University of Nottingham
UK | CHINA | MALAYSIA

Inference of transmission trees for epidemics using whole-genome sequence data

Rosanna Cassidy, BSc

Thesis submitted to The University of Nottingham
for the degree of Doctor of Philosophy

February 2019

Dedicated to my daughter, Anastasia, whose arrival made this process far more difficult, but far more rewarding.

For God all things are possible.

Abstract

Recently, collection of sequence data has become increasingly rapid and cost-efficient, prompting much research into using this kind of data in the analysis of infectious diseases. There is currently substantial interest in developing epidemic model frameworks which can incorporate this new abundance of data. Whole-genome sequence (WGS) data reveal to us the unique construction— the 'fingerprint'— of the DNA of a sample pathogen. These high resolution data introduce the possibility that we may be able to discover who infected whom in an epidemic outbreak, allowing us to better understand transmission dynamics and therefore design improved preventative and intervention measures. WGS data may prove useful in understanding how levels of infectiousness and susceptibility vary between individuals in a population, or patients on a hospital ward. Genetic data are becoming increasingly widely available, and it is now possible to sequence isolates of some pathogens in real-time in the field with mobile sequencing technologies. Therefore, developing the models and methods to best exploit this is of considerable importance.

The first focus of the research presented here is on antibiotic-resistant nosocomial infections, or 'hospital superbugs', as these still pose a significant problem in hospitals, especially in developing countries. Antibiotic resistance is estimated to kill 700 thousand people globally every year. Current public focus on the threat of an 'antibiotic apocalypse' focuses on the need to reduce the overuse of antibiotics, but another important strategy is to better understand the transmission of such pathogens in order that better prevention and intervention strategies can be designed. Hospital wards present a unique environment, data from which require their own models and methods to analyse outbreaks of infectious disease. Initial research in this thesis has concentrated on outbreaks of methicillin-resistant *Staphylococcus aureus* (MRSA), as it is the most widespread and most common antibiotic-resistant nosocomial infection.

In this thesis, discrete-time stochastic epidemic models are developed which can be

used to analyse both epidemiological and genetic data from an outbreak of MRSA on a hospital ward. These new models can be used to estimate routes of transmission through the hospital ward on the level of individual transmission events by harnessing the information available in the genetic distances between isolate sequences taken from colonised patients. The unobserved transmission dynamics in the models can be inferred using Bayesian inference in a data-augmented MCMC algorithm. Although techniques have been developed to assess the goodness-of-fit of epidemic models in Bayesian settings, they do not assess how well a model fits the genetic data. Methods for doing so are developed in this thesis. An outbreak of MRSA is analysed using the presented models, and the new goodness-of-fit techniques are used to suggest ways to improve the fit of models.

The ideas behind the models for genetic data from MRSA outbreaks are also applicable to other epidemic outbreaks for which genetic data are available. In this thesis we present continuous-time stochastic epidemic models for the spread of avian influenza. These models have a spatial aspect and can be used to estimate the transmission events between farms by analysing genetic and epidemiological data from each farm. Avian influenza is carried endemically by wild birds, so it is very difficult to prevent outbreaks entirely. Therefore, it is very useful to better understand the transmission dynamics of outbreaks and to be able to make predictions about the course of a future epidemic.

The combined analysis of both epidemiological and genetic data through novel models and methods allows transmission of pathogens in epidemic outbreaks to be investigated on the level of individuals in the population. This can have a great public health impact, as results about the routes of infection can inform prevention and control measures.

Acknowledgements

We thank Pramot Srisamang, Ben Cooper, Sharon Peacock, Matthew Holden, Emma Nickerson, Maliwan Hongsuwan and Julian Parkhill for collaborative support with the MRSA chapter. We thank Thomas Hagenaars and colleagues (Wageningen Biovet-erinary Research, The Netherlands) for sharing anonymized outbreak, culling and denominator data for the Dutch 2003 HPAI epidemic with us.

I would like to thank my supervisors, Philip O'Neill and Theodore Kypraios, for their unfailing insight and support. I would also like to thank my husband, Anthony, my parents and my brother, Louis, for their everlasting patience, love, and free childcare. My heartfelt gratitude goes to the friends who supported me in writing this thesis, especially Katie, who kept me focused and motivated, Chloe, who gave up her time to improve my writing skills, and Adam, Conor, Irene and Rachel who are wondrous always.

Contents

Chapter 1

Introduction to epidemic modelling and whole-genome sequence data

1.1 Introduction

This thesis aims to introduce and explore new methods for analysing whole-genome sequence data with stochastic models for the spread of pathogens during outbreaks. The idea is that by harnessing this relatively new type of data we can better estimate the route of transmission that disease takes within a population during an epidemic.

In this chapter we discuss the background to this work, in both epidemic modelling and in whole-genome sequencing. We introduce two specific pathogens which we will model in detail later in the thesis.

1.2 Epidemic models

Epidemiology is a well established field which is concerned with health conditions and diseases in populations of living things [1]. In terms of infectious diseases, epidemiology studies and analyses their occurrence, transmission and possible control measures. This can include mathematical modelling of diseases which can describe aspects of epidemics, including the spread of the disease, control mechanisms and the effect on the population. Epidemic models can be used to estimate the rate and route of transmission of the pathogen, the proportion of the population that is infected, and the effectiveness of intervention measures. The use of such models for communicable diseases can reveal the spatial spread of the disease, and can expose

mechanisms behind the spread which may be to do with the biology of the disease or population, or to do with how the population interacts. Models enable predictions to be made about future epidemics, and can allow for the assessment of how prevention and control measures would impact outbreaks. This can provide valuable public health information; for instance a model could be used to evaluate the benefits of a future vaccination program [2, 3]. Here we describe some common epidemic modelling techniques.

We use the word *pathogen* to refer to the bacterium, virus, or other microorganism which can cause *disease*, which is illness, in a host which carries it, although some pathogens may also be carried without causing disease. We use the words *epidemic* or *outbreak* to mean numerous episodes of disease within a population over a defined, usually short, period of time as opposed to a single *incidence* of the disease in one member of the population. A *population* is a defined group of individuals, which may be people, animals or plants which are at risk of acquiring the pathogen in question.

1.2.1 SIR or SEIR models

We now introduce two standard compartmental epidemic models which will underpin the models presented in this thesis. In either an SIR (*susceptible, infectious, removed*) model or an SEIR (*susceptible, exposed, infectious, removed*) model, the members of a population are each classified by a state, and we model the movement of individuals between states over time [4]. In the SIR model these states are S, susceptible, I, infectious, and R, removed. A susceptible individual is not carrying the pathogen but may acquire it later in time. An infectious individual is carrying the pathogen and may infect other individuals. A removed individual is no longer susceptible to infection, and is also no longer capable of transmitting the disease. This may be because they have left the population through death, or have been isolated so that they no longer have contact with other individuals, or they may have recovered and become immune. A transmission event occurs when there is contact between an infectious individual and a susceptible one which results in the susceptible individual becoming infectious. The basic SIR model is illustrated in figure 1.1. The SEIR model is similar, with the extra state E being exposed individuals that carry the pathogen but are not yet able to infect other individuals. In this model a transmission event occurs when contact between an infectious individual and a susceptible individual results in the susceptible individual becoming exposed. After a latent period in this state they become infectious and can pass on the infection.

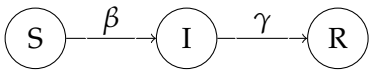


Figure 1.1: An illustration of an SIR model. An individual moves from state S, where they are susceptible, to state I, where they are infectious, at rate β. An individual moves from state I to state R, where they are removed through immunity or death, at rate γ.

Other variations of these compartmental models include an SI model where individuals remain infectious for the entire study length, an SIS model where recovered individuals become susceptible to infection again, or an SIRS model where individuals are removed by temporary immunity or isolation from the rest of the population and then become susceptible to infection again after this period.

The rate of transmission, β, in figure 1.1 measures how quickly susceptible individuals become infectious through contact with other infectious individuals. This rate depends on how the population mixes, so how likely it is that a susceptible individual will have an interaction with an infectious individual. In more complex models, it may also depend on how susceptible the specific individual is, and on how infective the specific infectious individual that they come into contact with is. Contact does not have to refer to the two individuals physically touching, it can mean that the pathogen transfers between them through the air or through some third party carrier that is not explicitly included in the model. The most common SIR model assumes that the mixing, susceptibility and infectivity of the population is homogenous, and therefore the rate of transmission is directly proportional to the size of the infected population. A specific susceptible individual becomes infectious at a rate $q(t) = \beta I(t)$. The coefficient β is the standard for transmission. Clearly homogenous mixing and population characteristics will not be appropriate in every outbreak, but this can be incorporated into this type of model easily by allowing the parameter β to depend upon a characteristic of the susceptible or infectious individual, or upon their geographic spacing. Homogeneous susceptibility and infectivity will suffice for the purposes of this thesis.

1.2.1.1 Stochastic SIR model

Here we describe a stochastic SIR model, in continuous time, which allows for randomness in the outcome of the epidemic. From the same initial situation an outbreak could die out quickly or could become much bigger, simply by chance. The initial population is considered to be made up of m susceptible individuals and n infectious individuals. The infectious individuals remain infectious for certain lengths of time (their infectious periods) which are independently and identically distributed

according to a random variable P which has a specified distribution. At the end of an individual's infectious period they are removed from the population. Whilst a given individual, i, is infectious they come into contact with another given individual, j, at the points of a Poisson point process which has rate β and is independent of the infectious periods and the Poisson processes governing contacts between other pairs of individuals. If the individual j is susceptible at the time that the contact takes place then j is infected and becomes infectious. Once there are no infectious individuals left in the population the epidemic immediately ends [5].

If the infectious periods of the infectious individuals are independently and identically distributed according to a random variable which is exponentially distributed, $X \sim \text{Exp}(\gamma)$, with PDF given by:

$$\text{f}_X(x) = \begin{cases} \gamma\, \text{e}^{-\gamma x} & x > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{1.2.1}$$

with parameter $\gamma > 0$, then this model is a model known as the *general stochastic epidemic model* [2]. In this case, if $S(t)$ and $I(t)$ are the number of susceptible and infectious individuals in the population at time t, then the process $(S, I) = \{(S(t), I(t)); t \geq 0\}$ is a Markov process. Assuming exponential infectious period distributions can therefore make the analysis much simpler.

The basic stochastic SIR model can be extended in a number of different ways. The rate β may be assumed to differ between different pairs of individuals, so that individual i comes into contact with individual j at the points of a Poisson process with rate $\beta_{i,j}$. This means that the model can allow for different types of individuals (eg. individuals in different age or sex categories) with different infectious period distributions and/or different susceptibilities. The rate $\beta_{i,j}$ may also rely on the spatial distance between the two individuals.

1.3 Bayesian inference

In order to be used in applications, epidemic models are fit to the specific data that are of interest. Estimates for the values of the parameters given the data can be obtained and will provide information about the epidemic. There are many techniques which can be used to find these estimates, but in this thesis we will exclusively focus on Bayesian inference. Unlike classical statistics, where parameters are assumed to have fixed values, Bayesian statistics assume that each parameter has a probability

distribution [6].

An advantage of Bayesian inference is that prior knowledge about the parameters can be included through what is called the *prior* distribution [7]. Equally, an uninformative distribution can be used for the prior if there is no previous knowledge about the parameter, or if we want to ignore this knowledge in order to investigate the parameter based solely on the data. The prior distribution comes from the fact that Bayesian inference is based on Bayes' Theorem which, for data x and parameters θ, states that

$$\pi(\theta|x) = \frac{\pi(x|\theta)\pi(\theta)}{\pi(x)} = \frac{\pi(x|\theta)\pi(\theta)}{\int \pi(x|\theta)\pi(\theta)\,\mathrm{d}\,\theta}. \tag{1.3.1}$$

Here $\pi(\theta)$ is the prior density of θ, $\pi(\theta|x)$ is the posterior density of θ given the data x, and $\pi(x|\theta)$ is the likelihood of the data x given the parameters θ. Thus features of the posterior distributions of the parameters which we wish to investigate, such as the mean, quantiles and moments, can be described using the prior and likelihood [6]. For example, the posterior mean is

$$\mathbb{E}[\theta|x] = \frac{\int \theta\pi(x|\theta)\pi(\theta)\,\mathrm{d}\,\theta}{\int \pi(x|\theta)\pi(\theta)\,\mathrm{d}\,\theta}. \tag{1.3.2}$$

However, the integrations required to calculate such expressions which would give the features of the posterior distributions are most often intractable, especially when the model has high dimensions [8]. Therefore Markov Chain Monte Carlo methods are commonly used to sample from the posterior densities of the parameters. These samples can be used to approximate expectations of the features of the posterior distribution.

1.3.1 Markov Chain Monte Carlo

We now briefly introduce Markov Chain Monte Carlo methods which are commonly used in Bayesian inference. A Markov Chain is a sequence of random states, $X_1, X_2, \ldots X_n$, where the current state is only dependent upon the previous state so the transition kernel is $\mathrm{P}(X_{i+1}|X_i)$. Monte Carlo integration repeatedly samples from a distribution and then uses these samples to approximate expectations of the distribution. Markov Chain Monte Carlo works by generating samples from a Markov Chain with limiting distribution which is the distribution that we are interested in, $\pi(\theta|x)$, and using these repeat samples to approximate expectations of functions of $\pi(\theta|x)$ [8].

1.3.1.1 Metropolois-Hastings algorithm

One way of producing a Markov Chain which has limiting distribution $\pi(\cdot)$ is to use the Metropolis-Hasting Algorithm [9]. This algorithm works by taking a target density $\pi(\theta)$, which need only be known in relation of proportionality, and a proposal distribution $q(\widetilde{\theta}|\theta)$ and at each iteration sampling an updated value, $\widetilde{\theta}$. The algorithm starts with initial parameter values θ^0 and then at each iteration, n, proposes a value $\widetilde{\theta}$ for θ^n from the proposal distribution. The proposed value is accepted with probability

$$\min\left(1, \frac{\pi(\widetilde{\theta})q(\theta^{n-1}|\widetilde{\theta})}{\pi(\theta^{n-1})q(\widetilde{\theta}|\theta^{n-1})}\right)$$

and $\theta^n = \widetilde{\theta}$ or else $\theta^n = \theta^{n-1}$. The Markov Chain produced here has a stationary distribution which is the target distribution if it converges. It will converge after a large number of iterations, as long as the proposal distribution can propose all points in the space of the true distribution. The period of iterations before convergence is known as the 'burn-in period'. The data from these iterations do not need to be stored, and the number of iterations in the burn-in period can depend on the efficiency of the proposal distribution [10]. When the target density is the posterior distribution of parameters θ only the prior density, $\pi(\theta)$, and likelihood, $\pi(x|\theta)$, need to be calculated at each step as the constant of proportionality cancels in the acceptance probability.

Random-walk proposal

The most common choice of proposal distribution for the Metropolis-Hastings algorithm is a random-walk proposal distribution. The idea is to sample from a standard symmetric distribution and add this sampled value, ε, to the current value of the parameters. So

$$\widetilde{\theta} = \theta^{n-1} + \varepsilon.$$

The Metropolis-Hastings acceptance probability is then

$$\min\left(1, \frac{\pi(\widetilde{\theta})}{\pi(\theta^{n-1})}\right).$$

1.3.1.2 Gibbs sampler

A special case of the Metropolis-Hastings algorithm is called the Gibbs sampler [11, 12], and is based on the target distribution $\pi(\theta)$. If the model has m parameters, so $\theta = (\theta_1, \theta_2, \ldots \theta_m)$, we consider one of the parameters singly, θ_i, from the parameter vector θ and then write its conditional distribution as $\pi(\theta_i|\theta_{-i}, x)$, where θ_{-i} is

the vector of parameters without θ_i. If this conditional distribution is a distribution which we can sample from then we may set this as the proposal distribution, $q(\widetilde{\theta}|\theta)$, in the Metropolis-Hastings distribution. The acceptance probability under this proposal distribution cancels to 1, so each sample of $\widetilde{\theta}$ will be accepted.

1.3.1.3 Data augmentation in MCMC

Here we introduce the technique of data augmentation. Data augmentation is very useful in the analysis of epidemics because the unobserved transmission events can be sampled over, making the likelihood tractable. Many analyses in the literature use this method [13–19].

Often, the evaluation of the posterior densities of the parameters is hindered by likelihoods that are made intractable by missing or unobserved data. In these cases data augmentation can sometimes be used to sample from the posterior distribution by including the missing data, T, as if it were another set of parameters [20], so $\pi(x|\theta) = \int_T \pi(x, T|\theta)\,\mathrm{d}\,T$. The parameter space is augmented to (θ, T) with the missing data so that the augmented likelihood $\pi(x, T|\theta)$ can be considered. From Bayes' Theorem we get

$$\pi(T, \theta|x) = \frac{\pi(x, T|\theta)\pi(\theta)}{\pi(x)},$$

which is proportional to the likelihood of x and T given the parameters θ multiplied by the prior distribution for the parameters.

A data-augmented MCMC algorithm follows these steps:

1. Initial values are set for the parameters, θ^0, and for the missing data, T^0.
2. The parameters, θ, are updated using Metropolis-Hastings or Gibbs steps.
3. Proposal values, $\widetilde{T}$, for the augmented data are generated using a sampling distribution, $q(\widetilde{T}|T, \theta)$.
4. The proposed values are accepted with probability

$$\min\left(1, \frac{\pi(\widetilde{T}|\theta)\pi(x|\widetilde{T}, \theta)\pi(\theta)q(T)}{\pi(T|\theta)\pi(x|T, \theta)\pi(\theta)q(\widetilde{T})}\right)$$

   and the current state of T, T^n, is set to $\widetilde{T}$, or else T^n is set as T^{n-1}.
5. Steps 2-4 are repeated until the desired number of iterations have been completed.

1.3.1.4 Posterior predictive distributions

Assessing the fit of a model to a particular set of data is often done in a Bayesian setting using the posterior predictive distribution [21]. We have defined the posterior density $\pi(\theta|x)$ for the parameters θ of a model given the data x. The posterior predictive distribution is used to make predictions about hypothetical future data x^{new}. This posterior predictive distribution is given by

$$\pi(x^{new}|x) = \int \pi(x^{new}|\theta)\pi(\theta|x)\,\mathrm{d}\,\theta.$$

A set of auxiliary statistics, $A(x)$, may be specified which are matched to the observed data when sampling the future data so $A(x^{new}) = A(x)$. These auxiliary statistics may include the length of the study, the size of the sample population, dates of sample collections and other such things that do not form part of the stochastic modelling framework. Then,

$$\pi(x^{new}|x, A(x)) = \int \pi(x^{new}|\theta, A(x))\pi(\theta|x)\,\mathrm{d}\,\theta.$$

In order to assess the goodness-of-fit of a model to the data this posterior predictive distribution may be used to repeatedly simulate hypothetical future datasets from which an approximation to the distribution of a summary statistic, $S(x)$, may be produced. This summary statistic must be chosen to represent the data and capture the variation between the hypothetical datasets.

The observed value of the summary statistic from the original data, $S(x)$, can be compared to the approximated distribution, and this gives a posterior predictive p-value [22],

$$p_s = \mathrm{P}(S(x) \geq S(x^{new})|x, \theta).$$

Extreme p-values which fall outside of the $2.5\% - 97.5\%$ interval give evidence against the fit of the model to that specific dataset.

1.4 Whole-genome sequence data

We now discuss genetic data, specifically pathogen genome data, because of the information which it can provide about transmission during epidemics. We will go on to introduce models for analysing such data in section 1.7. In the past decade the collection of genome sequence data has become increasingly rapid, accurate, and cost-efficient [23–25]. Genome sequence data are of interest to those studying population dynamics and evolution, as well as to those studying the evolution and transmission

of pathogens. For pathogens, genome sequence data can provide information about the mechanisms of evolution which allow the disease to become more infectious or drug-resistant. It can also give more information about the route that transmission of the pathogen takes through a population, prompting much research into using this kind of data in analysing epidemics. Whole-genome sequences (WGS) data display the unique construction- the 'fingerprint'- of the DNA of a sample of a pathogen. High resolution data allow for better identification and tracking of pathogens from an outbreak, meaning that we can better understand transmission dynamics and therefore design improved prevention and intervention measures. Here we introduce some of the basics of genome sequencing.

1.4.1 The structure of genomes

A genome is a DNA structure which is made up of two strands which are linked in the classic double-helix structure. Each strand of DNA is made up of a string of nucleotides, each of which is paired with a nucleotide in the opposite strand in what is called a *base pair*. Nucleotides can take four different bases, adenine (A), cytosine (C), guanine (G) or thymine (T). Base A will only pair with base T (and vice versa) and base C will only pair with base G (and vice versa), so only one strand of the DNA is necessary to reveal all of the information about the genome. A whole-genome sequence will vary in length, L, depending on the organism that it comes from, but can always be represented as a vector $B = (B_1, B_2, \ldots, B_L)$, where each $B_i \in \{A, C, G, T\}$.

Mutations occur in the DNA of an organism when there is a mistake made during the replication process. A mutation of a single base pair is called a single nucleotide polymorphism (SNP). Transitions happen when a nucleotide base is replaced by its pair, so A by G and G by A, or C by T and T by C. Transversions happen when a nucleotide base is replaced by one of the other two bases which are not its pair. It is generally accepted that transitions are more likely than transversions, although many models for mutation allow transitions and transversions with equal probability.

Diversity in genomes from an organism can also arise through recombination events, where portions of the genome break off and are reinserted at a different site, or through horizontal gene transfer, where portions of the genome are imported from another source either in the environment or from a plasmid or virus. This type of diversity accumulation produces more pronounced change in the organism than single mutations because more of the DNA is changed at once. It is likely that these events

are the origin of new traits in organisms such as resistance to antibiotics in bacteria [26].

In the following sections we will introduce two methods for collecting the information present within genome sequences. We will start with the more basic method, genotyping, and then will introduce whole-genome sequencing techniques.

1.4.2 Genotyping

Genotyping does not require sequencing the whole-genome for each sample, but instead focuses on discovering what separates one genome from another and assigning them to types or clusters. There are many methods for doing this, the earliest of which involved examining fragments of DNA or RNA. More recent techniques focus on regions of the genome that are known to produce variation, such as known repetitive regions, or specific nucleotides which are known to be variable. These genotyping methods can determine whether a pair of sequences belong to the same type, and therefore whether they are likely to be part of the same outbreak. However, these methods are limited as they depend on previous knowledge about the genomes of the specific pathogen being investigated. In order to look in detail at possible transmission events greater resolution is required.

1.4.3 Whole-genome sequencing

Most whole-genome sequencing techniques work by separating the genomes into overlapping fragments which are then sequenced and finally reassembled to give the full-length sequence. One of the first, and most well-known, of these was the Sanger method, which uses bacterial cloning and DNA polymerase to create a series of DNA fragments which are nested, and end in a known nucleotide. This is achieved by synthesising nucleotides to a single strand of the DNA to be sequenced using a solution which contains adapted versions of one of the four nucleotide bases. These adapted bases will terminate the pairing process. Since the length of the fragment can be determined, and the nucleotide base at the end of the fragment is known, from the whole series of fragments the whole-genome can be reassembled [27].

Next-generation sequencing (NGS) allows for higher-throughput pipelines of DNA sequencing. There are a range of methods currently available in commercial packages, the most widely used being Illumina with at least 90% of sequencing data

worldwide created through this platform [28]. Most platforms share the same basic preparation steps during which the DNA is fragmented and denatured resulting in sections of single stranded DNA. Adaptors are added to the ends and the fragments are amplified to create many copies which can be sequenced in parallel. Each method sequences these fragments by synthesis using either DNA polymerase or DNA ligase. The next nucleotide to be paired is signalled in some way, either through fluorescent or died nucleotides, or through a byproduct of the synthesis [24, 29].

