H., Ding, Y.-Z., and Liu, Y.-S. (2011). An overview on ELISA techniques for FMD. *Virology Journal*, 8(1):419.

[Malirat et al., 2011] Malirat, V., Bergmann, I. E., de Mendonça Campos, R., Salgado, G., Sánchez, C., Conde, F., Quiroga, J. L., and Ortiz, S. (2011). Phylogenetic analysis of foot-and-mouth disease virus type O circulating in the Andean region of South America during 2002-2008. *Veterinary Microbiology*, 152(1-2):74–87.

[Manning et al., 2010] Manning, S. D., Springman, A. C., Million, A. D., Milton, N. R., McNamara, S. E., Somsel, P. A., Bartlett, P., and Davies, H. D. (2010). Association of Group B Streptococcus colonization and bovine exposure: a prospective cross-sectional cohort study. *PLoS ONE*, 5(1):e8795.

[Mardones et al., 2013] Mardones, F. O., Zu Donha, H., Thunes, C., Velez, V., and Carpenter, T. E. (2013). The value of animal movement tracing: a case study simulating the spread and control of foot-and-mouth disease in California. *Preventive veterinary medicine*, 110(2):133–138.

[Martinez et al., 2000] Martinez, G., Harel, J., Higgins, R., Daignault, D., Gottschalk, M., and Lacouture, S. (2000). Characterization of *Streptococcus agalactiae* isolates of bovine and human origin by randomly amplified polymorphic DNA analysis. *Journal of Clinical Microbiology*, 38(1):71–78.

[Matthews et al., 2003] Matthews, L., Haydon, D. T., Shaw, D. J., Chase-Topping, M. E., Keeling, M. J., and Woolhouse, M. E. J. (2003). Neighbourhood control policies and the spread of infectious diseases. *Proceedings. Biological sciences / The Royal Society*, 270(1525):1659–1666.

[Morelli et al., 2012] Morelli, M. J., Thébaud, G., Chadœuf, J., King, D. P., Haydon, D. T., and Soubeyrand, S. (2012). A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS Computational Biology*, 8(11):e1002768.

[Mweu, 2013] Mweu, M. M. (2013). Streptococcus agalactiae *infection in the population of Danish dairy cattle herds: An epidemiological inquiry*. PhD thesis, University of Copenhagen.

[Mweu et al., 2014] Mweu, M. M., Nielsen, S. r. S., Halasa, T., and Toft, N. (2014). Spatiotemporal patterns, annual baseline and movement-related incidence of *Streptococcus agalactiae* infection in Danish dairy herds: 2000-2009. *Preventive Veterinary Medicine*, 113(2):219–230.

[Mweu et al., 2012] Mweu, M. M., Toft, N., Katholm, J., and Nielsen, S. S. (2012). Evaluation of two herd-level diagnostic tests for *Streptococcus agalactiae* using a latent class approach. *Veterinary Microbiology*, 159(1-2):181–186.

[Neave et al., 1969] Neave, F. K., Dodd, F. H., Kingwill, R. G., and Westgarth, D. R. (1969). Control of mastitis in the dairy herd by hygiene and management. *Journal of Dairy Science*, 52(5):696–707.

[Nöremark et al., 2011] Nöremark, M., Håkansson, N., Lewerin, S. S., Lindberg, A., and Jonsson, A. (2011). Network analysis of cattle and pig movements in Sweden: measures relevant for disease control and risk based surveillance. *Preventive Veterinary Medicine*, 99(2-4):78–90.

[Oliveira et al., 2006] Oliveira, I. C. M., de Mattos, M. C., Pinto, T. A., Ferreira-Carvalho, B. T., Benchetrit, L. C., Whiting, A. A., Bohnsack, J. F., and Figueiredo, A. M. S. (2006). Genetic relatedness between group B streptococci originating from bovine mastitis and a human group B streptococcus type V cluster displaying an identical pulsed-field gel electrophoresis pattern. *Clinical Microbiology and Infection*, 12:887–893.

[Olofsson et al., 2014] Olofsson, E., Nöremark, M., and Lewerin, S. S. (2014). Patterns of between-farm contacts via professionals in Sweden. *Acta Veterinaria Scandinavica*, 56(70).

[Ortiz-Pelaez et al., 2006] Ortiz-Pelaez, A., Pfeiffer, D. U., Soares-Magalhães, R. J., and Guitian, F. J. (2006). Use of social network analysis to characterize the pattern of animal movements in the initial phases of the 2001 foot and mouth disease (FMD) epidemic in the UK. *Preventive Veterinary Medicine*, 76:40–55.

[Orton et al., 2012] Orton, R. J., Bessell, P. R., Birch, C. P. D., O'Hare, A., and Kao, R. R. (2012). Risk of foot-and-mouth disease spread due to sole occupancy authorities and linked cattle holdings. *PLoS ONE*, 7(4):e35089.

[Pereira et al., 2010] Pereira, U. P., Mian, G. F., Oliveira, I. C. M., Benchetrit, L. C., Costa, G. M., and Figueiredo, H. C. P. (2010). Genotyping of *Streptococcus agalactiae* strains isolated from fish, human and cattle and their virulence potential in Nile tilapia. *Veterinary Microbiology*, 140:186–192.