The latest development in whole-genome sequencing has been the development of third-generation sequencing (TGS) methods [25]. These long-read methods do not need to fragment and amplify the genome into sections and the longer reads reduce the risk of mistakes being made in the reassembly of the genome due to repetitive sections. As this technology sequences single molecules it is faster than NGS methods and the DNA can be sequenced in real time instead of having to pause after each nucleotide read. This is achieved by sensors in the machine which record in real time the products, for example fluorescence or change in ionic current, of the reactions when each nucleotide is synthesised. The MinION platform for TGS has allowed real time sequencing of DNA in the field during ongoing outbreaks [30, 31] as it is small enough to be portable, and the cost-efficiency of such platforms makes such sequencing feasible.

1.5 Healthcare associated methicillin-resistant *Staphylococcus aureus*

We now discuss a specific pathogen from which WGS data may be collected for analysis. This thesis has a particular focus on outbreaks of methicillin-resistant *Staphylococcus aureus* (MRSA), which falls into the category of antibiotic-resistant nosocomial infections, or 'hospital superbugs'. These still pose a significant problem in hospitals [32], resulting in increased levels of illness and death and requiring patients to stay in hospital for longer periods with the associated costs for treatment and bed space. In the U.S. in 2018, it was estimated that more than 2 million people per year are infected by an antibiotic-resistant microbe, and 23,000 of these ultimately die [33]. In the European Union it is estimated that antimicrobial infections cause 25,000 deaths per year and result in 2.5 million days in hospital [34]. The situation is even worse in developing countries, with 58,000 babies estimated to die in India per year through infection with antibiotic-resistant bacteria [35], and 38,000 deaths per year estimated

in Thailand, along with 3.2 million extra hospital days [36].

There is currently much public focus on the threat of an 'antibiotic apocalypse' during which even routine infections will be untreatable due to completely antibiotic-resistant superbugs [37]. Although such public campaigns aim to reduce the overuse of antibiotics, another key strategy is to better understand the transmission of such pathogens in order that better preventative strategies can be designed. Hospital wards present a unique environment, data from which require their own models and methods to analyse outbreaks of infectious disease. The population on a hospital ward changes through admissions and discharges, and admissions can bring new importations of the disease onto the ward. Patient-to-patient transmission is often facilitated by the frequent contacts of healthcare workers, who may have temporarily contaminated hands, with each of the patients [38].

There is a substantial amount of literature that investigates MRSA as it is the most prominent and widespread of the 'superbugs'; MRSA causes more nosocomial infection than any other pathogen, and in 2004 it was estimated to cause disease in 2% of all patients in hospitals [39]. *Staphylococcus aureus* (SA) is a bacterium which is persistently and asymptomatically carried by 20% of the healthy population, and intermittently by 60% of the healthy population [40]. The bacterium is most often found within the nose, but can also be carried on the skin of the neck, forearm, hand, chest, abdomen, back, thigh and ankle, and in the perineum [41]. Methicillin-resistant *Staphylococcus aureus* has also been found in the urinary tract, groin and pharynx [42]. When SA or MRSA enters the bloodstream through a wound, broken skin or surgical site it can cause severe illness and even death. Mortality rates are higher for patients infected by MRSA than for those infected with methicillin-susceptible SA [43, 44] and on average they stay in hospital for longer [45].

1.6 Avian influenza

In this section we introduce another pathogen, avian influenza, from which WGS data may be, and has been, collected to assist in analysis of epidemics. The population setting for the study of this pathogen is very different from the nosocomial pathogen discussed earlier as avian influenza epidemics can cover a large geographical area so the population is often assumed to comprise individual farms or other groups of birds rather than individual birds.

Subtypes of the Type A virus which causes avian influenza, known as 'bird flu', are endemically carried by flocks of wild birds and in particular by waterfowl [46]. The specific H5 and H7 subtypes can mutate into highly pathogenic avian influenza (HPAI) which causes epidemics in commercial flocks [47] through infection of the tissues of the respiratory, digestive and nervous systems of poultry. HPAI is associated with high transmissibility and high rates of mortality, up to 100% [48]. Between 1996 and 2008, HPAI viruses are known to have caused 11 separate epidemics and four of these outbreaks involved several millions of poultry [49].

Avian influenza can also be transmitted to humans. In China there have been five epidemics of human infections of the strain H7N9 since 2013 with 1344 cases and 511 mortalities up to April 2017 [50, 51]. In order to prevent these outbreaks among humans and to limit the loss of commercial poultry it is vital that transmission during epidemics is better understood in order to enhance prevention and control strategies.

Since it is very difficult to prevent outbreaks of avian influenza entirely, due to its endemic carriage in wild birds and the fact that cases are not detected until they start to exhibit clinical symptoms, which is some time after the infection event, it is important to study the transmission dynamics of outbreaks in order to be able to make predictions about the course of a future epidemic whilst it is in progress. Knowledge about likely transmission routes could help to end an outbreak by minimising its spread through culling and other intervention measures [52].

1.7 Models to analyse genetic and epidemiological data

Now we discuss epidemic models which have been used to analyse genetic and epidemiological data from pathogens including, but not exclusive to, MRSA and avian influenza. The recent abundance of genetic data from such epidemics has lead to the proposal of many different methods and models which can be used to analyse this sort of data. These methods differ in terms of the type of data that they require from the epidemic and in what sort of outbreaks they are suitable for. Table 1.1 lists some of the well-known Bayesian inference methods and summarises the key similarities and differences between them. The first, third and fourth models in the table do not include a forward model for the spread of the pathogen, instead working backward to reconstruct the transmission tree, whereas the others do include a forward model for the outbreak through time.

In table 1.1:

- 'Multiple imports' means that the method allows for the disease to be imported into the outbreak from outside the population through more than one event.
- 'Multiple sequences per host' means that the method allows analysis of more than one sequence per host.
- 'Unsampled case' means that the method allows for there to have been cases of the disease which were not detected or sampled (not simply known cases with missing sequences).
- 'Infection times estimated' means that the method estimates the infection times as part of the method rather than assuming they are known beforehand.
- 'Susceptible population considered' means that the whole population is modelled rather than just those who were infected.
- 'Microevolution model' means that the genetic distances between hosts are modelled with some mechanism which produces mutations on the genome.
- 'Gen. and epi. data indep.' means that the genetic and epidemiological data are assumed to be independent.
- 'Phylogeny estimated' means that the method estimates the phylogenetic tree as well as the transmission tree.
- 'Spatial element' means that the physical distance between hosts is modelled.

We describe these models in more detail in the next sections.

Model	Multiple imports	Multiple sequences per host	Un-sampled cases	Infection times estimated	Susceptible population considered	Micro-evolution model	Gen. and epi. data indep.	Phylogeny estimated	Spatial element
Cottam et al. [53]	no	no	no	yes	no	yes	yes	yes	no
outbreaker [18]	yes	no	yes	no	no	yes	yes	no	no
Numminun et al. [54]	no	yes	no	no	yes	no	yes	yes	no
Hall et al. [55]	no	yes	yes	yes	yes	yes	yes	yes	yes
Ypma et al. [19]	no	no	no	no	no	yes	yes	no	yes
Morelli et al. [17]	no	no	yes	yes	yes	yes	no	no	yes
Mollentze et al. [56]	yes	no	yes	yes	yes	yes	no	no	yes
Worby et al. [14]	yes	yes	yes	yes	yes	no	no	no	no

Table 1.1: A summary of key similarities and differences between methods for analysing genetic and epidemiological data from epidemic outbreaks.

1.7.1 Phylogeny-based methods

Many models rely upon the construction of phylogenetic trees, where the sampled sequences are the external nodes and the internal nodes are the ancestors of the sequences going back to the most recent common ancestors so that the tree is fully connected. Although construction of these phylogenies provides information about how related the samples in the data are, it is not straightforward to link them to transmission trees. Cottam et al. [53] present a method which does not integrate the genetic and epidemiological data but uses them one after the other by first using statistical parsimony genealogies constructed from the genetic data to find the set of possible transmission trees. Then most likely transmission tree is found using the epidemiological data. This method was applied to data from the 2001 outbreak of foot-and-mouth disease in the UK and was successful in finding a tree which gave sources for each infection which were 80% more likely than any other source. The construction of the genealogies in this method is achieved using the software package BEAST [57] which performs Bayesian phylogenetic inference, and other analyses to do with the evolution of sequences, using MCMC algorithms. BEAST is widely used as it provides a number of different models for the evolution of sequences and for tree structures.

Numminen et al. [54] also use phylogenetic trees in their importance sampling scheme during which both the phylogeny and the transmission tree are sampled from importance distributions. Hall et al. [55] propose a method which also samples from the spaces of both trees at the same time using MCMC. The posterior probability of the trees is calculated using a model for the structure of the epidemic which models transmission at the individual host level and also models the DNA evolution process taking place within each host. This model is available in the software package BEAST.

1.7.2 Non-phylogeny-based methods

Other approaches avoid the use of phylogenies by using functions of the genetic distance between samples to weight the edges of a transmission tree. In the R package *outbreaker* Jombart et al. [18] presented the first method to be widely available as a software package. This approach does not use genealogies but treats the transmission tree as a network with edges corresponding to infection events. The simplest assumption of maximum parsimony is adopted: edges are weighted by the number of mutations between samples from the hosts at each node and the minimum weighted

tree is sought. This method allows multiple introductions of the pathogen and unobserved cases, but does not allow hosts to be infected before their sample is taken, so it requires a densely-sampled outbreak. This method was used to reanalyse the 2003 outbreak of Severe Acute Respiratory Syndrome (SARS) in Singapore and produce a transmission tree with one source case rather than the two which were inferred by the previous study. Investigations into the dynamics of the disease mutation supported this transmission tree.

Ypma et al. [19] propose a method which uses genetic data alongside spatial and temporal data to construct transmission trees for the 2003 outbreak of avian influenza in the Netherlands. This approach assumes that these three types of data are independent, so the likelihood that farm A infected farm B is simply the product of contributions from each data type. In the genetic contribution transitions and transversions in the DNA are assumed to occur at different rates, but the possibility of a nucleotide mutating twice between sequences is neglected. The probabilities of all possible transmission events are attained by averaging over the posterior density over the sample space.

The methods above rely on the assumption that the epidemiological and genetic data are independent when constructing the likelihood. Morelli et al. [17] present a method which estimates the likelihood of transmission trees using all sources of information (genetic, location and timing data) simultaneously. This method allows for multiple mutations at one nucleotide position using the Jukes-Cantor correction which allows for the fact that some nucleotides which appear unchanged may have actually changed multiple times before reverting to their original state, so less distance is observed between sequences. The Jukes-Cantor correction [58] states that the mean number of nucleotide differences that have actually occurred in a single position on the genome, μ is related to the mutation rate m by $\mu = \frac{3}{4}\ln\left(\frac{3}{(3-4m)}\right)$. This means that, in the Morelli et al. model, the conditional distribution of M, the number of substitutions between two sequences, given Δ, the sum of time intervals along the transmission chain, becomes:

$$M|\Delta \sim \text{Bin}\left[s, \frac{3}{4}\{1-\exp(-\frac{4}{3}m\Delta)\}\right]$$

where s is the length of an observed sequence and m is the mutation rate per nucleotide per day. The method gives the joint posterior distribution of the transmission tree, infection times, duration of latent periods, lag between infection and detection, and parameters (transmission and latency) given the observed genetic, spatial and

temporal data. Mollentze et al. [56] extend this framework to allow for there to be multiple, unconnected transmission trees which each begin with a separate importation of the disease. This method also estimates the total number of unobserved cases during the sample period.

Worby et al. [14] introduce two models which can be used to infer the transmission tree with an MCMC algorithm working forwards through time by deciding which of the previously colonised patients was most likely to have infected the next infected patient. In contrast to the other models discussed, these models simply use the SNPs between sequences as a measure of genetic distance rather than having a model of microevolution. The first model, the Importation Structure model, assumes that there are different MRSA types that isolates can belong to (the number of types is inferred during the MCMC algorithm) and the probability of the genetic distance, $\Psi_{i,j}$, between any pair of isolates is given by:

$$\mathrm{P}(\Psi_{i,j} = x) = \begin{cases} \mu(1-\mu)^x & \text{if } i \text{ and } j \text{ same type} \\ \mu_G(1-\mu_G)^x & \text{otherwise} \end{cases} \tag{1.7.1}$$

where μ, $\mu_G \in [0,1]$, and x is an integer value taking a value between zero and L, the length of the genome. The second model, the Transmission Chain Diversity model, assumes that SNPs accumulate over time. The probability of the genetic distance between a pair of sequences is then:

$$\mathrm{P}(\Psi_{i,j} = x) = \begin{cases} \mu\gamma^{t(r(i),r(j))}(1-\mu\gamma^{t(r(i),r(j))})^x & \text{if } i \text{ and } j \text{ in same transmission tree} \\ \mu_G(1-\mu_G)^x & \text{otherwise} \end{cases} \tag{1.7.2}$$

where $r(k)$ is the patient whom the kth sequence belongs to, $t(r(i), r(j))$ is the time between colonisation of the patients and $\gamma \in (0,1)$ accounts for increasing distance over multiple transmission events. These models are the only ones that we have found which allow for multiple importations, multiple sequences per host and unsampled cases within the same framework. They also have the advantages of estimating the infection times and explicitly modelling the susceptible population.

1.7.3 Strengths and weaknesses of models to analyse genetic and epidemiological data

The methods and models for analysing genetic and epidemiological data which we have discussed vary a lot in the approaches taken. The phylogeny-based methods work backwards in time, and they rely upon inferring common ancestors among the

sampled genomes which provide the tips of the phylogenetic tree. Therefore these methods are not applicable when the set of sampled genomes contains both ancestors and their descendants [59].

Of the non-phylogeny-based models, the Worby et al. model is the only one which allows for multiple importations of the pathogen to the population, unsampled cases within the population, and multiple sequences per host. The necessity of allowing for multiple importations will depend on the pathogen being studied, but for a nosocomial pathogen such as MRSA, for example, it is very important to have a model which allows for more than one individual to introduce the disease into the ward. Modelling unsampled cases will be advantageous in most situations, as it is rare for an outbreak to be fully sampled. Allowing for multiple sequences per host is important in order to account for within-host diversity. The impact of within-host diversity will again depend upon the specific pathogen, but studies of MRSA have shown that a single host can carry multiple sequence types [60, 61]. Another advantage of the Worby et al. model is that it estimates the infection times instead of relying on them being known, and it explicitly models the likelihood of the susceptible population avoiding infection, which allows for estimation of the transmission rate as well as the transmission tree.

The lack of a microevolution model in the Worby et al. model means that there are no complicated equations governing the accumulation of genetic diversity in the sequences. Instead this model works simply by assuming that the genetic distances between sequences are drawn from certain distributions depending on the relationship between the hosts that they are sampled from. This reduces the dimensions of the model, and by using the genetic distances it can allow for genetic diversity gained from SNPs or recombination. However, this model does assume that the genetic distances between each pair of individuals are all independent, which is not accurate for a transmission chain, eg. in the chain $a \rightarrow b \rightarrow c$ the genetic distance between the isolates from individuals a and c will not be independent of the genetic distance between the isolates from a and b, and that between b and c.

The Worby model also does not include a spatial model since it was designed for nosocomial infections where this would not be applicable. This thesis will present models which will harness the advantages of the Worby et al. model, which have been discussed here, but which will not incur the same disadvantages. Relaxing the limiting assumption of independence between the genetic distances will be useful for

all applications of such models. Defining a model which also has a spatial element will be useful for applications to pathogens in populations which are geographically spread.

1.8 Aims and structure of the thesis

In this thesis the aims will be:

1. To assess the validity of common assumptions in the modelling of genetic distances and to present new models which relax the most restrictive assumptions.
2. To present a stochastic model for the spread of disease which incorporates these new genetic distance models and to show how it may be used to analyse genetic and epidemiological data.
3. To present a new method for assessing the goodness-of-fit of genetic distance models.
4. To investigate an outbreak of MRSA in Thailand, including fitting of transmission trees.
5. To investigate an outbreak of avian influenza in the Netherlands, including fitting of transmission trees.

Chapter 2 investigates relaxing common assumptions in the modelling of genetic distances, and introduces three new models. These models are described in the context of a full stochastic model for the spread of a disease in a population. Chapter 3 describes methods for assessing the goodness-of-fit of epidemic models to epidemiological data and presents a method for also assessing the goodness-of-fit of a genetic model to data consisting of a genetic distance matrix. Chapter 4 fits the three new models presented in chapter 2 to an outbreak of MRSA using a discrete-time data-augmented MCMC algorithm. The performance of the algorithm is assessed using a simulation study, and the goodness-of-fit of the model is investigated using the model assessment techniques introduced in chapter 3. New versions of the models are suggested. Chapter 5 fits the three genetic models within a continuous-time epidemic model with a spatial kernel to an outbreak of avian influenza using a data-augmented MCMC routine. The performance of the algorithm is again assessed using a simulation study, and the goodness-of-fit of the model is investigated using our model assessment techniques.

Chapter 2

Models for epidemics to analyse whole-genome sequence data

2.1 Motivation

In the past decade technology for the collection of sequence data has been continuously improving in speed, cost and accuracy, prompting much research into using this type of data in the field of epidemiology. There has been a focus on developing models and methods for analysing epidemics which exploit this new abundance of data. The increasing availability of whole-genome sequence (WGS) data introduce the possibility that we may be able to infer who-infected-whom in an epidemic outbreak, allowing for better understanding of transmission dynamics. This can inform the design of improved preventative and intervention measures. WGS data may prove useful in understanding how levels of infectiousness and susceptibility vary between individuals in a population, or patients on a hospital ward. Genetic data are becoming increasingly widely available, with it not unlikely that within the next decade all cases of an emerging pathogen could be sequenced during an outbreak [62], so developing models and methods to best take advantage of this is of significant importance.

The starting focus of this research is concerned with nosocomial infections, or 'hospital superbugs', as these are still a cause of increased illness and mortality in hospitals [32]. Hospital wards present a unique environment which requires its own models and methods because the population dynamics are very different to other settings, with patients admitted and discharged from the population. Patients remain on the ward constantly during their hospital stay, so there is much opportunity for contact between patients, either directly or indirectly through healthcare workers. After

developing our new models we proceed to apply them to an outbreak of methicillin-resistant *Staphylococcus aureus* (MRSA) in a hospital in Thailand in chapter 4. There is a substantial amount of literature concerned with MRSA transmission due to it being the most prominent and widespread of the 'superbugs'. MRSA causes more nosocomial infection than any other pathogen, and in 2004 it was estimated to cause disease in 2% of all patients in hospitals [39]. More recently, the highest rates of MRSA infection have been reported in Asia and North and South America, where it is estimated to cause >50% of all healthcare-associated infection, although countries in Africa as well as China, Australia and some European countries also report rates from 25-50% [63].

In order to investigate how well our models, which focus on the joint modelling of genetic and epidemiological data, can capture the dynamics of other types of outbreaks, we adapt and apply our models to an epidemic of highly-pathogenic avian influenza which affected the Netherlands in 2003 in chapter 5. Highly-pathogenic zoonotic diseases such as this are important to study as they are often characterised by fast transmission and large losses of commercial animals. In the last two decades, a number of outbreaks of these types of pathogens, including avian influenza [64, 65], swine influenza [66, 67], and foot-and-mouth disease [53, 68], have occurred in different countries, with widespread economic impact and concern for public health. Therefore, control measures have an important role to play in lessening the effects of such epidemics, and the better we can understand the dynamics of outbreaks, the better we can design the control measures.

2.2 Introduction

Many studies have investigated the transmission of pathogens, especially in hospital settings, using whole-genome sequence data alongside traditional epidemiological data such as admission and discharge times, and pathogen swab test results. Many different models have been presented to infer transmission trees for outbreaks of such pathogens some of which have been presented in chapter 1. These models aim to construct a transmission tree which shows the source of every infected individual. If an individual was colonised by someone before they entered the study population we call them an importation. If they were colonised by another individual in the sample during the course of the study we refer to this as a direct transmission event. A sequence of direct transmission events such that individual a infects individual b who infects individual c etc. is referred to as a transmission chain, and we say that c

was indirectly infected by patient a. The transmission tree may be made up of more than one transmission chain. In almost all models it is assumed that if a patient was directly or indirectly infected by another patient this pair of individuals will have sequences that are more genetically similar than patients who are in distinct chains of transmission, so the probability that a transmission event occurred between two patients who are colonised depends on the genetic diversity observed between their sequences. There is much disparity between different models as to how the variation in genetic distances between patients' sequences is modelled.

In section 2.3 we describe the Worby et al. model for inferring transmission trees [14]. Section 2.4 looks at some of the assumptions made in the Worby et al., and many other, models and explores the impact of these assumptions and how they might be relaxed. This provides the context for the model for genetic distances that we develop in section 2.6. Section 2.7 introduces a full discrete-time stochastic epidemic model which uses the new models for the genetic distances. Section 2.8 discusses inference of the model parameters.

2.3 Inferring transmission trees

In this section we will introduce the Worby et al. model [14] for inference of transmission trees. The Worby et al. model is a discrete time stochastic model designed to describe an outbreak of a pathogen on a hospital ward, which includes a genetic model to describe the genetic distances between sequenced isolates from colonised patients. This model can be used to construct a possible transmission tree, with estimated transmission times, from an outbreak of a communicable pathogen in a hospital ward setting, as well as estimating parameters such as the transmission rate, test sensitivity and probability of importation.

2.3.1 The Worby et al. model

Since data from a hospital setting are often collected daily, events such as admission and discharge of patients, and transmission of the pathogen are modelled to occur daily. The length of the study is L and the time $t = 0, 1, \ldots, L$, where the initial day of the study is considered to be $t = 0$. Over this period n patients are admitted to the ward, with the ward assumed to be empty at $t = 0$, and all the patients to have left by $t = L$. A specific patient, i, is admitted to the ward at time t_i^a and discharged at time t_i^d.

Each patient on the ward at time t is either susceptible or colonised. Each patient who is admitted to the ward is either admitted colonised, with probability p, independently of all other patients, or susceptible, with probability $1-p$. Colonised patients are those carrying the pathogen. Each patient, i, is subject to a number, ν_i, of tests at times $t_i^t = t_{i,1}^t, t_{i,2}^t, \ldots, t_{i,\nu_i}^t$, which give a set of results, $X_i = X_{i,1}, X_{i,2}, \ldots, X_{i,\nu_i}$, which are either positive or negative. Each pathogen test is assumed, independently, to have sensitivity z and specificity 1, meaning that a colonised patient is tested positive with probability z and an uncolonised patient is always screened negative.

The probability of a specific susceptible becoming colonised on day t depends on the number of colonised patients on the ward and it is assumed that all colonised patients are equally infective, and all uncolonised patients are equally susceptible. The probability that a specific susceptible patient avoids colonisation on day t is $P(\text{avoid(t)}) = \exp(-\beta C(t))$ where $C(t)$ is the number of colonised patients on the ward on day t. If the patient does not avoid colonisation on day t then a source for their colonisation can be picked uniformly at random from the set of colonised patients, and therefore the probability of a given susceptible patient being colonised by a given carrier of the pathogen is given by

$$\frac{1-\exp(-\beta C(t))}{C(t)}.$$

The number of colonised patients, $C(t)$, is the total of all patients on the ward on day t who were colonised on or before day $t-1$ and those who are imported on day t or before. Once a patient is colonised they stay colonised for the rest of their stay on the ward and are included in the colonised population from the day after their colonisation, $t_i^c + 1$, until the day of their discharge, t_i^d.

The model also includes a genetic model for any genetic sequences which are obtained from the pathogen isolates taken from colonised patients. Each patient, i, who has one or more positive test result may have ζ_i isolates sampled and sequenced on days $t_i^s = t_{i,1}^s, t_{i,2}^s, \ldots, t_{i,\zeta_i}^s$. The model describes the differences between these sequences rather than the sequences themselves, so for each sequence from each patient a genetic distance is drawn to each other sequence (from this patient and each other patient) that was sampled at an earlier time. Each of these distances is modelled as being drawn from a probability distribution which depends on where the two patients from whom the sequences were sampled are in the transmission tree in relation to each other. The probability distributions for the two versions of the Worby et al. model are outlined in section 2.5.

2.3.2 Assumptions in the Worby et al. model

This Worby et al. model makes a number of assumptions:

- It is assumed that a transmission event is a population bottleneck, so that only a very small number of the bacterial population present in the source patient are transmitted to the newly infected patient. A population bottleneck is an event during which only one strain of the genetically diverse bacterial population which exists in the source patient is transmitted to the newly colonised patient.
- However, genetic diversity is modelled as a consequence of transmission, so it is assumed that diversity is gained at, or shortly after, the transmission event. Therefore if we observe the sequences of both the patient who first had the infection, and the patient whom they colonised in a ward, we would expect them to have very similar, but not necessarily identical, genetic sequences.
- It is assumed that each sequence that we observe is representative of the particular patient's colonisation as a whole.
- It is further assumed that the genetic distances between sequences from patients in a connected transmission chain are independent of each other.

2.3.3 Advantages of modelling the genetic distances rather than sequences

Genetic sequences are compared by aligning them and counting the number of positions on the nucleotide that differ between each sequence. We call the resulting number a genetic distance (also sometimes referred to as a 'genetic difference'). The Worby et al. model and our work use just these genetic distances rather than the observed sequences themselves. Working simply with the genetic distances from the observed sequences is beneficial because modelling how the mutations in the sequences actually occur with a microevolution model would require far more assumptions about unobserved underlying processes.

The distance parameters that we use in modelling the genetic distances can represent diversity between sequences due to any of a number of factors: SNPs, recombination events, or even colonisation by multiple strains in one host. More complex microevolution models, however, often ignore the possibility of recombination and multiple infections in order to make modelling the mutation dynamics possible. It is

also impractical to work with the sequences themselves due to the large number of nucleotide states that would need to be stored and modelled in order to model the likelihood of observing each particular sequence. For example, a typical sequence from an isolate of MRSA is 2.8- 2.9 million base pairs in length [39] and, although most of these will not mutate, even in the fairly small dataset which we introduce in chapter 4 there are 2591 locations where the base pairs do differ between sequences. In order to use the sequences themselves we would need to work with vectors for each sample which were at least this long.

2.4 Assessing the assumptions made in the modelling of genetic distances

The Worby et al. model for the genetic distances between sequences in a transmission chain assumes that the genetic distances between pairs of sequences are independent of each other. In reality this can not be the case because in a chain of three sequences (see figure 2.1), the distance between the first and third sequence, denoted by $d_{1,3}$, is bounded by the distance between the first and the second ($d_{1,2}$), and the second and third ($d_{2,3}$). Clearly $d_{1,3}$ can be at most the sum of $d_{1,2}$ and $d_{2,3}$ (this maximum distance is achieved when different nucleotides mutate between each pair of consecutive sequences, ie. x nucleotides mutate between sequence 1 and 2, and y different nucleotides mutate between sequences 2 and 3, so $d_{1,3} = x + y$), and $d_{1,3}$ can not be smaller than the difference between $d_{1,2}$ and $d_{2,3}$ (this minimum distance is achieved when we have the largest possible number of reversions, such as nucleotide 1 in figure 2.1, happening during the second transmission). Hence,

$$|d_{1,2} - d_{2,3}| \leq d_{1,3} \leq \min(d_{1,2} + d_{2,3}, N) \tag{2.4.1}$$

where N is the number of nucleotides in each sequence.

Another common assumption in models [18, 19, 53] for genetic variation between sequences is that it is appropriate to neglect the possibility that the same nucleotide will mutate more than once during a chain of transmission. This assumption is made because the probability of any mutation happening is so small, so the probability that a mutation will happen twice in the same position during the period of study is deemed unlikely enough to be ignored. Hence, the possibility of a nucleotide mutating during an unobserved transmission and then reverting to its original state in the next transmission is discounted. Although this is not an assumption made ex-

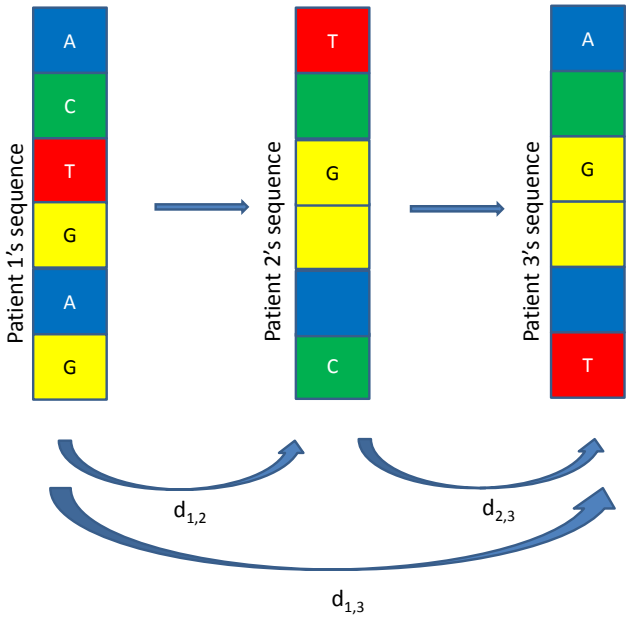


Figure 2.1: A diagram to show a chain of transmission between patients 1, 2 and 3 where each patient has a sequenced isolate which is 6 nucleotides long. In this case $d_{1,2} = 3$, $d_{2,3} = 2$ and $d_{1,3} = 2$. The nucleotide in position 1 changes once and then reverts to its original state, the nucleotide in position 3 changes once and stays changed, and the nucleotide in position 6 changes twice, each time to a different base.

plicitly in the Worby et al. model we now explore the validity and impact of the two assumptions stated here both together and separately.

2.4.1 The impact of common assumptions

We now investigate how much effect the assumptions that (i) nucleotides can only change once and (ii) genetic distances are independent have in models of genetic variation by defining a full joint probability distribution for the observed distances and comparing it to the distributions obtained under these assumptions. We assume that we have a chain of transmission in which consecutive patients infect each other, so patient 1 infects patient 2 who infects patient 3 and so on. We assume that we observe each of these patients, but some of the transmissions may be indirect, meaning that there could have been, say, an unobserved transmission event between patients 1 and 2 so that patient 1 directly infected patient 1_b who in turn directly infects patient 2. We assume that each nucleotide in the sequence mutates between two observed sequences i and $i+1$ with a probability θ_i which depends on the number of unobserved transmission events between these sequences. If we have N nucleotides in a sequence that are assumed to mutate independently of each other, we thus have N

independent Bernoulli trials, so we assume a Bin (N, θ_i) distribution for the number of mutations between these two consecutively observed sequences. Clearly the first step in the chain has probability P [Bin $(N, \theta_1) = d_{1,2}$]. If a mutation to each different base is assumed to be equally likely then at the following stage in the chain, each nucleotide which has already changed can do one of three things: revert to its previous state with probability $\frac{\theta_{i+1}}{3}$ (see first nucleotide in figure 2.1), make a further change to one of the two bases which it has not been with probability $\frac{2\theta_{i+1}}{3}$ (see sixth nucleotide in figure 2.1), or remain in its changed state with probability $1 - \theta_{i+1}$ (see third nucleotide in figure 2.1).

The genetic distance between consecutive sequences i and $i+1$ is modelled by the random variable $D_{i,i+1}$. The probability of observing the three genetic distances $D_{1,2} = d_{1,2}, D_{2,3} = d_{2,3}$ and $D_{1,3} = d_{1,3}$ becomes

$$\mathrm{P}\left[(D_{1,2}, D_{2,3}, D_{1,3}) = (d_{1,2}, d_{2,3}, d_{1,3})\right] = \mathrm{P}\left[\mathrm{Bin}\left(N, \theta_1\right) = d_{1,2}\right] \times f\left(d_{1,2}, d_{2,3}, d_{1,3}, \theta_2\right), \tag{2.4.2}$$

where, for $d_{1,3} \in \mathbb{Z}_+$ and $|d_{1,2} - d_{2,3}| \leq d_{1,3} < (d_{1,2} + d_{2,3} + |d_{1,2} - d_{2,3}|)/2$,

$$\begin{aligned} & f\left(d_{1,2}, d_{2,3}, d_{1,3}, \theta_2\right) = \\ & \sum_{j=0}^{(d_{1,3}-|d_{1,2}-d_{2,3}|)/2} \left[\binom{d_{1,2}}{\max\left(d_{1,2}, d_{2,3}\right) - d_{1,3} + j} \left(\frac{\theta_2}{3}\right)^{\max(d_{1,2}, d_{2,3}) - d_{1,3} + j} \right. \\ & \times \binom{d_{1,2} - \max\left(d_{1,2}, d_{2,3}\right) + d_{1,3} - j}{d_{1,3} - |d_{1,2} - d_{2,3}| - 2j} \left(\frac{2\theta_2}{3}\right)^{d_{1,3} - |d_{1,2} - d_{2,3}| - 2j} \\ & \left. \times \left(1 - \theta_2\right)^{|d_{1,2} - \min(d_{1,2}, d_{2,3})| + j} \mathrm{P}[\mathrm{Bin}\left(N - d_{1,2}, \theta_2\right) = |d_{1,2} - \max\left(d_{1,2}, d_{2,3}\right)| + j] \right], \end{aligned}$$

and for $d_{1,3} \in \mathbb{Z}_+$ and $(d_{1,2} + d_{2,3} + |d_{1,2} - d_{2,3}|)/2 \leq d_{1,3} \leq d_{1,2} + d_{2,3}$,

$$\begin{aligned} & f\left(d_{1,2}, d_{2,3}, d_{1,3}, \theta_2\right) = \\ & \sum_{j=0}^{(d_{1,2}+d_{2,3}-d_{1,3})/2} \left[\binom{d_{1,2}}{d_{1,2} + d_{2,3} - d_{1,3} - 2j} \left(\frac{2\theta_2}{3}\right)^{d_{1,2}+d_{2,3}-d_{1,3}-2j} \right. \\ & \times \binom{d_{1,3} + 2j - d_{2,3}}{j} \left(\frac{\theta_2}{3}\right)^{j} \\ & \left. \times \left(1 - \theta_2\right)^{d_{1,3}+j-d_{2,3}} \mathrm{P}\left[\mathrm{Bin}\left(N - d_{1,2}, \theta_2\right) = d_{1,3} - d_{1,2} + j\right] \right]. \end{aligned}$$

The sum over j in the formulae accounts for the fact that if $d_{1,3} - |d_{1,2} - d_{2,3}| > 1$ then there is more than one way to achieve the specific $d_{1,2}$, $d_{2,3}$ and $d_{1,3}$. We sum over the probabilities of each of these happening to give the total $\Pr[d_{1,2}, d_{2,3}, d_{1,3}]$. If the difference between the first and third sequences, $d_{1,3}$, is small ($< \frac{d_{1,2}+d_{2,3}+|d_{1,2}-d_{2,3}|}{2}$) then that will have been achieved mostly by the same nucleotides changing in the second transmission as in the first transmission, whereas if $d_{1,3}$ is larger ($\geq \frac{d_{1,2}+d_{2,3}+|d_{1,2}-d_{2,3}|}{2}$) there will have been more 'new' nucleotides changing on the second transmission, and hence we have different formulas for the probability of each case. In each case we construct each way in which the combination $d_{1,2}$, $d_{2,3}$, $d_{1,3}$ could have occurred by choosing, in the second transmission, the number of nucleotides which changed again, the number that reverted to the original state, and the number which stayed as they were. Clearly any further mutations happened in previously unchanged nucleotides and are therefore modelled by the binomial distribution with adjusted N.

In order to compare this joint distribution $\Pr[D_{1,2}, D_{2,3}, D_{1,3}]$ to the distribution implied by the model where each genetic difference between a pair of sequences is assumed independent, we compare the distributions of $\Pr[D_{1,3}|D_{2,3}, D_{1,2}]$ and $\Pr[D_{1,3}]$ under the binomial assumption. The conditional probability is easily extracted from the joint probability given above through $\Pr[D_{1,3}|D_{1,2}, D_{2,3}] = \frac{\Pr[D_{1,2},D_{2,3},D_{1,3}]}{\Pr[D_{1,2}]\Pr[D_{2,3}]}$:

$$
\begin{aligned}
&\mathrm{P}[D_{1,3} = d_{1,3} | D_{1,2} = d_{1,2}, D_{2,3} = d_{2,3}] \\
&= \frac{\mathrm{P}[(D_{1,2}, D_{2,3}, D_{1,3}) = (d_{1,2}, d_{2,3}, d_{1,3})]}{\mathrm{P}[D_{1,2} = d_{1,2}]\,\mathrm{P}[D_{2,3} = d_{2,3}]} \\
&= \frac{\mathrm{P}[\mathrm{Bin}(N, \theta_1) = d_{1,2}] f\,(d_{1,2}, d_{2,3}, d_{1,3}, \theta_2)}{\mathrm{P}[\mathrm{Bin}(N, \theta_1) = d_{1,2}]\,\mathrm{P}[\mathrm{Bin}(N, \theta_2) = d_{2,3}]} \\
&= \frac{f\,(d_{1,2}, d_{2,3}, d_{1,3}, \theta_2)}{\mathrm{P}[\mathrm{Bin}(N, \theta_2) = d_{2,3}]}.
\end{aligned}
$$

Figure 2.2 shows the conditional distribution for $d_{1,3}$ in an example where $d_{1,2} = 9$ and $d_{2,3} = 9$, $\boldsymbol{\theta} = (0.0002615, 0.0002615)$ and $N = 5354$ (the values for $\boldsymbol{\theta}$ and N here are taken from Ypma et al. [19]) and figure 2.3 shows the independent distribution for the same case. For all credible values of the parameters (ie. N large enough to be the number of nucleotides in a genetic sequence, and θ small enough to be the probability of a nucleotide mutating) over a wide range of distances the independent distribution always displays far more variability than the conditional. In tables 2.1 and 2.2 we compare the mean and variance for the conditional and independent distributions over different values for $D_{1,2}$ and $D_{2,3}$ for two possible combinations of N and θ.

$N = 5354$, $\theta = 0.0002615$	$D_{1,2} = 1$, $D_{2,3} = 9$	$D_{1,2} = 9$, $D_{2,3} = 9$	$D_{1,2} = 20$, $D_{2,3} = 9$	$D_{1,2} = 70$, $D_{2,3} = 50$	$D_{1,2} = 1$, $D_{2,3} = 0$
Conditional mean	9.997759	17.97983	28.95517	119.1284	1
Independent mean	1.400071	1.400071	1.400071	1.400071	1.400071
Conditional variance	0.003417146	0.03017241	0.06692719	1.28174	0
Independent variance	1.399705	1.399705	1.399705	1.399705	1.399705

Table 2.1: The mean and variance of $D_{1,3}$ when $N = 5354$ and $\theta = 0.0002615$ conditional of different values of $D_{1,2}$ and $D_{2,3}$ compared to the mean and variance of the probability mass function for $D_{1,3}$ when we assume that it is independent of the other two distances.

Table 2.1 takes values for N and θ from Ypma et al. [19] and table 2.2 takes values from Worby [69].

$N = 2591$, $\theta = 0.022$	$D_{1,2} = 1$, $D_{2,3} = 9$	$D_{1,2} = 9$, $D_{2,3} = 9$	$D_{1,2} = 20$, $D_{2,3} = 9$	$D_{1,2} = 70$, $D_{2,3} = 50$	$D_{1,2} = 1$, $D_{2,3} = 0$
Conditional mean	9.995369	17.95832	28.90737	118.1989	1
Independent mean	57.002	57.002	57.002	57.002	57.002
Conditional variance	0.006925675	0.06216	0.06692719	2.592574	0
Independent variance	53.747956	53.747956	53.747956	53.747956	53.747956

Table 2.2: The mean and variance of $D_{1,3}$ when $N = 2591$ and $\theta = 0.022$ conditional of different values of $D_{1,2}$ and $D_{2,3}$ compared to the mean and variance of the probability mass function for $D_{1,3}$ when we assume that it is independent of the other two distances.

2.4.2 Assessing the validity of the assumptions

Figures 2.2 and 2.3 show that the conditional distribution placed the vast majority of the probability on the value $d_{1,3} = d_{1,2} + d_{2,3}$. This lends credence to the assumption mentioned earlier: that the probability of the same nucleotide changing twice in a transmission chain is so small that it can be discounted. This assumption would

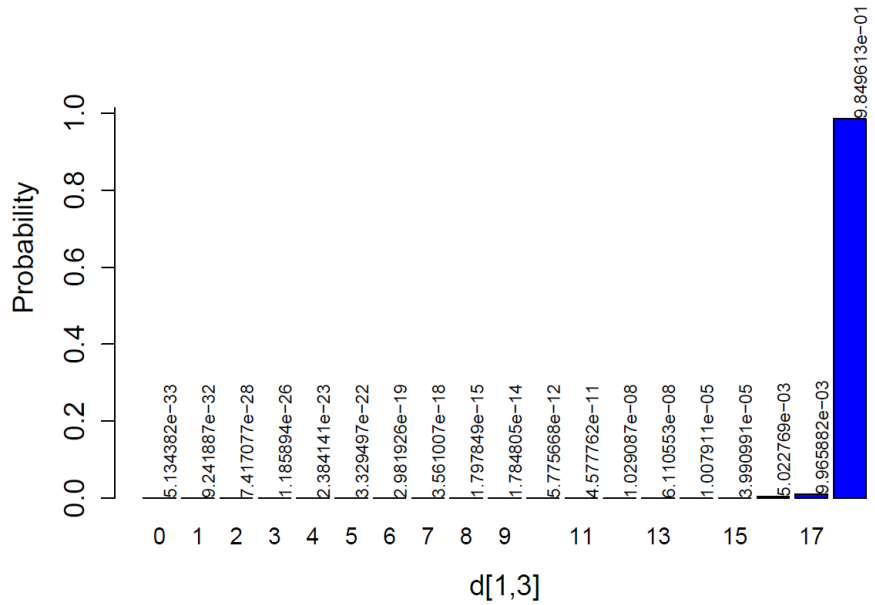


Figure 2.2: $\Pr(D_{1,3}|D_{1,2} = 9, D_{2,3} = 9)$ The probability mass function for the genetic distance between sequences from the first and third patient in a transmission chain given that the distances between the sequences from the first and second, and second and third patient are both 9. We assume that each of the 5354 nucleotides in the genetic sequence can mutate between two consecutive sequences with probability $\theta = 0.0002615$.

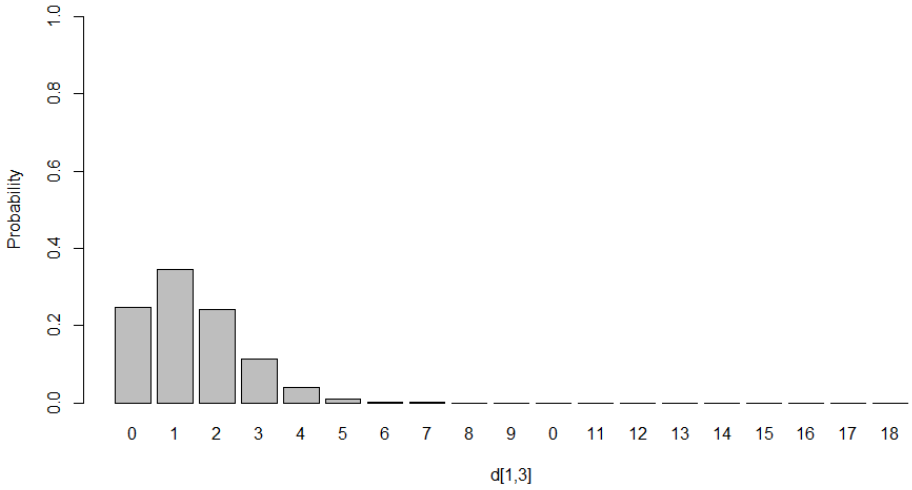


Figure 2.3: The probability mass function for $D_{1,3}$ when we assume that the distance between sequences from the first and third patient is independent of the distances between sequences from the first and second patient, and second and third patient. We show $D_{1,3}$ in the range $0, 1, \ldots, 18$ which are the theoretically possible values when $D_{1,2} = 9$ and $D_{2,3} = 9$

.

greatly simplify our problem, and the joint probability in this case is simply

$$\mathrm{P}[(D_{1,2}, D_{2,3}, D_{1,3}) = (d_{1,2}, d_{2,3}, d_{1,3})] = \mathrm{P}[\mathrm{Bin}\,(N, \theta_1) = d_{1,2}] \times \mathrm{P}[\mathrm{Bin}\,(N - d_{1,2}, \theta_2) = d_{2,3}], \tag{2.4.3}$$

which can easily be generalised to a k-sequence transmission chain by

$$\begin{aligned}\mathrm{P}[(D_{1,2},&D_{2,3}, D_{1,3}, ..., D_{1,k}, ..., D_{k-1,k}) = (d_{1,2}, d_{2,3}, d_{1,3}, ..., d_{1,k}, ..., d_{k-1,k})] \\ &= \mathrm{P}[\mathrm{Bin}\,(N, \theta_1) = d_{1,2}] \cdot \mathrm{P}[\mathrm{Bin}\,(N - d_{1,2}, \theta_2) = d_{2,3}]... \\ &\times \mathrm{P}[\mathrm{Bin}\,(N - d_{1,2} - d_{2,3} - ... - d_{k-2,k-1}, \theta_{k-1}) = d_{k-1,k}].\end{aligned}$$

From this joint probability we see that the conditional probability of, say, $D_{1,3}$ given $D_{1,2}$ and $D_{2,3}$ is $\mathrm{P}[(D_{1,3}|D_{1,2}, D_{2,3}) = (d_{1,3}|d_{1,2}, d_{2,3})] = \mathbb{1}_{\{d_{1,3}=d_{1,2}+d_{2,3}\}}$. This is close to the random variable whose probability mass function is shown in 2.2 so this assumption appears much more plausible than the assumption of independent distances. It is notable that even for values for which the means of the two distributions are close, the variation of the independent distribution is always far higher than the conditional variance.

Having seen that the assumption of independence between distances is not realistic, we are motivated to create a new model for genetic distances to improve upon the Worby et al. model.

2.5 The genetic distance models in the Worby et al. model

There are two versions of the Worby et al. model, which were briefly introduced in chapter 1. We recap them here for ease. The two models differ in the modelling of the genetic distances. The principle behind both is that the genetic distances between sequences from two patients who are closely linked are more likely to be similar than between sequences from two patients who are not closely linked. Exactly what constitutes 'closely linked' is what separates the two models. In each model the genetic distances are assumed to be drawn from a different set of probability distributions which depend on the inferred relationship between the two patients from whom the sequences were collected. Both models are briefly outlined here.