[Pybus and Rambaut, 2009] Pybus, O. G. and Rambaut, A. (2009). Evolutionary analysis of the dynamics of viral infectious disease. *Nature reviews: Genetics*, 10(8):540–550.

[Rivas et al., 1997] Rivas, A. L., González, R. N., Wiedmann, M., Bruce, J. L., Cole, E. M., Bennett, G. J., Schulte, H. F., Wilson, D. J., Mohammed, H. O., and Batt, C. A. (1997). Diversity of *Streptococcus agalactiae* and *Staphylococcus aureus* ribotypes recovered from New York dairy herds. *American Journal of Veterinary Research*, 58:482–487.

[Robinson and Christley, 2007] Robinson, S. E. and Christley, R. M. (2007). Exploring the role of auction markets in cattle movements within Great Britain. *Preventive Veterinary Medicine*, 81:21–37.

[Rweyemamu et al., 2008] Rweyemamu, M., Roeder, P., MacKay, D., Sumption, K., Brownlie, J., Leforban, Y., Valarcher, J. F., Knowles, N. J., and Saraiva, V. (2008). Epidemiological patterns of foot-and-mouth disease worldwide. *Transboundary and Emerging Diseases*, 55(1):57–72.

[Savill et al., 2006] Savill, N. J., Shaw, D. J., Deardon, R., Tildesley, M. J., Keeling, M. J., Woolhouse, M. E. J., Brooks, S. P., and Grenfell, B. T. (2006). Topographic determinants of foot and mouth disease transmission in the UK 2001 epidemic. *BMC Veterinary Research*, 2:3.

[Scudamore et al., 2002] Scudamore, J. M., Trevelyan, G. M., Tas, M. V., Varley, E. M., and Hickman, G. A. W. (2002). Carcass disposal: lessons from Great Britain following the foot and mouth disease outbreaks of 2001. *Revue scientifique et technique (International Office of Epizootics)*, 21(3):775–787.

[Song and Kulldorff, 2003] Song, C. and Kulldorff, M. (2003). Power evaluation of disease clustering tests. *International Journal of Health Geographics*, 2(1):9.

[Spada et al., 2004] Spada, E., Sagliocca, L., Sourdis, J., Garbuglia, A., Poggi, V., De Fusco, C., and Mele, A. (2004). Use of the minimum spanning tree model for molecular epidemiological investigation of a nosocomial outbreak of hepatitis C virus infection. *Journal of Clinical Microbiology*, 42(9):4230.

[Spratt, 1999] Spratt, B. G. (1999). Multilocus sequence typing: molecular typing of bacterial pathogens in an era of rapid DNA sequencing and the internet. *Current opinion in microbiology*, 2(3):312–316.

[Spratt and Maiden, 1999] Spratt, B. G. and Maiden, M. C. (1999). Bacterial population genetics, evolution and epidemiology. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 354(1384):701–710.

[Thalmann and Nöckler, 2001] Thalmann, G. and Nöckler, A. (2001). Occurrence of foot and mouth disease — a historical survey. *DTW. Deutsche tierarztliche Wochenschrift*, 108(12):484–494.

[Tildesley et al., 2008] Tildesley, M. J., Deardon, R., Savill, N. J., Bessell, P. R., Brooks, S. P., Woolhouse, M. E. J., Grenfell, B. T., and Keeling, M. J. (2008). Accuracy of models for the 2001 foot-and-mouth epidemic. *Proceedings. Biological sciences / The Royal Society*, 275(1641):1459–1468.

[Tildesley and Keeling, 2009] Tildesley, M. J. and Keeling, M. J. (2009). Is R(0) a good predictor of final epidemic size: foot-and-mouth disease in the UK. *Journal of Theoretical Biology*, 258(4):623–629.

[Wallinga and Teunis, 2004] Wallinga, J. and Teunis, P. (2004). Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology*, 160(6):509–516.

[Yang et al., 2013] Yang, Y., Liu, Y., Ding, Y., Yi, L., Ma, Z., Fan, H., and Lu, C. (2013). Molecular characterization of *Streptococcus agalactiae* isolated from bovine mastitis in Eastern China. *PLoS ONE*, 8(7):1–8.

[Ypma et al., 2012] Ypma, R. J. F., Bataille, A. M. A., Stegeman, A., Koch, G., Wallinga, J., and van Ballegooijen, W. M. (2012). Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proceedings. Biological sciences / The Royal Society*, 279(1728):444–450.

[Zadoks and Fitzpatrick, 2009] Zadoks, R. N. and Fitzpatrick, J. L. (2009). Changing trends in mastitis. *Irish Veterinary Journal*, 62(Suppl 4):S59.

[Zadoks et al., 2011] Zadoks, R. N., Middleton, J. R., McDougall, S., Katholm, J. r., and Schukken, Y. H. (2011). Molecular epidemiology of mastitis pathogens of dairy cattle and comparative relevance to humans. *Journal of Mammary Gland Biology and Neoplasia*, 16(4):357–372.