2.5.1 The Importation Structure Worby et al. model

The Importation Structure model assumes that there are different strains of the pathogen that isolates can belong to. It is assumed that a patient who is colonised during their

stay on the ward has a strain of the pathogen which belongs to the same group as the pathogen of the patient who infected them. Patients who arrive on the ward already colonised have a pathogen which belongs to a group which is already represented with probability c (where c is called the 'clustering' parameter), or which belongs to a new group which has not yet been observed on the ward. The total number of groups, and the number of patients in each group, is unknown and needs to be inferred. The probability of the genetic distance, $\Psi_{i,j}$, between any pair of isolates is given by

$$\mathrm{P}(\Psi_{i,j} = x) = \begin{cases} \mu(1-\mu)^x & \text{if } i \text{ and } j \text{ are same type} \\ \mu_G(1-\mu_G)^x & \text{otherwise} \end{cases}$$

where $x = 0, 1, \ldots$ and μ, $\mu_G \in [0, 1]$. In reality x can not be larger than the length of the genome L, but since L is very large this has no effect because $\sum_{x=o}^{L} \mathrm{P}(\Psi_{i,j} = x) \approx 1$.

2.5.2 The Transmission Chain Diversity Worby et al. model

The second model, the Transmission Chain Diversity model, assumes that SNPs accumulate over time and that isolates in separate transmission chains are unrelated, so any two importations will be unrelated to each other. The probability of the genetic distance between a pair of sequences is then

$$\mathrm{P}(\Psi_{i,j} = x) = \begin{cases} \mu\gamma^{t(r(i),r(j))}(1-\mu\gamma^{t(r(i),r(j))})^x & \text{if } i \text{ and } j \text{ are in the same tree} \\ \mu_G(1-\mu_G)^x & \text{otherwise} \end{cases}$$

where $r(i)$ is the patient whom the ith sequence belongs to, $t(r(i), r(j))$ is the time between colonisation of the patients, γ is the transmission diversity factor, and $\gamma \in (0, 1)$ accounts for increasing distance over multiple transmission events.

2.6 Relaxing the assumption of independence

In both variations of the Worby et al. model it is assumed that the genetic distances along a chain are all drawn independently from geometric distributions. We have previously discussed the limitations of this assumption of independence, and therefore we introduce three new models for the genetic distances that take into account the dependence between distances in the same chain. The rationale behind the models, that the genetic distance between two related patients is likely to be smaller than the genetic distance between two unrelated patients, remains the same, but the precise distributions from which these distances are drawn differ from model to model.

2.6.1 Three new models for a genetic distance matrix

These new models define the distributions from which the genetic distances between sequences are assumed to be drawn. We will refer to the patient from which a specific sequence i was taken as H_i, so H_i was the host for sequence i. Alternatively, we can refer to the sequences taken from a specific patient j as $Q_j = \{Q_{j,1}, \ldots, Q_{j,\zeta_j}\}$, so sequences Q_j were sampled from patient j. In defining these new models we use the terms 'distinct' or 'separate' transmission chains to refer to transmission chains which have different roots, so each transmission chain originates with a different patient who is colonised before admission to the ward. Patients in the same transmission chain are those who directly or indirectly infect each other eg. $H_i \to H_j$ or $H_i \to \cdots \to H_j$.

The models use Poisson distributions because if the number of mutations is assumed to be binomially distributed $\sim \text{Bin}(N, p)$ then due to the large value of N and small value of p we can approximate this as a Poisson distribution with parameter $\lambda = Np$ since:

$$\begin{aligned}
\binom{N}{x} p^x (1-p)^{N-x} &= \frac{N(N-1)\ldots(N-x+1)}{x!} \left(\frac{\lambda}{N}\right)^x \left(1-\frac{\lambda}{N}\right)^{N-x} \\
&= \frac{N(N-1)\ldots(N-x+1)}{N^x} \frac{\lambda^x}{x!} \left(1-\frac{\lambda}{N}\right)^N \left(1-\frac{\lambda}{N}\right)^{-x} \\
&\approx \frac{\lambda^x}{x!} \left(1-\frac{\lambda}{N}\right)^N \quad \text{if } N \text{ is much larger than } x, \\
&\approx \frac{\lambda^x}{x!} \exp(-\lambda) \quad \text{if } N \text{ is large.}
\end{aligned}$$

2.6.1.1 The Chain Error model

The first new model, the Chain Error model, is based on the idea that patients who are in distinct transmission chains, and are thus unrelated, have the genetic distances between their sequences drawn from a Poisson distribution with parameter θ_{gl} and patients who share a direct transmission event have the genetic distances between their sequences drawn from a Poisson distribution with parameter θ. If there are patients who had more than one isolate sampled then the genetic distance between the within-host sequences is drawn from a Poisson distribution with parameter θ_i.

Patients who are in the same transmission chain but are separated by more than one transmission event have genetic distances between their sequences which are on average equal to the sum of the underlying distances in the chain which separates them,

with variance that will increase with the length of the chain. For sequences i and j which are taken from patients who are separated by $k > 1$ transmission events, we define $D_{i,j}$ to be the sum of the genetic distances between sequences from consecutive patients who make up the underlying transmission chain. If H_i and H_j are separated by k transmission events such that H_i colonises p_1 who colonises p_2 etc. ($H_i \to p_1 \to p_2 \to \cdots \to p_{k-1} \to H_j$) then $D_{i,j} = \sum_{r=0}^{k-1} \Psi_{Q_{p_r,1},Q_{p_{r+1},1}}$ where $p_0 = H_i$, $p_k = H_j$ and $\Psi_{Q_{p_1,1},Q_{p_2,1}}$ is the genetic distance between sequences $Q_{p_1,1}$ and $Q_{p_2,1}$ which are the first sequences taken from patients p_1 and p_2. The genetic distance between i and j is defined as $D_{i,j} + \xi W$ where $\mathrm{P}(\xi = 1) = \mathrm{P}(\xi = -1) = 0.5$. W is a Poisson random variable with parameter $k\gamma$ truncated at $D_{i,j}$, in order to ensure that the distance can not be negative. The random variables ξ and W are independent.

To derive the conditional probability distribution for $\Psi_{i,j}$ when $k > 1$ we define

$$\begin{aligned} p_j = \mathrm{P}(W = j) &= \frac{\mathrm{P}(\mathrm{Pois}(k\gamma) = j)}{\sum_{l=0}^{D_{i,j}} \mathrm{P}(\mathrm{Pois}(k\gamma) = l)} \\ &= \frac{(k\gamma)^j}{j! \sum_{l=0}^{D_{i,j}} \frac{(k\gamma)^l}{l!}}. \end{aligned}$$

Therefore

$$\mathrm{P}(X = D_{i,j}) = P(W = 0) = p_0, \tag{2.6.1}$$

$$\mathrm{P}(X = D_{i,j} + q) = \frac{1}{2} p_q = \frac{1}{2} p_{(x - D_{i,j})} \qquad (q = 1, 2, \ldots, D_{i,j}), \tag{2.6.2}$$

$$\mathrm{P}(X = D_{i,j} - q) = \frac{1}{2} p_q = \frac{1}{2} p_{(D_{i,j} - x)} \qquad (q = 1, 2, \ldots, D_{i,j}). \tag{2.6.3}$$

Equations 2.6.2 and 2.6.3 can be combined to give $\mathrm{P}(X = x) = \frac{1}{2} p_{|D_{i,j} - x|}$ for $x \neq D_{i,j}$. Therefore, adding an indicator function gives us $\mathrm{P}(X = x) = p_{|D_{i,j} - x|} \left(\frac{1}{2}\right)^{\mathbb{1}_{\{x \neq D_{i,j}\}}}$ for all $x \leq 2D_{i,j}$. This restriction on x is due to the truncation which ensures that the genetic distance can not be negative.

If $\Psi_{i,j}$ is the genetic distance between sequences i and j, and k is the number of transmission events that separates the patients from which the sequences i and j were taken ($k = \infty$ if the patients are not in the same chain), then the genetic distances for pairs of sequences which are from the same patient, or share a direct transmission event, or are unrelated are drawn, independently, from the following distributions:

$$\mathrm{P}(\Psi_{i,j} = x) = \begin{cases} (\theta_{gl}^x / x!) \exp(-\theta_{gl}) & \text{if } k = \infty \\ (\theta_i^x / x!) \exp(-\theta_i) & \text{if } k = 0 \\ (\theta^x / x!) \exp(-\theta) & \text{if } k = 1 \end{cases} \tag{2.6.4}$$

where $x = 0, 1, \ldots$. Again, x in practice can not be larger than the length of the genome L, but the fact that L is so large means that $\sum_{x=o}^{L} \mathrm{P}(\Psi_{i,j} = x) \approx 1$.

The conditional probability distribution for genetic distances for pairs of sequences which are in the same chain but separated by more than one transmission event is:

$$\mathrm{P}(\Psi_{i,j} = x | \Psi_{a,b} = \psi_{a,b} : \mathrm{trans}(a,b) = 1) = \frac{(k\gamma)^{|D_{i,j}-x|}}{|D_{i,j}-x|! \left(\sum_{l=0}^{D_{i,j}} (k\gamma)^l / l!\right)} \left(\frac{1}{2}\right)^{\mathbb{1}_{\{x \neq D_{i,j}\}}} \mathbb{1}_{\{x \leq 2D_{i,j}\}} \quad \text{if } k > 1, \tag{2.6.5}$$

where $\mathrm{trans}(a,b)$ is the number of transmission events separating sequences a and b and $\psi_{a,b}$ is the genetic distance between sequences a and b. The number $D_{i,j}$ is the sum of underlying genetic distances in the chain between i and j for which $\mathrm{trans}(a,b) = 1$.

The joint distribution for all genetic distances in the transmission tree is simply the product of, firstly, the marginal distributions for those genetic distances for which $\mathrm{trans}(i,j) \in \{0, 1, \infty\}$, which are given in equation 2.6.4, and, secondly, the marginal distributions for those genetic distances for which $\mathrm{trans}(i,j) > 1$, which are given in equation 2.6.5 and are conditional on the first set of marginal distributions. Therefore, the joint distribution is:

$$\mathrm{P}\left(\bigcap_{(i,j)\in G} \{\Psi_{i,j} = \psi_{i,j}\}\right) = \left(\prod_{(i,j)\in G_1} \mathrm{P}\left(\Psi_{i,j} = \psi_{i,j}\right)\right) \left(\prod_{(i,j)\in G_2} \mathrm{P}\left(\Psi_{i,j} = \psi_{i,j} | \Psi_{a,b} = \psi_{a,b} : (a,b) \in \{0,1,\infty\}\right)\right), \tag{2.6.6}$$

where $G = \{(i,j) : i < j, j \leq n_{seqs}\}$ is the set of indices for all genetic distances, including the unobserved distances, with n_{seqs} being the total number of sequences in the genetic distance matrix, including the unobserved sequences. The sets G_1 and G_2 are: $G_1 = G \cap \{(i,j) : \mathrm{trans}(i,j) \in \{0,1,\infty\}\}$ and $G_2 = G \cap \{(i,j) : \mathrm{trans}(i,j) > 1\}$. The terms in the products are given by equations 2.6.4 and 2.6.5.

2.6.1.2 The Chain Poisson model

The second new model can be thought of as a variation of the Chain Error model described above; under this model the genetic distance between two sequences from

patients who are in the same chain but separated by more than one transmission event will equal $D_{i,j}$ on average, with a variance that will increase with $D_{i,j}$. These genetic distances are drawn from a Poisson distribution with parameter equal to the sum of the underlying distances in the transmission chain, $D_{i,j}$. As before, the genetic distances between sequences from patients who are in separate transmission chains are drawn from a Poisson distribution with parameter θ_{gl}, and the genetic distances between sequences from those patients who share a direct transmission event are drawn from a Poisson distribution with parameter θ. If there are patients who had more than one isolate sampled then the genetic distance between the within-host sequences is drawn from a Poisson distribution with parameter θ_i.

Under this Chain Poisson model, the the genetic distances for pairs of sequences which are from the same patient ($k = 0$), or share a direct transmission event ($k = 1$), or are unrelated ($k = \infty$) are drawn, independently, from the following:

$$\mathrm{P}(\Psi_{i,j} = x) = \begin{cases} (\theta_{gl}^x / x!) \exp(-\theta_{gl}) & \text{if } k = \infty \\ (\theta_i^x / x!) \exp(-\theta_i) & \text{if } k = 0 \\ (\theta^x / x!) \exp(-\theta) & \text{if } k = 1. \end{cases} \tag{2.6.7}$$

The conditional probability distribution for genetic distances for pairs of sequences which are in the same chain but separated by more than one transmission event is:

$$\mathrm{P}(\Psi_{i,j} = x | \Psi_{a,b} = \psi_{a,b} : \mathrm{trans}(a,b) = 1) = (D_{i,j}^x / x!) \exp(-D_{i,j}) \quad \text{if } k > 1, \tag{2.6.8}$$

where $\mathrm{trans}(a,b)$ is the number of transmission events separating sequences a and b and $\psi_{a,b}$ is the genetic distance between sequences a and b. The number $D_{i,j}$ is the sum of underlying genetic distances in the chain between i and j for which $\mathrm{trans}(a,b) = 1$.

Here we simply draw the genetic distances for sequences in the same chain from a Poisson distribution with parameter $D_{i,j} = \sum_{r=0}^{k-1} \Psi_{Q_{p_r,1},Q_{p_{r+1},1}}$ where $p_0 = H_i$, $p_k = H_j$ and $\Psi_{Q_{p_r,1},Q_{p_{r+1},1}}$ is the genetic distance between the first sequence taken from patient p_r and the first sequence taken from patient p_{r+1}. So $D_{i,j}$ is the sum of the distances between the sequences from patients involved in the consecutive transmission events between the patients with sequences i and j. In the Chain Error model we assume that the distance will be equal to $D_{i,j}$ with some error, the size of which depends on the length of the underlying chain, whereas in this version the mean is equal to $D_{i,j}$ and the variance around this grows as the distance grows (as the mean and variance of a Poisson distribution are equal). Therefore the length of the chain is implicitly included in this model, as we assume that the total distance will increase as the length

of the chain increases. This version of the model has fewer parameters and leads to simpler likelihood expressions when fitting the model to data.

The joint distribution for all genetic distances in the transmission tree is simply the product of, firstly, the marginal distributions for those genetic distances for which $\text{trans}(i,j) \in \{0,1,\infty\}$, which are given in equation 2.6.7, and, secondly, the marginal distributions for those genetic distances for which $\text{trans}(i,j) > 1$, which are given in equation 2.6.8 and are conditional on the first set of marginal distributions. Therefore, the joint distribution is again given by equation 2.6.6 with the terms in the products given by equations 2.6.7 and 2.6.8.

2.6.1.3 The Time Dependent Distances model

In the two models above it is indirectly assumed that the size of the difference between two genetic sequences will depend on the time that elapses between the patients' infections as well as their relative positions in the transmission tree. In the Chain Error model the inclusion of the parameter k (which is the number of transmission events between patients with sequences i and j) in the distribution for the genetic distance between two sequences from patients in the same transmission chain who are separated by at least two transmission events means that the variance of the size of the genetic distances will increase as the length of the chain between the two patients increases. In the Chain Poisson model the variance of the size of the genetic distances between sequences from patients who are separated by more than one transmission in the same chain increases as the sum of the underlying distances in the chain increases, which intuitively suggests that the distance will increase as the time between the infections of the patients increases. Therefore, we introduce a model which includes time dependence explicitly in the model for the genetic distance between sequences from an infector and its direct infectee.

The Time Dependent Distances model is based on the idea that the genetic distance between sequences from a patient and the patient who infected them will have some dependence on the time between the sampling of the isolates from the patients. This is because the pathogen is multiplying and mutating within the first host, H_i, from the time of sampling, s_i, until it is transmitted to the second host, H_j, during the transmission event at time I_{H_j}, and there it continues to multiply and mutate until it is sampled at time s_j. As we have modelled the time dependence at the level of the individual transmission events we keep the distribution from the Chain Poisson

model for the genetic distances between sequences from patients who are separated by more than one transmission event. As we observe the underlying distances that make up the transmission chain we want to utilise this information rather than throwing it away in order to model time dependence on this level explicitly. Patients who are in different transmission chains again have the genetic distance between their sequences drawn from a Poisson distribution with parameter θ_{gl} and any within host genetic distances are drawn from a Poisson distribution with parameter θ_i. Therefore the Time Dependent Distances model assumes that the genetic distances for pairs of sequences which are from the same patient ($k = 0$), or share a direct transmission event ($k = 1$), or are unrelated ($k = \infty$), are drawn, independently, from the following distributions:

$$\mathrm{P}(\Psi_{i,j} = x) = \begin{cases} (\theta_{gl}^x / x!) \exp(-\theta_{gl}) & \text{if } k = \infty \\ (\theta_i^x / x!) \exp(-\theta_i) & \text{if } k = 0 \\ ((t_{ij}\theta)^x / x!) \exp(-(t_{ij}\theta)) & \text{if } k = 1. \end{cases} \tag{2.6.9}$$

The conditional probability distribution for genetic distances for pairs of sequences which are in the same chain but separated by more than one transmission event is:

$$\mathrm{P}(\Psi_{i,j} = x | \Psi_{a,b} = \psi_{a,b} : \mathrm{trans}(a,b) = 1) = (D_{i,j}^x / x!) \exp(-D_{i,j}) \quad \text{if } k > 1, \tag{2.6.10}$$

where $\mathrm{trans}(a,b)$ is the number of transmission events separating sequences a and b and $\psi_{a,b}$ is the genetic distance between sequences a and b. The number $D_{i,j}$ is the sum of underlying genetic distances in the chain between i and j for which $\mathrm{trans}(a,b) = 1$. So $D_{i,j} = \sum_{r=0}^{k-1} \Psi_{Q_{p_r,1},Q_{p_{r+1},1}}$ where $p_0 = H_i$, $p_k = H_j$ and $\Psi_{Q_{p_r,1},Q_{p_{r+1},1}}$ is the genetic distance between the first sequence taken from patient p_r and the first sequence taken from patient p_{r+1}. So $D_{i,j}$ is the sum of the distances between the sequences from the patients involved in the consecutive transmission events between patient H_i and H_j. Some measure of the time difference related to the two sequences is included in this model through t_{ij}. This quantity can be defined differently depending on the setting of the outbreak to be modelled. For nosocomial infections it could be $t_{ij} = |t_i^s - t_j^s|$ where t_i^s and t_j^s are the sampling times of sequences i and j. If sampling times were not available the difference between the times of infections of the patients could be used as the measure of the time between sequences i and j.

The joint distribution for all genetic distances in the transmission tree is simply the product of, firstly, the marginal distributions for those genetic distances for which $\mathrm{trans}(i,j) \in \{0, 1, \infty\}$, which are given in equation 2.6.9, and, secondly, the marginal distributions for those genetic distances for which $\mathrm{trans}(i,j) > 1$, which are given in

equation 2.6.10 and are conditional on the first set of marginal distributions. Therefore, the joint distribution is again given by equation 2.6.6 on page 36, with the terms in the products given by equations 2.6.9 and 2.6.10.

In the next sections we describe in detail how this model for the genetic distances fits within the wider model for the spread of a pathogen.

2.7 The model for the spread of a pathogen

The three genetic models described in section 2.6 give the distributions from which we can assume the genetic distances between patients' sequences are drawn. This is one part of the whole stochastic model which describes the spread of a pathogen. The stochastic model which the genetic distance distribution fits into will vary depending on the setting and type of pathogen which is being modelled. Here we will describe such a model for the spread of a pathogen within a single hospital ward in discrete time over a study of length L. The initial day of the study is set as $t = 0$ and therefore, $t = 0, 1, \ldots, L$. For ease we assume that at time $t = 0$ the number of patients present on the ward is $n_t = 0$, although this could be relaxed by allowing patients present on the ward before the study began to have an 'admission' time equal to $t = 0$. Over the course of the study the total number of patients to be admitted and discharged is n. The model describes the dynamics of the spread of the pathogen on the level of individuals and can be used to estimate the transmission of the disease from patient to patient throughout the ward, as well as the times of transmission events and values of the parameters.

Each patient, i, is admitted to the ward at time t_i^a, colonised at time t_i^c ($t_i^c = \infty$ if the patient remains susceptible for the duration of their stay) and discharged at time t_i^d. A colonised patient is a patient who is carrying the pathogen at any body site, either asymptomatically or symptomatically. The model assumes that each patient is either positive (positive will mean that the patient is carrying the pathogen) on admission with probability p, or negative (not carrying the pathogen) with probability of $1 - p$, independently of all other patients. It is assumed that there is no background transmission, meaning that there is no ongoing contamination in the ward and there are no persistent carriers elsewhere in the hospital, including the staff on the ward, although this could easily be incorporated into the model if required.

All uncolonised patients are assumed to be equally susceptible, and all colonised pa-

tients to be equally infective. The probability of a specific susceptible patient avoiding colonisation on day t depends on the number of colonised patients on the ward, $C(t)$, so $\mathrm{P}(\mathrm{avoid}(\mathrm{t})) = \exp(-\beta C(t))$. Therefore if the patient does not avoid colonisation then a source can be picked uniformly at random from the set of colonised patients, and the probability of a given susceptible patient being infected by a given carrier of the pathogen is defined to be

$$\frac{1-\exp(-\beta C(t))}{C(t)}.$$

A colonised patient is regarded as infective from the day after infection, $t = t_i^c + 1$, until discharge from the ward at $t = t_i^d$. Each patient, i, has a number, ν_i, of test results which are either positive or negative, $X_i = X_{i,1}, X_{i,2}, \ldots, X_{i,\nu_i}$, which are taken at times $t_i^t = t_{i,1}^t, t_{i,2}^t, \ldots, t_{i,3}^t$. The sensitivity of the screening test is represented by the parameter z, so a colonised patient has a probability z of being screened positive each day that they receive a test independently of all other tests. The specificity of the test is assumed to be 100%, so all negative patients are screened negative.

Those patients who receive one or more positive test results may also have one or more isolates sequenced. A patient, i, with ν_i positive swabs has ζ_i sampled isolates sequenced on days $t_i^s = t_{i,1}^s, t_{i,2}^s, \ldots, t_{i,\zeta_i}^s$. The distributions from which the genetic distances between these sequences are assumed to be drawn have been given in section 2.6. The parameter vector, ρ, for our model is $\rho = \{p, z, \beta, \boldsymbol{\Theta}\}$ where $\boldsymbol{\Theta}$ is the vector of genetic diversity parameters.

2.8 Inference of parameters of the model for the spread of a pathogen

The model introduced in section 2.7 can be used to infer the transmission dynamics, transmission times and values of the parameters for data collected from a hospital ward. Such a data set should contain admission and discharge times for patients, pathogen swab results which are either positive or negative (a positive result means they are carrying the pathogen, since we do not allow for false-positives, although we do allow for false-negatives), and the distances between the genetic sequences taken from the patients' pathogen isolates.

Inferring the transmission dynamics means inferring the admission states of the patients and which patient was likely to have been the 'source' of a particular positive

patient's colonisation. The admission state, ϕ_i, of a patient, i, is whether the patient is already colonised when they are admitted to the ward ($\phi_i = 1$), or is susceptible on admission ($\phi_i = 0$). If there has been a contact (which may be indirect contact, not necessarily direct, physical contact) between two patients that has resulted in patient j being colonised by the pathogen which patient i was carrying then we define patient i to be the source for patient j. In order to completely specify the transmission tree we include $T = \{t^c, \phi, s, \Psi^a\}$, the vector of unobserved data consisting of unobserved colonisation times $t^c = (t^c_1, t^c_2, \ldots, t^c_{n_{pos}})$ for all positive patients n_{pos}, admission states $\phi = (\phi_1, \phi_2, \ldots, \phi_n)$ for all n patients, sources $s = (s_1, s_2, \ldots, s_{n_{acq}})$ for all patients who acquire the pathogen whilst on the ward n_{acq}, and unobserved genetic distances Ψ^a which are the distances which would be included in the genetic distance matrix if sequences had been observed for every colonised patient.

2.8.1 Model likelihood

Using our model, we now derive the likelihood of observing the set of genetic distances, Ψ, between patients' isolates, and the results of the screening tests carried out for each patient, X. This will allow us to estimate the transmission tree for the spread of the pathogen through the ward, which includes times and sources of each transmission event. Using the matrix of genetic distances, Ψ, which is straightforward to recover from a dataset that contains genome sequence data, seems a more intuitive method than considering the very small probabilities of observing each particular genetic sequence.

The model likelihood that we are interested in, $\pi(X, \Psi|\rho)$, is intractable because it requires integration over all possible colonisation times and sources. Therefore we augment the parameter space with unobserved data, T. The vector T consists of the times of every colonisation, t^c, sources of every colonisation, s, and every patient's importation status, ϕ. As we are working with the genetic distances between isolates instead of the sequences themselves, this augmented likelihood is the likelihood of the observed matrix of distances, Ψ and MRSA screening results, X, given the model parameters, ρ, conditional on Z, the vector of observed data such as admission and

discharge times which are not explicitly modelled. From Bayes' Theorem we get:

$$\begin{aligned}\pi(\rho, T|X, \Psi, Z) &= \frac{\pi(\rho, T, X, \Psi, Z)}{\pi(X, \Psi, Z)} \\ &= \frac{\pi(X, \Psi|T, \rho, Z)\pi(T|\rho, Z)\pi(Z|\rho)\pi(\rho)}{\pi(X, \Psi|Z)\pi(Z)} \\ &= \frac{\pi(X, \Psi|T, \rho, Z)\pi(T|\rho, Z)\pi(\rho)}{\pi(X, \Psi|Z)}\end{aligned} \tag{2.8.1}$$

since $\pi(Z|\rho) = \pi(Z)$ because Z, the observed data that is not explicitly included in our model, is independent of the parameters, ρ. Equation 2.8.1 shows that the likelihood $\pi(\rho, T|X, \Psi, Z)$ is proportional to the likelihood of observing the distance matrix and screening results given the unobserved dynamics and parameters, $\pi(X, \Psi|T, \rho, Z)$, multiplied by the likelihood of the unobserved data given the parameters, $\pi(T|\rho, Z)$, multiplied by the prior distribution of the parameters, $\pi(\rho)$. In order to infer the whole transmission process a data-augmented MCMC routine can be used to sample the parameters ρ and the transmission dynamics, T, from $\pi(X, \Psi|T, \rho, Z)\pi(T|\rho, Z)$.

2.8.1.1 Genetic part of the model likelihood

$\pi(X, \Psi|T, \rho, Z)$ is the likelihood of observing the distance matrix and screening results given the unobserved dynamics and parameters and therefore this genetic part of the likelihood varies for each of the three models. The three variations are given below. We define n_{seqs} as the number of genetic sequences in the genetic distance matrix. The number of transmission events between patient H_i and patient H_j, which are the patients which give sequences i and j, is given by trans(i, j). This is calculated by looking at the sources of H_i and H_j and any intermediate patients in the transmission chain. If patients H_i and H_j are the same patient then trans$(i, j) = 0$, and if they are in separate transmission chains then trans$(i, j) = \infty$. If trans$(i, j) > 1$ then $D_{i,j} = \sum_{r=0}^{k-1} \Psi_{Q_{p_r,1}, Q_{p_{r+1},1}}$ where $p_0 = H_i$ and $p_k = H_j$ so $D_{i,j}$ is the sum of the genetic distances between sequences from consecutive patients in the transmission chain between H_i and H_j.

Chain Error model

$$\begin{aligned}\pi(X,\Psi|T,\rho,Z) = z^{\mathrm{TP}(X)}(1-z)^{\mathrm{FN}(X,T)} \\ \times \prod_{j=2}^{n_{seqs}}\prod_{i=1}^{j}\Bigg[\mathbb{1}_{\mathrm{trans}(i,j)=1}\frac{\theta^{\Psi_{i,j}}\exp(-\theta)}{\Psi_{i,j}!} \\ + \mathbb{1}_{\mathrm{trans}(i,j)>1}\frac{(k\gamma)^{|D_{i,j}-\Psi_{i,j}|}}{|D_{i,j}-\Psi_{i,j}|!\left(\sum_{l=0}^{D_{i,j}}\frac{(k\gamma)^l}{l!}\right)}\left(\frac{1}{2}\right)^{\mathbb{1}_{\{\Psi_{i,j}\neq D_{i,j}\}}}\mathbb{1}_{\{\Psi_{i,j}\leq 2D_{i,j}\}} \\ + \mathbb{1}_{\mathrm{trans}(i,j)=\infty}\frac{\theta_{gl}^{\Psi_{i,j}}\exp(-\theta_{gl})}{\Psi_{i,j}!} + \mathbb{1}_{\mathrm{trans}(i,j)=0}\frac{\theta_i^{\Psi_{i,j}}\exp(-\theta_i)}{\Psi_{i,j}!}\Bigg]\end{aligned}$$

where $\mathrm{TP}(X)$ is the number of true positive screening results given the swab results X, and $\mathrm{FN}(X,T)$ is the number of false negative screening results, given the swab results X and augmented data T.

Chain Poisson model

$$\begin{aligned}\pi(X,\Psi|T,\rho,Z) = z^{\mathrm{TP}(X)}(1-z)^{\mathrm{FN}(X,T)} \\ \times \prod_{j=2}^{n_{seqs}}\prod_{i=1}^{j}\Bigg[\mathbb{1}_{\mathrm{trans}(i,j)=1}\frac{\theta^{\Psi_{i,j}}\exp(-\theta)}{\Psi_{i,j}!} + \mathbb{1}_{\mathrm{trans}(i,j)=0}\frac{\theta_i^{\Psi_{i,j}}\exp(-\theta_i)}{\Psi_{i,j}!} \\ + \mathbb{1}_{\mathrm{trans}(i,j)>1}\frac{D_{i,j}^{\Psi_{i,j}}\exp(-D_{i,j})}{\Psi_{i,j}!} + \mathbb{1}_{\mathrm{trans}(i,j)=\infty}\frac{\theta_{gl}^{\Psi_{i,j}}\exp(-\theta_{gl})}{\Psi_{i,j}!}\Bigg]\end{aligned}$$

where $\mathrm{TP}(X)$ is the number of true positive screening results given the swab results X, and $\mathrm{FN}(X,T)$ is the number of false negative screening results, given the swab results X and augmented data T.

Time Dependent Distances model

$$\begin{aligned}\pi(X,\Psi|T,\rho,Z) = z^{\mathrm{TP}(X)}(1-z)^{\mathrm{FN}(X,T)} \\ \times \prod_{j=2}^{n_{seqs}}\prod_{i=1}^{j}\Bigg[\mathbb{1}_{\mathrm{trans}(i,j)=1}\frac{(t_{ij}\theta)^{\Psi_{i,j}}\exp(-(t_{ij}\theta))}{\Psi_{i,j}!} + \mathbb{1}_{\mathrm{trans}(i,j)=0}\frac{\theta_i^{\Psi_{i,j}}\exp(-\theta_i)}{\Psi_{i,j}!} \\ + \mathbb{1}_{\mathrm{trans}(i,j)>1}\frac{D_{i,j}^{\Psi_{i,j}}\exp(-D_{i,j})}{\Psi_{i,j}!} + \mathbb{1}_{\mathrm{trans}(i,j)=\infty}\frac{\theta_{gl}^{\Psi_{i,j}}\exp(-\theta_{gl})}{\Psi_{i,j}!}\Bigg]\end{aligned}$$

where t_{ij} is a measure of the time difference between the two patients which in the case where we have swab times is $t_{i,j} = |t_i^s - t_j^s|$, $\mathrm{TP}(X)$ is the number of true positive screening results given the swab results X, and $\mathrm{FN}(X,T)$ is the number of false negative screening results, given the swab results X and augmented data T.

2.8.1.2 Epidemiological part of the model likelihood

The likelihood of the unobserved data given the parameters is $\pi(T|\rho, Z)$, where Z, the vector of observed dynamics, consists of admission, discharge and screening times. The number of patients who are admitted to the ward during the course of the study is given by n. The set of all patients in the study is given by P and the set of patients who were colonised whilst on the ward is given by $P^w = \{i \in P : \phi_i = 0, t_i^c \neq \infty\}$. The epidemiological likelihood is given by

$$\begin{aligned}\pi(T|\rho, Z) = {} & p^{\sum_i \phi_i}(1-p)^{n-\sum_i \phi_i} \\ & \times \prod_{i=1}^{n}\left[\mathbb{1}_{t_i^c = t_i^a} + \mathbb{1}_{t_i^c \neq t_i^a}\exp\left(-\sum_{t=t_i^a}^{min(t_i^c-1, t_i^d)}\beta C(t)\right)\right] \\ & \times \prod_{j\in P^w}\frac{\left(1-\exp\left(-\beta C(t_j^c)\right)\right)}{C(t_j^c)}\mathbb{1}_{\{s_j \in C(t)\}}\end{aligned}$$

where ϕ_i is the admission state of the ith patient, so $\phi_i = 1$ if the patient was colonised before admission and $\phi_i = 0$ if they were susceptible on admission, s_j is the source of patient j's colonisation, and $C(t)$ is the number of colonised individuals present on the ward on day t. The admission, colonisation and discharge times of patient i are given by t_i^a, t_i^c, and t_i^d respectively.

2.9 Discussion

In this chapter we have investigated the validity of two assumptions that are often made when modelling genetic distances between sequenced isolates. The first common assumption is that each nucleotide in a sequence can only mutate once between isolate samples. Thus there is no probability of a nucleotide being observed as the same base in both sequences but actually having mutated to another base and then back to the original base. We found that this assumption was credible in a setting where the number of nucleotides is large and the mutation rate is low.

The second common assumption that we looked at was the assumption that the pairwise genetic distances between patients in a transmission chain are independent of each other. We found that this assumption was not realistic and therefore proposed to develop a model which relaxed this assumption. We introduced three new models for the genetic distances between sampled isolates. Each of these models includes dependence upon the underlying genetic distances from the chain in the distribution

for the genetic distance between sequences from two patients in a transmission chain who are separated by more than one transmission event.

We introduced a discrete-time stochastic model for an epidemic which includes any one of the three models for genetic distances. In chapter 3 we will discuss methods for assessing the goodness-of-fit of these models to data, and in chapters 4 and 5 we will fit these models to datasets from outbreaks of two very different pathogens, MRSA and avian influenza.

Chapter 3

Model assessment for models used to analyse whole-genome sequence data

3.1 Motivation

In chapter 1 we discussed some of the many different models which have been proposed which can make use of the new abundance of genetic data available in order to analyse epidemic outbreaks. Many of these models are tailored to fit the dynamics of a specific disease (eg. [70–72]), although some have wider application (eg. [17–19]). However, the goodness-of-fit of these models is difficult to test, and where there is more than one model for a disease, it is often difficult to say which model is a better fit for the data. Epidemic model assessment is of great importance because predictions about future outbreaks and control strategies for them can be sensitive to the chosen model. The field of model assessment for these types of models is underdeveloped, and although there has recently been a focus in the literature on model criticism for stochastic epidemic models [73], methods which consider the genetic data specifically have not been developed.

In chapter 2 we proposed a model for the spread of disease which can take one of three separate models for the genetic data from the epidemic. When we apply this model to data we must be able to assess the fit of the whole model for the spread of the pathogen, and also to distinguish which of the three genetic models is the best fit for the specific data that we are using. In this chapter we propose novel methods for

doing this. First, in section 3.2, we explore established methods for assessing the fit of an epidemic model to the epidemiological data from an outbreak of disease. Then, in section 3.3, we propose to extend these methods in order to assess the fit of an epidemic model to whole-genome sequence data.

3.2 Model assessment for epidemic models

In this section we will discuss the established method of posterior predictive checking for assessing the goodness-of-fit of models to epidemiological data. We use the term epidemiological data to refer to data collected from an outbreak which gives us information about the transmission of the disease. For example, these could be the results of tests which are carried out on the population to assess who has the disease, or data concerned with the dynamics of the population. In section 3.3 we will expand the idea of posterior prediction to the model for the genetic data also.

3.2.1 Posterior predictive checks

In a Bayesian setting, posterior predictive checks are often carried out in order to assess the goodness-of-fit of an epidemic model to a specific set of data [73, 74]. The posterior predictive distribution uses the posterior density of the model parameters, $\pi(\theta|x)$, which is obtained when a model has been fitted to data x in order to make predictions about hypothetical data, x^{new}, which might be observed in the future. The posterior predictive distribution is defined by

$$\pi(x^{new}|x) = \int \pi(x^{new}|\theta)\pi(\theta|x)\mathrm{d}\theta.$$

In addition, as proposed by Gelman et al. [21], we may define a set of auxiliary statistics, $A(x)$, which are to be matched when sampling the future data, so that $A(x^{new}) = A(x)$. In this case,

$$\pi(x^{new}|x, A(x)) = \int \pi(x^{new}|\theta, A(x))\pi(\theta|x)\mathrm{d}\theta.$$

In order to use this distribution to assess the goodness-of-fit of the model, often a summary statistic $S(X)$ is used to represent the dataset. In order to assess the fit of the model to the observed data, many hypothetical datasets may be drawn from the posterior predictive distribution to provide an approximation to the distribution of the summary statistic. Meng [22] shows that the observed value of the summary statistic, $S(X)$, may be compared to this approximate distribution in order to find a

posterior predictive p-value, defined by

$$p_S = \mathrm{P}(S(x) \geq S(x^{new})|x, \theta).$$

3.2.2 Summary statistics

In order to use posterior predictive checking to assess the goodness-of-fit of the models in chapter 2 we need to introduce some summary statistics which give a good representation of the epidemiological data which we are fitting our model to. Worby et al. [14] use the number of patients who are importations (the first swab after their arrival on the ward is positive) and the number of patients who are acquisitions (they have at least one negative swab before having a positive swab) as their summary statistics for assessing the goodness-of-fit of their model to data from an outbreak of MRSA in a hospital setting. These summary statistics will also work for assessing our model. We will also look at the number of patients ever to have a positive swab, and the variation in the number of patients with a positive swab present on the ward over the timescale of the outbreak.

3.2.3 An example of epidemic model assessment using simulated data

In order to illustrate how we can use posterior predictive checking to assess the goodness-of-fit of the models introduced in chapter 2 we simulated a dataset from our Chain Error model (see 2.6.1.1) for a population of 100 patients on a hospital ward over a period of 150 days. The values of the parameters used to simulate this dataset can be found in table 3.1 on page 50 under 'Sim 1'. The simulated dataset consisted of admission and discharge times for each of the patients, as well as results for swabs taken from each patient on the ward every other day for the length of their stay. We also simulated a set of genetic distances between sequences taken from patients with positive swabs according to the Chain Error model. In the next section we describe our simulation method.

3.2.3.1 Simulation method

In order to simulate an outbreak of a pathogen on a hospital ward we first specify the number of patients in the study, n, the length of the study, L, and the average length of stay for patients on the ward, A. The ward is assumed to be empty at $t = 0$ which is the first day of the study. We assume tests are taken every κ days from all patients who are presents on the ward on that day, so the set of test days t^t can be generated

	Total number of patients	Duration of study	Average length of patient stay	Test frequency	p	z	β	θ	θ_{gl}	γ
Sim 1	100	150	7	2	0.06	0.8	0.015	40	100	10
Sim 2	100	100	7	2	0.06	0.8	0.01	40	200	30
Sim 3	200	200	5	2	0.06	0.8	0.02	40	200	-
Sim 4	100	100	7	2	0.06	0.8	0.02	2	200	40

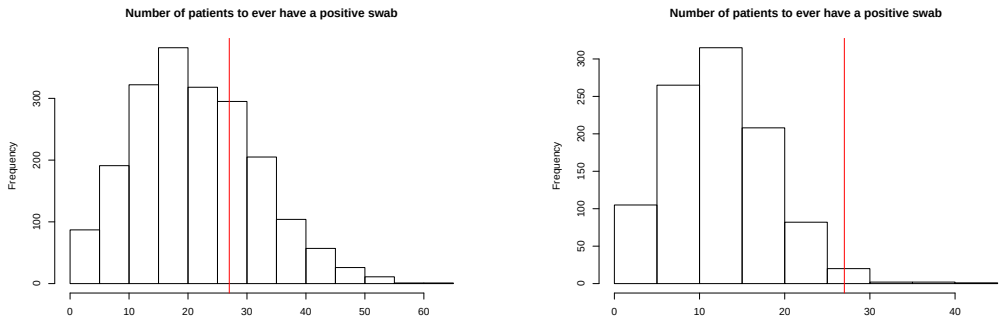
Table 3.1: Details for the simulations referred to in this chapter, including values of the model parameters.

independently of the patient stays from the test frequency parameter κ.

Each of the n patients is independently admitted colonised with probability p. For each we draw a date of admission to the ward uniformly at random from time 0 to time L and draw their length of stay from a Poisson distribution with parameter A.

Patients who are admitted to the ward in a susceptible state either remain susceptible for their whole stay, or become colonised through contact with another infectious patient. A susceptible patient, i, avoids colonisation on day t with probability $\mathrm{P}(\mathrm{avoid}(t)) = \exp(-\beta C(t))$, where $C(t)$ is the number of colonised patients present on the ward on day t. If patient i does not avoid colonisation on day t then they acquire the pathogen and $t_i^c = t$. A source of colonisation is drawn for this patient's colonisation uniformly at random from the $C(t_i^c)$ patients available to colonise them. $C(t)$ consists of the number of importation patients who have arrived on or before day t and are discharged after day t, plus the number of patients who acquire the pathogen before day t and are discharged after day t.

For each patient i who is colonised we generate a test result for each of the test days, t^t, that patient i was present on the ward for. The test sensitivity is z and the test specificity is 1, so when positive patients are tested the result of their test is positive with probability z, and negative with probability $1 - z$, independently of all other tests and when negative patients are tested their tests are always negative. Therefore we only need to simulate test results for those patients who are colonised during their stay on the ward. We assume that each positive swab result a patient receives leads to a genetic sequence being observed on that test day. We also assume that a patient i who is colonised but never receives a positive test result has an unobserved sequence on their day of colonisation t_i^c which has a genetic distance to each sequence (observed or unobserved) on day t_i^c or earlier. We assume each sequence taken on day t has a genetic distance to each sequence (observed or unobserved) on day t or earlier.



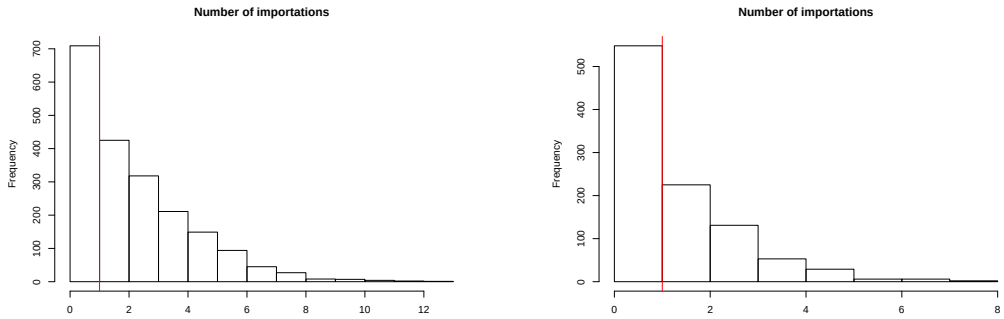
(a) (b)

Figure 3.1: Posterior predictive distributions for the number of patients ever to have a positive swab under (a) the true Chain Error model and (b) the wrong Chain Poisson model. The observed value from the original simulation is marked in red.

We draw these genetic distances from the distributions specified by the model, according to the relative positions on the transmission tree of the patients who have sequences i and j. For sequences from patients who are in distinct transmission chains, all models draw the genetic distance from a Poisson(θ_{gl}) distribution. For sequences from patients who share a direct transmission event, the Chain Error model and Chain Poisson model draw the genetic distance from a Poisson(θ) distribution and the Time Dependent Distances model draws the genetic distance from a Poisson$(t_{i,j}\theta)$ distribution. For sequences from patients who are in the same transmission chain but are separated by more than one transmission event the Chain Poisson model and the Time Dependent Distances model draw the genetic distance from a Poisson$(D_{i,j})$ distribution where $D_{i,j}$ is the sum of the underlying distances in the transmission chain, so $D_{i,j} = \sum_{r=0}^{k-1} \Psi_{Q_{p_r}, Q_{p_{r+1}}}$ where $p_0 = H_i$, $p_k = H_j$. The Chain Error model draws this distance by adding or subtracting from $D_{i,j}$, with probability 0.5, an error term drawn from a Poisson distribution with parameter $k\gamma$ which is truncated at the value $D_{i,j}$. The genetic distance between two sequences taken from the same patient is drawn from a Poisson distribution with parameter θ_i under each of the three models.

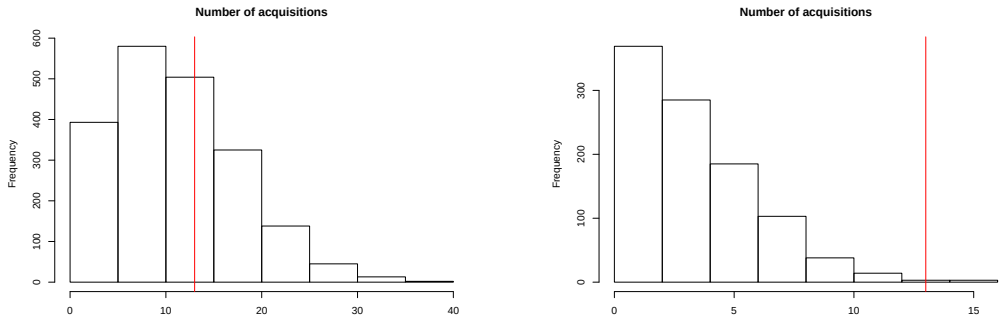
3.2.3.2 Using the simulated dataset for posterior prediction

To this dataset we fitted both the Chain Error model and the Chain Poisson model (see 2.6.1.2) with 100,000 iterations of our MCMC algorithm. A full description of the MCMC algorithm can be found in section 4.5. The prior distribution used for both parameter p and parameter z was U$(0,1)$, which is a Uniform distribution with parameters $a = 0$ and $b = 1$ which has probability density function $f(x) = \frac{1}{b-a}$ for



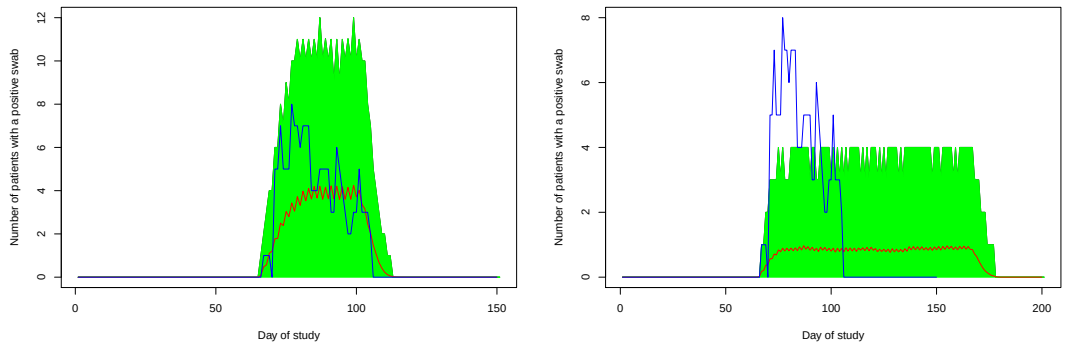
(a) (b)

Figure 3.2: Posterior predictive distributions for the number of patients whose first swab was positive under (a) the true Chain Error model and (b) the wrong Chain Poisson model. The observed value from the original simulation is marked in red.



(a) (b)

Figure 3.3: Posterior predictive distributions for the number of patients who had a positive swab after having had one or more negative swabs under (a) the true Chain Error model and (b) the wrong Chain Poisson model. The observed value from the original simulation is marked in red.



(a) (b)

Figure 3.4: Figures which show, in green, the 95% highest density region for the posterior predictive distribution of the number of patients with a positive swab present on the ward over the time of the study under (a) the true Chain Error model and (b) the wrong Chain Poisson model. The mean of the distribution is shown in red, and the observed data from the original simulation is shown in blue.

$a \leq x \leq b$. For the genetics parameters, θ, θ_{gl} and θ_i, a $\Gamma(1, 10^{-6})$ distribution was used as the prior distribution, which is a Gamma distribution, with parameters $\nu = 1$ and $\lambda = 10^{-6}$, which has probability density function $f(x) \propto x^{\nu-1}\exp(-\lambda x)$ for $x > 0$. We used improper uniform distributions on the set of positive real numbers as prior distributions for parameters β and γ. We initialised the infection times by giving each patient who had received a positive swab a colonisation time of the day before their first positive swab. If a patient's first positive swab was on their day of admission they were assigned as an importation. For patients who were not importations and had a positive test we drew a source uniformly at random from the set of other colonised patients on the ward on the day of colonisation. If no source was available we reassigned that patient as an importation. We initialised the missing sequences by drawing a genetic distance between each patient who had a positive swab but no sequence and each other patient sequence from a Poisson distribution with mean 30. We used initial values for the parameters that were based on the results of Worby [69]. We ran the MCMC algorithm a number of times with different initial values to check that it converged to the same mode, and we examined the traceplots to check that it had converged.

For our posterior predictive checking, 1000 datasets were simulated using values of the parameters for the model drawn from the posterior densities given by the MCMC algorithm output. For each of these 1000 datasets we recorded the values of the summary statistics to be used for model assessment: the number of patients ever to have

a positive swab, the number of patients whose first swab was positive (importations), the number of patients who had a positive swab after one or more negative swabs (acquisitions), and the number of patients present on the ward each day with a positive swab on that day or earlier. The first three of these summary statistics are single values, so from the 1000 simulations we can approximate the distribution of the statistic and find where the 'observed' value from our original simulation fits in it. As the data were simulated from the Chain Error model, we expect that when we assess the goodness-of-fit of this model we should see that it fits well, whereas we do not expect to necessarily find that the Chain Poisson model is a good fit for these data.

Figure 3.1 shows the posterior predictive distribution for the number of patients who had a positive swab under each of the two models that we used. It is clear that the observed value from the original simulation lies within the 95% highest density region (HDR) under the true Chain Error model, but it falls outside that HDR under the Chain Poisson model. This shows that there is evidence that the Chain Poisson model is not a good fit for these data, as we expected. Similarly, figure 3.3 shows that the observed number of patients who are 'acquisitions' in the original dataset falls within the 95% HDR of the posterior predictive distribution produced under the Chain Error model, but it falls outside of that HDR under the Chain Poisson model, giving more evidence against the use of this model for these data. The number of patients who are 'importations' falls within the 95% HDR for both models, so this does not provide us with evidence against either model. Figure 3.4, however, gives us more evidence that the Chain Poisson model is not a good fit for these data. The blue line, which represents the number of patients present on the ward with a positive swab each day in the original simulation data, stays within the green area, which is the 95% HDR of the posterior predictive distribution, for the true Chain Error model, but it departs significantly from this area for the Chain Poisson model.

The use of a simulated dataset has illustrated how posterior predictive checking of summary statistics to do with the transmission of the epidemic can help to assess the goodness-of-fit of a particular model to a specific dataset. In this case posterior prediction showed the expected result that there was no evidence of lack of fit for the Chain Error model, but the posterior predictive distributions of a number of the summary statistics indicated that the Chain Poisson model was not a good fit.

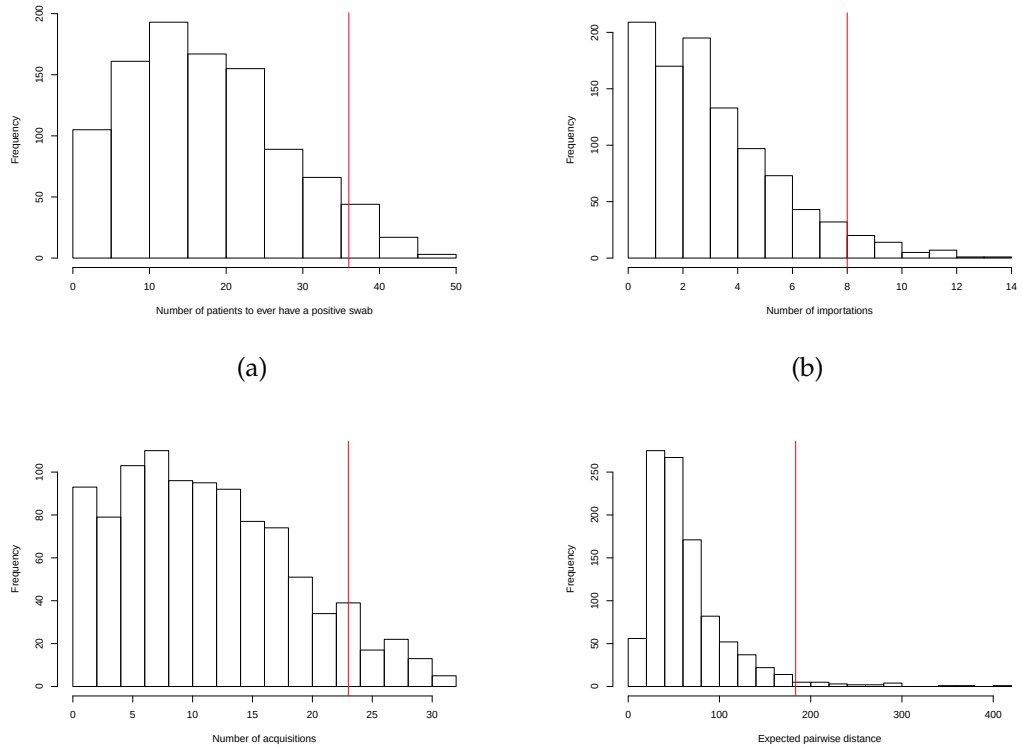
3.3 Model assessment for epidemic models which model genetic data

In section 3.2 we described methods for assessing the goodness-of-fit of our epidemic model, but these methods did not assess the fit of the model to the genetic data. This is an area of model assessment which has not yet been developed. Although there has recently been a focus on model assessment for stochastic epidemic models [73], the only method which explicitly assesses the fit of a model to genetic data was proposed by Worby et al. [14]. Here we introduce this method and then develop our own methods which build upon the posterior prediction methods used in section 3.2.

3.3.1 Posterior prediction using a summary statistic for the genetic data

Worby et al. [14] use posterior prediction to assess the goodness-of-fit of their model to the genetic data simply by using a single summary statistic for the genetic distance matrix. They are able to compare the observed value of this statistic to its distribution as approximated by simulating multiple datasets using parameter values from the posterior densities. The summary statistic which they use is the average genetic distance between any two sequences from patient isolates. This is a sensible starting point for genetic model assessment, but it does have limitations, as the expected genetic distance between two sequences is highly dependent upon the unknown transmission tree, and this method ignores the tree inferred by the model. Simulating multiple transmission trees from the same parameter values can give a set of very different tree structures, so the set of expected pairwise distances will also have a large range, as it is affected by how many chains of transmission there are, how long each of these chains are, and how many importations of the disease occur. Similarly, when we simulate epidemics from the posterior densities of the parameters, if we only record the expected genetic distance between each pair of sequences then we lose much information about the tree structure that has influenced that number.

Figure 3.5 illustrates one situation in which simply using the mean of genetic distances for a posterior predictive check can give misleading information about the fit of the model. Here we simulated a dataset from the Chain Error model, and then fit this true model to the data. The values of the parameters used to simulate this dataset can be found in table 3.1 on page 50 under 'Sim 2'. We fit the Chain Error model to these data using 100,000 iterations of our MCMC algorithm with prior distributions of $\mathrm{U}(0,1)$ for parameters p and z and $\Gamma(1,10^{-6})$ for the genetics parameters θ, θ_{gl} and



(a) (b)

(c) (d)

Figure 3.5: Posterior predictive distributions obtained by fitting the Poisson Error model to data simulated from the same model. Figure (a) shows the distribution of the number of patients to ever have a positive swab over the course of the epidemic, figure (b) shows the distribution of the number of patients whose first swab is positive, and figure (c) shows the distribution of the number of patients who had a negative swab before a positive swab. Figure (d) shows the distribution of the mean of the genetic distance matrix. In each case the value from the original simulation is marked in red.

θ_i. We used improper uniform distributions on the set of positive real numbers as prior distributions for parameters β and γ. The chain was initialised as described in section 3.2.3.2, and convergence was checked using the traceplots of the output of the parameter values.

All of the epidemic summary statistics in figure 3.5 (the number of positive swabs, the number of importations and acquisitions) show that the 'observed' values from the original simulation are well within the 95% highest density region of the posterior predictive distribution, and even within the 90% HDR, but the mean genetic distance is outside the 95% HDR of its posterior predictive distribution. Since the model that we fit to the data was the same model that the dataset was simulated from we do not expect it to be a poor fit. We see that the number of importations and number of acquisitions in our original simulation are above the mean of their posterior predictive distributions. This suggests that in the dataset which we simulated we have an above average number of chains (meaning more between chain distances), and possibly above average chain lengths (also meaning more large genetic distances), which would both increase the mean genetic distance. It is clear that the mean genetic distance is being influenced by the range of different tree structures that are simulated to produce the posterior predictive distribution. In the next sections we explore different ways of using posterior predictive checks to assess the goodness-of-fit of a model to genetic data.

3.3.2 Summary statistics of the genetic matrix for posterior predictive checks

We have seen that using the mean of the genetic distance matrix as a summary statistic in order to check the fit of a model using posterior prediction does not always accurately assess the goodness-of-fit. Therefore we investigated other summary statistics for the genetic distance matrix, which were:

- The median of the genetic distance matrix
- The sum of the genetic distance matrix
- The range of the genetic distance matrix
- The interquartile range of the genetic distance matrix

We again used simulated datasets, to which we fitted the true model, and a wrong model. We fit each model to these datasets using 100,000 iterations of our MCMC algorithm with prior distributions of $\mathrm{U}(0,1)$ for parameters p and z and $\Gamma(1,10^{-6})$ for

the genetics parameters θ, θ_{gl} and θ_i. We used improper uniform distributions on the set of positive real numbers as prior distributions for parameters β and γ. Each chain was initialised as described in section 3.2.3.2, and convergence was checked using the traceplots of the output of the parameter values. We found that using the posterior predictive checks listed above provided evidence against the fit of the wrong model, but also, for some simulations, gave evidence against the fit of the true model. This is the same problem that we had in using the mean of the genetic distance matrix: the structures of the trees being simulated to approximate the posterior predictive distribution are not always comparable to the structure of the true tree and the summary statistics do not capture this. The genetic distances are conditional upon the transmission tree and even simulating the correct model with the true parameter values may only rarely lead to a compatible tree to the data.

In an attempt to resolve this problem with the posterior predictive checks we proposed somehow to constrain the simulations from the posterior densities to be more similar to the original dataset. One way in which we tried to do this was to fix the times of admission, discharge and testing of the patients to those in the original dataset, since these do not form part of the model framework. Previously, the admission time for a patient had been drawn uniformly at random from time 0 to time L, and the length of stay drawn from a Poisson distribution with parameter A. The times of the test were determined by the test frequency parameter. Another way in which we tried to use simulations from the posterior densities that were more comparable to the original data was to keep only those simulations which had the same number of patients with a positive swab during the outbreak. Fixing the times for the simulations without fixing the number of patients with positive swabs did not improve the goodness-of-fit assessments, suggesting that these times do not hugely impact the tree structure. Fixing both the times and the number of patients with positive swabs meant that the process of posterior predictive checking took much longer, and although we did see improvement there were still some cases in which the summary statistic checks suggested a lack of fit under the true model.

We concluded that using summary statistics of the genetic distance matrix for posterior predictive checks was not always an accurate way of assessing the goodness-of-fit of a model to genetic data because such summary statistics can not capture information about the underlying transmission tree structure. In section 3.3.3 we discuss methods for including information from the posterior densities about the tree structure when performing model assessment.

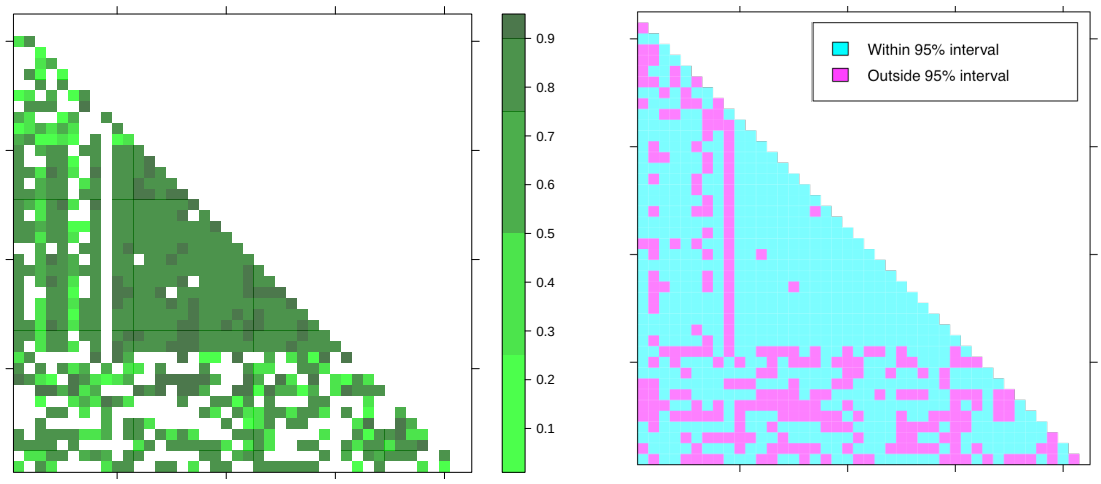
3.3.3 Posterior prediction for the whole genetic matrix

As discussed in section 3.3.2, using a summary statistic for posterior predictive checks of the genetic model fit does not use all the available information about the tree structure that the model has inferred for the data. We are interested in checking how well the model can estimate a transmission tree for the data, as a key aim of using a model such as ours is to produce a transmission tree which shows the likelihood of specific routes of transmission and infers who infected whom. This essentially means that we require a method of performing model assessment for matrices, as the genetic data are in the form of a matrix, so we need a way to assess the posterior distribution of matrices without resorting to summary statistics which result in a loss of information.

In order to achieve this we propose a method which uses the posterior density of the transmission tree, T, for posterior prediction as well as the posterior densities of the parameters, ρ. Therefore, instead of simulating the entire epidemic for each draw from the posteriors, we are simply able to simulate a genetic distance matrix from each draw since the posterior density of the transmission tree defines the relationship between each pair of individuals. Thus we can draw each $\widetilde{\Psi}_{i,j}$ from the distribution specified by the model. As a result we can approximate the posterior predictive distribution for each genetic distance separately. The step-by-step process of this model assessment method for a model which has been fitted using an MCMC algorithm is outlined here for clarity:

1. For each of the required k simulations, draw one of the m iterations of the MCMC algorithm output uniformly at random. Call this iteration i.
2. From this iteration i, record the posterior values of the genetic parameters, ρ^i, and the set of infection times, $t^{c\,i}$, and sources of infection, s^i, for the infected population.
3. Simulate a genetic distance matrix, $\widetilde{\Psi}^i$, from the genetic model using the sampled structure of the transmission tree and values of the parameters. Record this genetic distance matrix for each simulation.
4. Once this process has been repeated for the required number of simulations we have a set of k genetic distance matrices, $\widetilde{\mathbf{\Psi}}$, from which we can estimate the posterior predictive distribution for each of the genetic distances between pairs of sequences.

The simulated genetic distance matrices may differ slightly in dimension, as the MCMC



(a) (b)

Figure 3.6: Two different ways of plotting the results of our posterior predictive checks of the fit of a model to a genetic distance matrix. The layout of the matrices corresponds to the layout of the original genetic distance matrices, so the x-axis, left to right, runs from sequence 1 to sequence n_{seqs}, whereas the y-axis, from bottom to top, runs backwards from sequence n_{seqs} to sequence 1. Figure (a) fills the cells with a colour gradient which represents which level of highest density region of the posterior predictive distribution the observed value falls into. So the darkest green cells show that the distance only falls within the 90% highest density region, and the distances corresponding to the lightest green cells fall within the 25% highest density region. The distances corresponding to white cells fall outside the 90% highest density region. Figure (b) is a binary matrix which simply shows whether the observed value of each genetic distance was within the 95% highest density region of the posterior predictive matrix (blue cells) or not (pink cells). This plot is clearly easier to read.

algorithm can add and delete patients who did not have a positive swab from the tree. However, we will discard any added patients and concentrate on the 'core' sequences from patients who do have a positive swab in the data, and therefore appear in every genetic distance matrix. We originally tried plotting the posterior predictive distribution for each pairwise distance from the simulated values, and marking on the observed value, as we did for the posterior predictive checks in section 3.2. However, for genetic distance matrices of more than three or four sequences this becomes an unwieldy tool which is difficult to interpret. Instead, we propose to plot a matrix with a cell for each genetic distance which is coloured according to where the observed distance falls in the posterior predictive distribution. Two ways of doing this are shown in figure 3.6. Figure 3.6a shows a matrix with the cell for each pairwise genetic distance coloured on a gradient which indicates whether the observed distance falls outside of the 90% highest density region of the posterior predictive distribution (white), or falls within the 90%, 75%, 50% or 25% HDR. Figure 3.6b shows a matrix with the cell for each pairwise genetic distance coloured according to whether the observed distance falls within the 95% HDR of the posterior predictive distribution (blue) or not (pink). This version of the matrix plot is much clearer and visually presents the proportion of observed distances that fall within the 95% posterior predictive HDR. It also has the advantage that we may consider it as a binary matrix and assign each cell a 0 if it is pink and a 1 if it is blue, and therefore we can also give a percentage of these 'well-fitted' distances, which will give us an idea of the strength of the genetic model fit. We will call this percentage a *posterior predictive matrix score*.

3.3.3.1 Examples of genetic model assessment using simulated data

In order to demonstrate our novel method for assessing the goodness-of-fit of a genetic model to data we used datasets which were simulated from the Chain Error model (see 2.6.1.1) and the Chain Poisson model (see 2.6.1.2), including the dataset used in section 3.3.1. The details of the simulated datasets which feature in the following figures can be found in table 3.1 on page 50 under 'Sim 2', 'Sim 3' and 'Sim 4'. We then fitted three different models to these datasets in order to compare the fit. First, we fitted the true model, either the Chain Poisson model or the Chain Error model, to the data and then we fitted the other model to the same data. Finally, we fitted the Chain Poisson model again but with Geometric distributions in place of the Poisson distributions, in order to have an example of assessing the goodness-of-fit when the fitted model is vastly different from the true model. This model uses the

following distributions for the genetic distances between sequences:

$$\mathrm{P}(\Psi_{i,j} = x) = \begin{cases} (1-\mu_{gl})^x \mu_{gl} & \text{if } k = \infty \\ (1-\mu)^x \mu & \text{if } k = 1 \end{cases} \tag{3.3.1}$$

and

$$\mathrm{P}(\Psi_{i,j} = x | D_{i,j}) = (D_{i,j}^x / x!) \exp(-D_{i,j}) \quad \text{if } k > 1. \tag{3.3.2}$$

We fit each model to each dataset using 100,000 iterations of our MCMC algorithm with prior distributions of $\mathrm{U}(0,1)$ for parameters p and z and $\Gamma(1, 10^{-6})$ for the genetics parameters θ, θ_{gl} and θ_i. We used improper uniform distributions on the set of positive real numbers as prior distributions for parameters β and γ. Each time the chain was initialised as described in section 3.2.3.2, and convergence was checked using the traceplots of the output of the parameter values.

Using our method for assessing the goodness-of-fit of the genetic models we were able to identify the true model as the best fit for each simulation that we tried. The lack of fit identified for the wrong models varied depending on how different the models were: for example, the Geometric version of the model was a much worse fit for data simulated under the Chain Poisson model than the Chain Error model was. The composition of the genetic distance matrix in the dataset, which is affected by the values of the parameters, also had some impact upon the difference in the goodness-of-fit of the three models. Figures 3.7, 3.8 and 3.9 present some typical examples from our set of simulations.

Figure 3.7 shows the results of our method for assessing the fit of the genetic models for the simulated dataset (Sim 2) discussed in section 3.3.1 in a series of binary matrices. Despite the posterior predictive check of the mean genetic distance for the true model giving an extreme posterior predictive p-value, as shown in figure 3.5d, these model assessment matrices clearly show that the true model is the best fit for the data, with 88% of the 'observed' genetic distances falling within the 95% highest density regions from the posterior predictive distributions. The Chain Poisson model also appears to fit fairly well, with 78% falling within the 95% HDRs. This is not surprising, given that these models share common distributions for modelling genetic distances between sequences from patients who are separated by one transmission event, and for modelling genetic distances between sequences from patients who are in different transmission chains. The difference in these two models is in the way in which genetic distances between sequences from patients who are in the same trans-

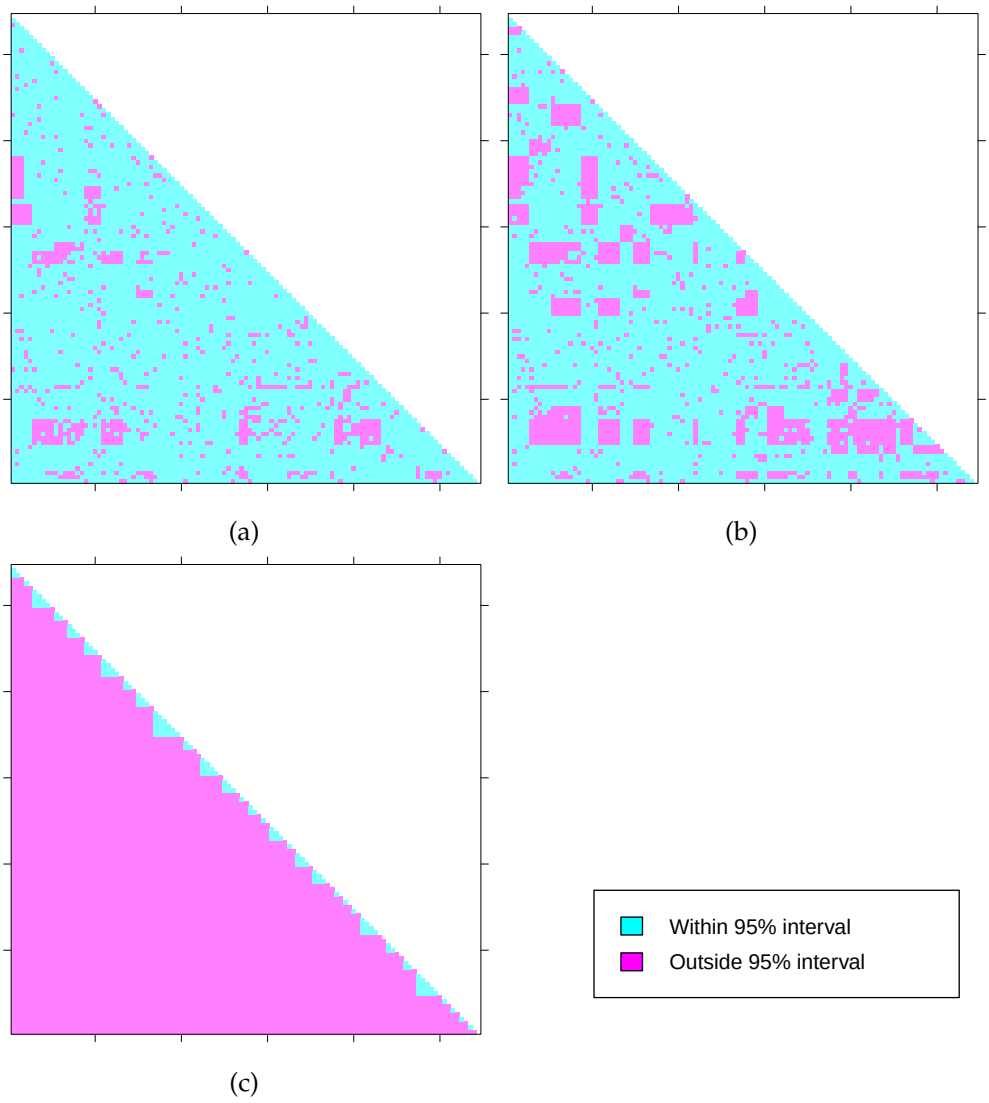


Figure 3.7: Sim 2: Binary matrices showing which of the 'observed' genetic distances fall within the 95% HDR of the posterior predictive distribution given by (a) the true Chain Error model, (b) the Chain Poisson model, and (c) the Geometric Chain Poisson model. The layout of the matrices corresponds to the layout of the original genetic distance matrices, so the x-axis, left to right, runs from sequence 1 to sequence n_{seqs}, whereas the y-axis, from bottom to top, runs backwards from sequence n_{seqs} to sequence 1.

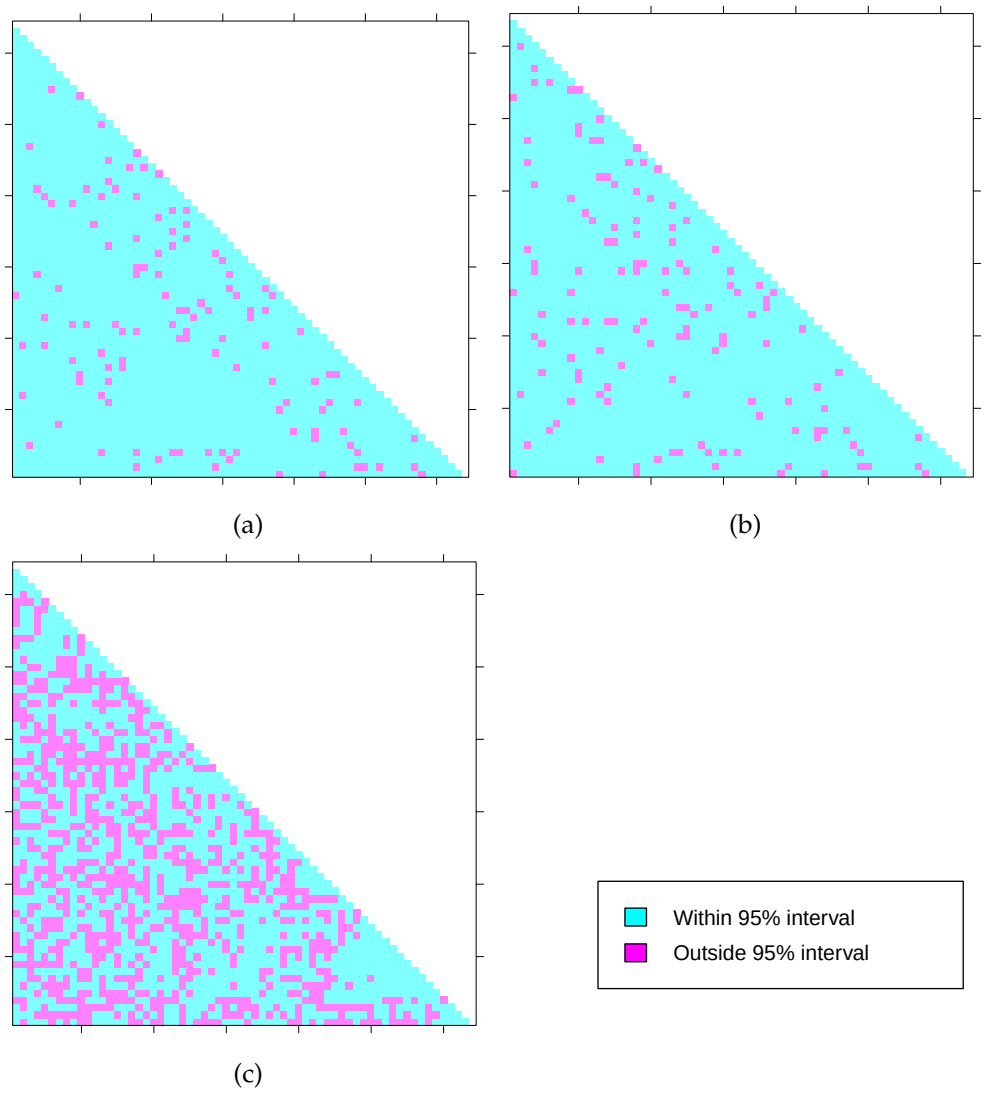


Figure 3.8: Sim 3: Binary matrices showing which of the 'observed' genetic distances fall within the 95% HDR of the posterior predictive distribution given by (a) the true Chain Poisson model, (b) the Chain Error model, and (c) the Geometric Chain Poisson model. The layout of the matrices corresponds to the layout of the original genetic distance matrices, so the x-axis, left to right, runs from sequence 1 to sequence n_{seqs}, whereas the y-axis, from bottom to top, runs backwards from sequence n_{seqs} to sequence 1.

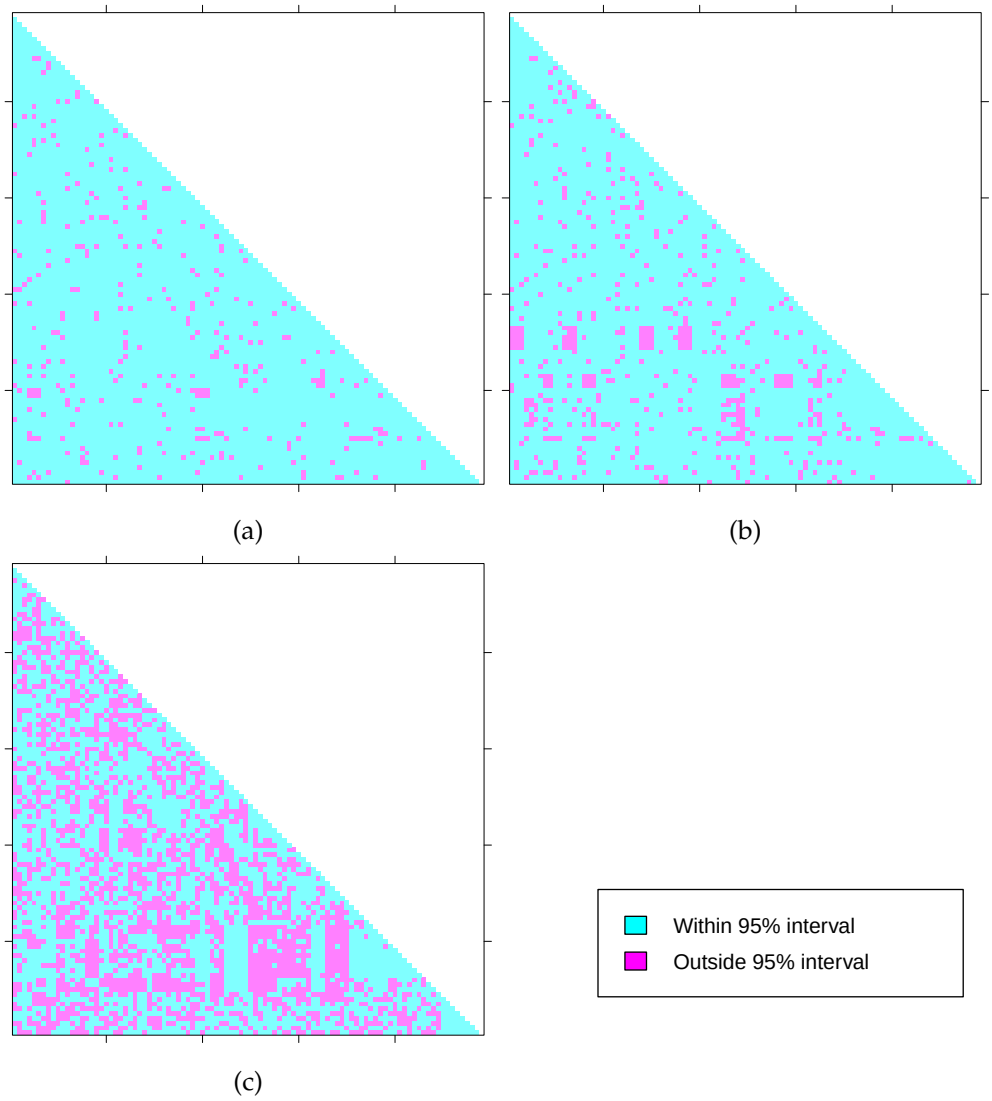


Figure 3.9: Sim 4: Binary matrices showing which of the 'observed' genetic distances fall within the 95% HDR of the posterior predictive distribution given by (a) the true Chain Error model, (b) the Chain Poisson model, and (c) the Geometric Chain Poisson model. The layout of the matrices corresponds to the layout of the original genetic distance matrices, so the x-axis, left to right, runs from sequence 1 to sequence n_{seqs}, whereas the y-axis, from bottom to top, runs backwards from sequence n_{seqs}

to sequence 1.

mission chain but are separated by more than one transmission event are modelled. Figure 3.7c shows that when we use a model which uses different distributions for all types of genetic distances, only 2.5% fall within the 95% HDRs from the posterior predictive distributions. This is much lower than for the true model and the similar model, which shows that the model is not a good fit for the data, as we expected.

Figure 3.8 (Sim 3) and 3.9 (Sim 4) give some other examples of using our method for assessing the goodness-of-fit of different models to the genetic data from a simulation. In figure 3.8 the true model is the Chain Poisson model and under this model 94% of the observed genetic distances fall within the 95% HDR of the posterior predictive distributions. Under the Chain Error model 93% of the observed distances fall within the 95% HDR, which shows that the Chain Error model fits almost as well as the true model in this case. As the Chain Poisson model is a simplified version of the Chain Error model this is not surprising. Under the Geometric Chain Poisson model, however, only 60% of the observed distances fall within the 95% HDR of the posterior predictive distributions. In this case the very different model has actually done reasonably well with the genetic distance matrix if we compare this 60% to the 2.5% for this model in figure 3.7. Of all our simulations, this was the one for which the Geometric model performed the best, but the true model and the Chain Error model are still clearly found to fit the data better.

In figure 3.9 the true model is the Chain Error model and under this model 94% of the observed pairwise genetic distances fall within the 95% HDR of the posterior predictive distribution. The similar Chain Poisson model was also found to fit reasonably well, with 89% of the observed pairwise genetic distances falling within the 95% HDR of the posterior predictive distributions. The very different model, the Geometric Chain Poisson model, was found to fit significantly less well, with only 57% of the observed pairwise genetic distances falling within the 95% HDR of the posterior predictive distributions. For all the simulated datasets to which we fitted different models we were able to correctly identify the true model using this method of posterior predictive matrices.

3.4 Discussion

In this chapter we have discussed a method for assessing the goodness-of-fit of models which are fitted to both epidemiological and genetic data. We have described, in section 3.2, how established methods for posterior predictive checking can be used

to assess the fit of the three models which were introduced in chapter 2 to epidemiological data from an epidemic. In section 3.3 we introduced our novel method for assessing the goodness-of-fit of our models, or similar models, to the genetic data from an epidemic. Our method results in a percentage, or *posterior predictive matrix score*, which describes the proportion of the genetic distance matrix that is 'well-fit' under the model used. It also produces a binary matrix as a visual representation of this percentage, which allows us to see whether there are specific areas of the genetic distance matrix which are poorly captured by the model. It is clearly important that the fit of the genetic part of a model for an epidemic is examined as well as the epidemiological part in order to assess the goodness-of-fit of the model as a whole. Our new tool allows this to be done for any model from which genetic distance matrices can be simulated, as long as the algorithm used to fit the model records the samples from the posterior density of the transmission tree at each iteration. It would be of interest to develop a tool like this one which could also assess the predictive ability of a genetic models.

Chapter 4

Analysis of an outbreak of methicillin-resistant *Staphylococcus aureus* in a hospital setting

4.1 Introduction

In this chapter we introduce a nosocomial dataset for an outbreak of methicillin-resistant *Staphylococcus aureus* in a hospital in Thailand. We outline how our new models for the genetic distances between patients, introduced in chapter 2, provide a full model for the outbreak. We set out the MCMC routine which is used to fit the model to the data, and analyse the results. A simulation study is carried out to assess the performance of the MCMC algorithm, and the methods for model assessment which were introduced in chapter 3 are used to assess the goodness-of-fit of the model to the data. A table of the notation used in this chapter can be found on pages 124-125.

4.2 Thai data

Here we introduce the dataset collected from an outbreak of methicillin-resistant *Staphylococcus aureus* (MRSA) in a hospital in Thailand in 2008. These data were collected by Tong et al. [75] who performed the study over a three-month period on two intensive care units in the same 1000-bed hospital in northeast Thailand. The dataset includes 83 MRSA genome sequences from 51 unique patients, which were aligned to a reference genome of the dominant lineage (ST 239 strain TW20) of MRSA in the

hospital. ST 239 is a global lineage of MRSA that has been discussed by Harris et al. [76]. In this Thai hospital dataset a total of 2591 nucleotides changed from the reference genome in at least one patient sequence. These data were collected by repeat screening for MRSA of patients on two intensive care units (ICUs), one surgical ICU and one paediatric ICU.

4.2.1 Overview of the dataset for each ward

The following table summarises the data that were observed for each ICU ward.

	ICU 1	ICU 2
Ward type	Surgery	Paediatric
Number of patients admitted	170	114
Number of unique patients	169	98
Number of patients with at least one positive swab	20	29
Total number of positive swabs collected	51	89
Total number of swabs sequenced	43	40

Figure 4.1 shows the number of transitions and transversions observed in the dataset. Interestingly, only 5 of the 2591 nucleotides underwent both a transition and a transversion in different sequences in the data. By this we mean nucleotides which had a base, A say, in the reference genome, and were observed to have the opposite base, G in this case, in one or more sequences in the data and to have one of the other bases, C or T here, in one or more other sequences. For a summary of the timelines of patients with positive swabs in each ward see figures 4.3 and 4.4. Each bar represents a patient's stay on the hospital ward. The red portion shows the time that the patient was in the ward without having had a positive swab result and the time at which the timeline turns to green is the date of the first positive swab result. The text at the ends of the bars lists the other patients who were on the ward and had had a positive swab on the day when the first positive swab was taken from the patient under consideration, and the genetic distance between these two patients.

Distances < 60 are marked in red on the timelines because Tong et al. [75] estimated that the maximum genetic distance between two sequences in the same cluster is 60, and we can see from figure 4.3 that the distances over 60 in ward 1 are all over 200, and likewise figure 4.4 shows that in ward 2 the smallest distances over 60 are 103 and 199, with all other distances over 200. Figures 4.2a and 4.2b show heatmaps of the genetic distance matrix for each ward, which again show the clear gap between

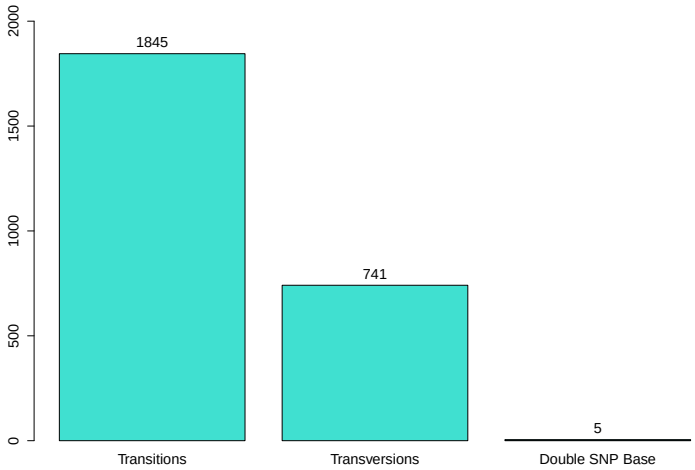
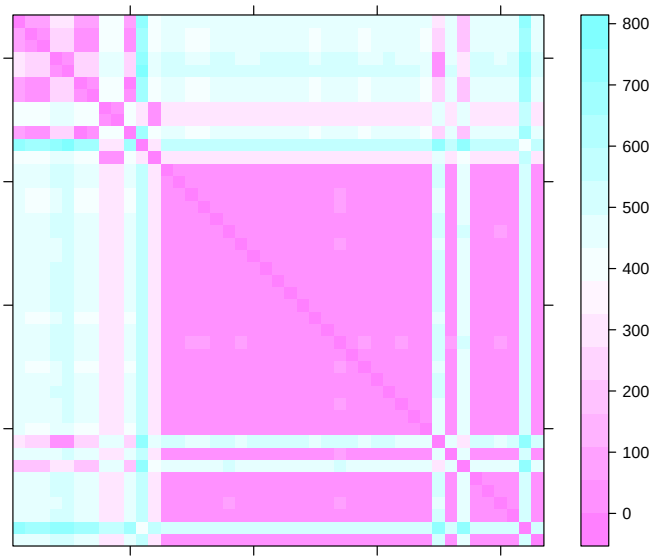
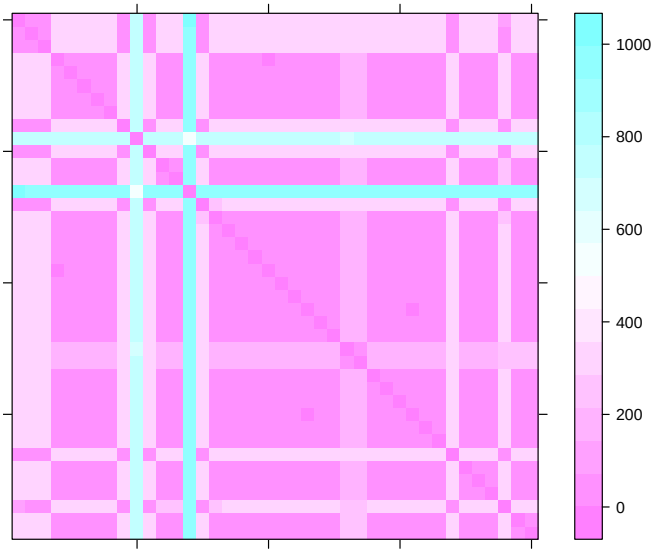


Figure 4.1: Changes from reference genome observed in the Thai data. Transitions occur when a nucleotide that was base A mutates to base G (or vice versa) or a nucleotide that was base C mutates to base T (or vice versa). All other possible mutations (A to C, C to G, A to T, T to G, or vice versa) are transversions. The 5 nucleotides that have a double SNP base are those which changed to two distinct bases which were both different from the reference genome in the sequences in the Thai data.

small distances of < 60 SNPs and larger distances. Patients within the same cluster are more likely to be in the same transmission chain, so, using the admission and discharge times and times of positive swabs we created two diagrams (figure 4.5a and figure 4.5b) which show what the transmission tree would look like if we assumed that two patients who were observed to be MRSA positive at the same time and whose sequences had a genetic distance of under 60 SNPs shared a direct transmission event. This crude way of creating transmission trees gives us a rough idea what the genetic distance data can tell us about the relationships between patients' sequences.



(a) ICU 1



(b) ICU2

Figure 4.2: Heatmaps, which are visual representations of the genetic distance matrices for each ward, with the size of the genetic distances between pairs of sequences represented by colours. It is clear that there is a large jump between the small distances of < 60 SNPs and the larger ones. The layout of the matrices corresponds to the layout of the original genetic distance matrices, so the x-axis, left to right, runs from sequence 1 to sequence n_{seqs}, whereas the y-axis, from bottom to top, runs backwards from sequence n_{seqs} to sequence 1.

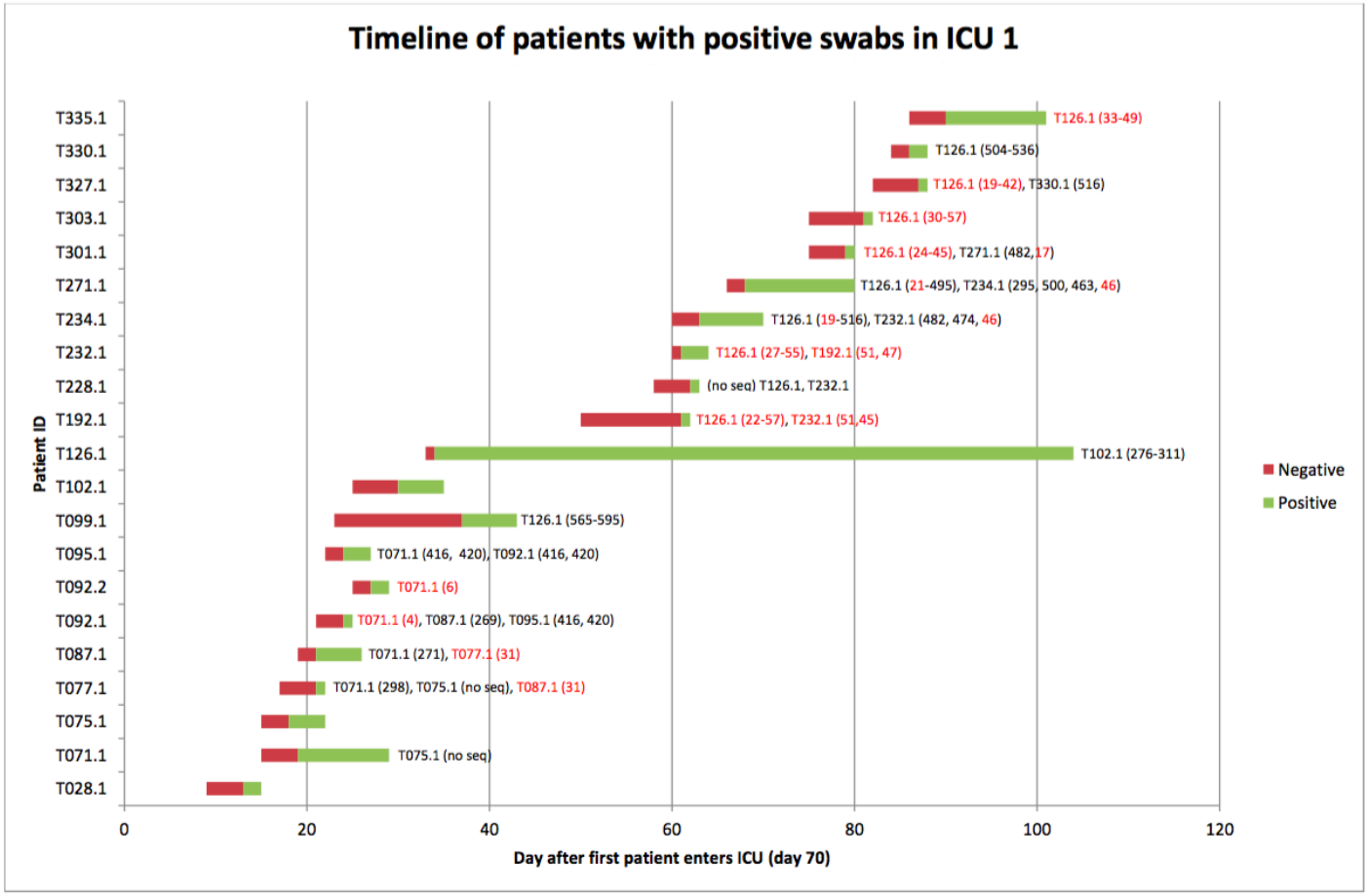


Figure 4.3: Timeline for patients observed positive in ward 1. Red shows the time that the patient was in the ward without having had a positive swab result and the time at which the timeline turns to green is the date of the first positive swab result. The labels are the other patients who are positive on the ward on the day when the first positive swab was taken from the patient under consideration, and the genetic distance between these two patients. Distances < 60 are in red.

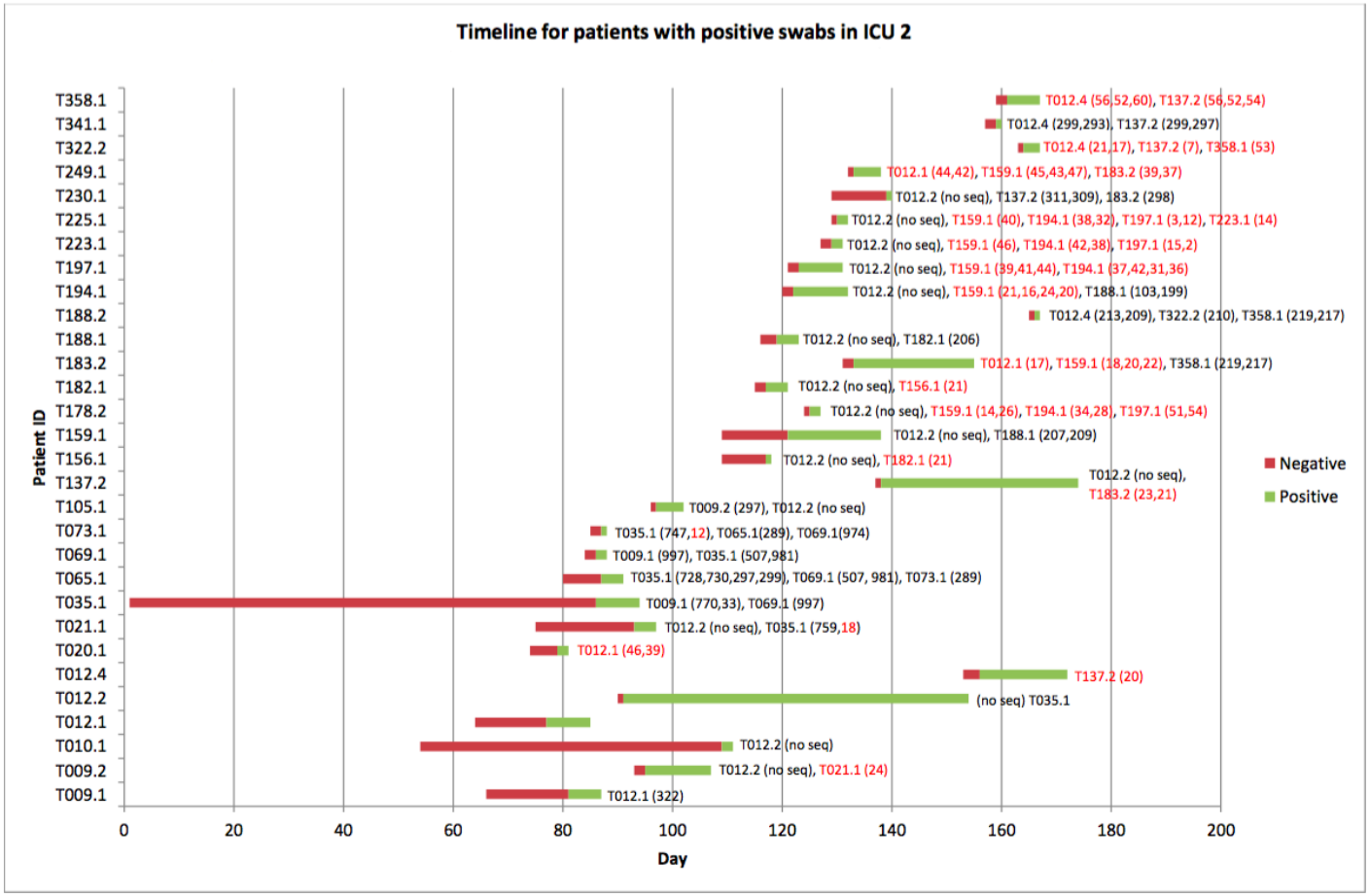
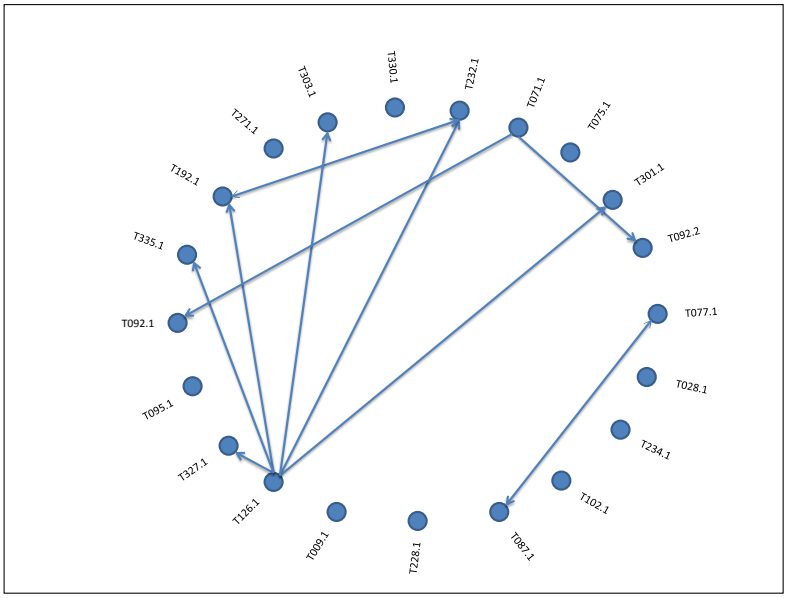
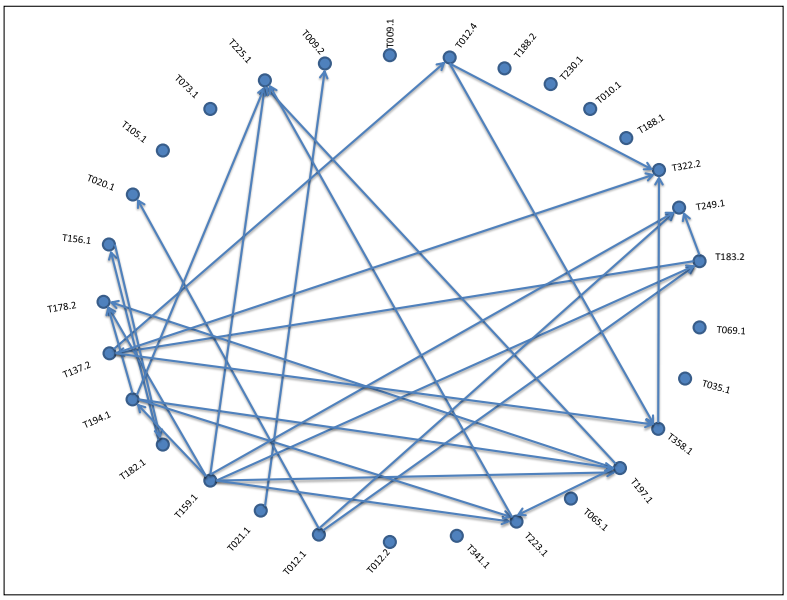


Figure 4.4: Timeline for patients observed positive in ward 2. Red shows the time that the patient was in the ward without having had a positive swab result and the time at which the timeline turns to green is the date of the first positive swab result. The labels are the other patients who are positive on the ward on the day when the first positive swab was taken from the patient under consideration, and the genetic distance between these two patients. Distances < 60 are in red.



(a) Ward 1



(b) Ward 2

Figure 4.5: Possible transmissions between patients positive at the same time with genetic distances < 60 in each ward. In Tong et al. [75] it is estimated that the maximum genetic distance between two sequences in the same cluster is 60. Patients within the same cluster are more likely to be in the same transmission chain.

4.3 The model for the spread of MRSA on a hospital ward

Here we recap the stochastic model introduced in chapter 2 for the spread of a pathogen within a single hospital ward in discrete-time. This model describes the dynamics of the spread of the pathogen on the level of individuals. Such a model can be used in

order to construct a transmission tree to show the spread of the pathogen from patient to patient throughout the ward. Since, in this nosocomial setting, we may have more than one importation of the pathogen to the ward we technically construct a *transmission forest* if we have two or more importation patients who start different trees. However, we will use the term *transmission tree* to refer to this potentially disjoint union of trees.

The model assumes that there is a hospital ward with a fixed number of beds to which patients can be admitted and discharged during the study over time $t = 0, 1, \ldots, L$. Any patients still on the ward at time $t = L$ are assumed to have a discharge day equal to L. Any patients already present on the ward at $t = 0$ have their 'admission' date set as $t = 0$. The day of admission of each patient, i, is given by t_i^a, and the day of discharge is given by t_i^d. These days are deterministic and do not form part of the 'stochastic' aspects of the model. Each patient may receive a number, ν_i, of screening tests, $t_i^t = t_{i,1}^t, t_{i,2}^t, \ldots, t_{i,\nu_i}^t$, which also occur at deterministic times. These tests produce a set of results $X_i = X_{i,1}, X_{i,2}, \ldots, X_{i,\nu_i}$ which are either negative or positive. The sensitivity of the screening test is represented by the parameter z, so a colonised patient has a probability of z of being screened positive, independently of all other screening tests. The specificity of the test is assumed to be 100%, so all uncolonised patients are screened negative.

The model assumes that each patient is colonised on admission with probability p, independently of all other patients. All uncolonised patients are assumed to have homogeneous susceptibility, and all colonised patients to have homogeneous infectivity. Since MRSA may be carried asymptomatically, we refer to patients being 'colonised' rather than 'infected'. Each susceptible patient on the ward avoids colonisation at time t with probability $\mathrm{P}(\mathrm{avoid}(t)) = \exp(-\beta C(t))$, where $C(t)$ is the number of colonised patients on the ward on day t. If the given patient does not avoid colonisation at time t then the source of their colonisation is picked uniformly at random from the set of colonised patients on day t, so the probability that a given susceptible patient is colonised by a given infectious patient at time t is defined as:

$$\frac{1 - \exp(-\beta C(t))}{C(t)}.$$

The day on which a patient, i, becomes colonised is given by t_i^c. For patients who remain susceptible $t_i^c = \infty$. A colonised patient is regarded as infectious from the day after colonisation, $t_i^c + 1$, until discharge from the ward at time t_i^d.

A colonised patient, i, who receives one or more positive screening test results may also have a number, ζ_i, of isolates of their pathogen sequenced. Genetic sequences, $Q_i = \{Q_{i,1}, Q_{i,2}, \ldots, Q_{i,\zeta_i}\}$, for patient i are produced from isolates sampled at times $t^s_{Q_i} = t^s_{Q_{i,1}}, t^s_{Q_{i,2}}, \ldots, t^s_{Q_{i,\zeta_i}}$. Each genetic sequence from each patient produces a set of genetic distances to every other sequence collected on the same day or earlier. The distributions from which the genetic distances between pairs of sequences are assumed to be drawn are given by the three different genetic models. These are briefly recalled here.

The Chain Error model

Under this model the genetic distance between a pair of sequences from patients who are separated by $k \leq 1$ or $k = \infty$ transmission events are drawn from the following distributions:

$$\mathrm{P}(\Psi_{i,j} = x) = \begin{cases} (\theta_{gl}^x / x!) \exp(-\theta_{gl}) & \text{if } k = \infty \\ (\theta_i^x / x!) \exp(-\theta_i) & \text{if } k = 0 \\ (\theta^x / x!) \exp(-\theta) & \text{if } k = 1, \end{cases} \tag{4.3.1}$$

where $x = 0, 1, \ldots$. The conditional probability distribution for genetic distances for pairs of sequences which are in the same chain but separated by more than one transmission event is:

$$\mathrm{P}(\Psi_{i,j} = x | D_{i,j}) = \frac{(k\gamma)^{|D_{i,j}-x|}}{|D_{i,j}-x|! \left(\sum_{l=0}^{D_{i,j}} (k\gamma)^l / l!\right)} \left(\frac{1}{2}\right)^{\mathbb{1}_{\{x \neq D_{i,j}\}}} \mathbb{1}_{\{x \leq 2D_{i,j}\}} \quad \text{if } k > 1. \tag{4.3.2}$$

Here $D_{i,j}$ is the sum of the consecutive distances between the isolates from the patients that compose the transmission chain between H_i and H_j. If H_i and H_j are separated by k transmission events such that H_i colonises p_1 who colonises p_2 etc. ($H_i \to p_1 \to p_2 \to \cdots \to p_{k-1} \to H_j$) then $D_{i,j} = \sum_{r=0}^{k-1} \Psi_{Q_{p_r,1}, Q_{p_{r+1},1}}$ where $p_0 = H_i$ and $p_k = H_j$. The genetic distance between sequences i and j is equal to $D_{i,j}$ with an error term which has a Poisson distribution with parameter $k\gamma$. This error term is added to $D_{i,j}$ with probability 0.5 or subtracted from $D_{i,j}$ with probability 0.5.

The Chain Poisson model

Genetic distances between a pair of sequences which belong to patients who are separated by $k \leq 1$ or $k = \infty$ transmission events under this model are drawn from the following distributions:

$$\mathrm{P}(\Psi_{i,j} = x) = \begin{cases} (\theta_{gl}^x / x!) \exp(-\theta_{gl}) & \text{if } k = \infty \\ (\theta_i^x / x!) \exp(-\theta_i) & \text{if } k = 0 \\ (\theta^x / x!) \exp(-\theta) & \text{if } k = 1. \end{cases} \tag{4.3.3}$$

The conditional probability distribution for genetic distances for pairs of sequences which are in the same chain but separated by more than one transmission event ($k > 1$) is:

$$\mathrm{P}(\Psi_{i,j} = x | D_{i,j}) = (D_{i,j}^x / x!) \exp(-D_{i,j}) \quad \text{if } k > 1. \tag{4.3.4}$$

Here we draw the pairwise genetic distances for sequences from patients in the same chain from a Poisson distribution with parameter $D_{i,j} = \sum_{r=0}^{k-1} \Psi_{Q_{p_r}, Q_{p_{r+1}}}$ where $p_0 = H_i$, $p_k = H_j$ and $\Psi_{Q_{p_r,1}, Q_{p_{r+1},1}}$ is the genetic distance between the first sequence from patient p_r and the first sequence from patient p_{r+1}. So $D_{i,j}$ is the sum of the genetic distances between the consecutive transmission events between the patients who have sequences i and j.

The Time Dependent Distances model

The Time Dependent Distances model uses the following distributions for the genetic distances between sequences from pairs of patients separated by $k \leq 1$ or $k = \infty$ transmission events:

$$\mathrm{P}(\Psi_{i,j} = x) = \begin{cases} (\theta_{gl}^x / x!) \exp(-\theta_{gl}) & \text{if } k = \infty \\ (\theta_i^x / x!) \exp(-\theta_i) & \text{if } k = 0 \\ ((t_{i,j}\theta)^x / x!) \exp(-(t_{i,j}\theta)) & \text{if } k = 1. \end{cases} \tag{4.3.5}$$

Here $t_{i,j} = |t_i^s - t_j^s|$ where t_i^s and t_j^s are the sampling times for sequences i and j. The conditional probability distribution for genetic distances for pairs of sequences which are in the same chain but separated by more than one transmission event is:

$$\mathrm{P}(\Psi_{i,j} = x | D_{i,j}) = (D_{i,j}^x / x!) \exp(-D_{i,j}) \quad \text{if } k > 1. \tag{4.3.6}$$

Again, $D_{i,j} = \sum_{r=0}^{k-1} \Psi_{Q_{p_r}, Q_{p_{r+1}}}$ where $p_0 = H_i$, $p_k = H_j$ and $\Psi_{Q_{p_r,1}, Q_{p_{r+1},1}}$ is the genetic distance between the first sequence from patient p_r and the first sequence from patient p_{r+1}. Therefore, $D_{i,j}$ is the sum of the distances between the consecutive transmission events between the patients from whom sequences i and j were taken.

The parameter vector, ρ, for our model is $\{p, z, \beta, \boldsymbol{\Theta}\}$ where $\boldsymbol{\Theta}$ is the vector of genetic diversity parameters.

4.4 Inference of parameters of the model for the spread of MRSA on a hospital ward

We can infer the parameters of the model in section 4.3 for the dataset introduced in section 4.2 which contains admission and discharge times for patients, pathogen screening test results, and the distances between the genetic sequences taken from the patients' pathogen isolates.

As well as inferring the parameters we can use this model to infer the unobserved transmission dynamics which are the admission states of the patients and the 'sources' and times for each patient's colonisation. The admission states $\phi = (\phi_1, \phi_2, \ldots, \phi_n)$ for each patient have the value 1 if the patient was already colonised when they were admitted to the ward, and value 0 if they were susceptible on admission. The sources for each of the n_{acq} patients who become colonised whilst on the ward are $s = (s_1, s_2, \ldots, s_{n_{acq}})$. For a patient i, who was admitted to the ward in a susceptible state and subsequently came into contact (directly or indirectly) with an infectious patient, j, which resulted in i becoming colonised, $s_i = j$. These unobserved colonisation events occur at times $t^c = (t_1^c, t_2^c, \ldots, t_{n_{pos}}^c)$. If a patient was colonised before admission to the ward their colonisation time is $t_i^c = -1$. Therefore, in order to completely specify the transmission tree we consider $T = \{t^c, \phi, s, \Psi^a\}$, the vector of these unobserved data and the unobserved genetic distances, Ψ^a, which are pairwise genetic distances between all patients' sequences and unobserved sequences from those patients for whom an isolate was not collected despite them carrying the pathogen.

4.4.1 Model likelihood

Using the model for the spread of MRSA in a hospital ward, we derive the likelihood of observing the set of screening test results, X, and the set of genetic distances, Ψ, between the genetic sequences from patients' isolates. The model likelihood that we are interested in, $\pi(X, \Psi|\rho, Z)$, is intractable, so we augment the parameter space with unobserved data, $T = \{t^c, \phi, s, \Psi^a\}$. This gives $\pi(X, \Psi|\rho, Z) = \sum_T \pi(X, \Psi, T|\rho, Z) = \sum_T \pi(X, \Psi|T, \rho, Z)\pi(T|\rho, Z)$, where Z is the vector of deterministic admission, discharge and screening times that are observed. We can not evaluate this sum since T is complicated and high-dimensional, but we can sample the parameters, ρ, and the transmission dynamics, T, from $\pi(X, \Psi|T, \rho, Z)\pi(T|\rho, Z)\pi(\rho)$, where $\pi(\rho)$ is the joint prior distribution of the parameters, using an MCMC routine. The contribution $\pi(X, \Psi|T, \rho, Z)$ is the likelihood of observing the distance matrix and screening re-

sults given the unobserved dynamics and parameters, and $\pi(T|\rho, Z)$ is the likelihood of the unobserved data given the parameters.

The likelihoods for the three different models are given below. Here, n_{seqs} will be the number of genetic sequences in the genetic distance matrix. The number of transmission events between the patients H_i and H_j from whom a pair of sequences, i and j, were taken, is given by trans(i, j). This number is calculated by working backwards from the patient with the later colonisation time and adding the number of sources along the transmission chain until either the other patient in the pair is reached, or an importation patient is reached. If an importation is reached in this calculation then the two sequences are from patients who are in unrelated transmission trees and trans$(i, j) = \infty$. If sequences i and j are from the same patient then trans$(i, j) = 0$. If trans$(i, j) > 1$ then $D_{i,j} = \sum_{r=0}^{k-1} \Psi_{Q_{p_r}, Q_{p_{r+1}}}$ where $p_0 = H_i$, $p_k = H_j$ and $\Psi_{Q_{p_r,1}, Q_{p_{r+1},1}}$ is the genetic distance between the first sequence from patient p_r and the first sequence from patient p_{r+1}. So $D_{i,j}$ is the sum of the genetic distances between sequences from consecutive sources working backwards along the transmission chain from H_i to H_j (or H_j to H_i if H_j has a later colonisation time).

Chain Error model The augmented likelihood for the Chain Error model is

$$\begin{aligned}
\pi(X, \Psi|T, \rho, Z)\pi(T|\rho, Z) = {} & z^{\mathrm{TP}(X)}(1-z)^{\mathrm{FN}(X,T)} p^{\sum_i \phi_i}(1-p)^{n-\sum_i \phi_i} \\
& \times \prod_{j=2}^{n_{seqs}} \prod_{i=1}^{j} \Bigg[\mathbb{1}_{\mathrm{trans}(i,j)=1} \frac{\theta^{\Psi_{i,j}} \exp(-\theta)}{\Psi_{i,j}!} \\
& + \mathbb{1}_{\mathrm{trans}(i,j)>1} \frac{(k\gamma)^{|D_{i,j}-\Psi_{i,j}|}}{|D_{i,j}-\Psi_{i,j}|! \left(\sum_{l=0}^{D_{i,j}} \frac{(k\gamma)^l}{l!}\right)} \left(\frac{1}{2}\right)^{\mathbb{1}_{\{\Psi_{i,j} \neq D_{i,j}\}}} \mathbb{1}_{\{\Psi_{i,j} \leq 2D_{i,j}\}} \\
& + \mathbb{1}_{\mathrm{trans}(i,j)=\infty} \frac{\theta_{gl}^{\Psi_{i,j}} \exp(-\theta_{gl})}{\Psi_{i,j}!} + \mathbb{1}_{\mathrm{trans}(i,j)=0} \frac{\theta_i^{\Psi_{i,j}} \exp(-\theta_i)}{\Psi_{i,j}!} \Bigg] \\
& \times \prod_{i=1}^{n} \left[\mathbb{1}_{t_i^c = t_i^a} + \mathbb{1}_{t_i^c \neq t_i^a} \exp\Big(- \sum_{t=t_i^a}^{min(t_i^c-1, t_i^d)} \beta C(t)\Big) \right] \\
& \times \prod_{\substack{j: t_j^c \neq \infty \\ \phi_j = 0}} \frac{(1 - \exp(-\beta C(t_j^c)))}{C(t_j^c)} \mathbb{1}_{\{s_j \in C(t)\}}, \qquad (4.4.1)
\end{aligned}$$

where the parameter vector, ρ, is $\{p, z, \beta, \mathbf{\Theta}\}$, s_j is the source of patient j's colonisation, and $\mathbf{\Theta} = \{\theta, \theta_i, \theta_{gl}, \gamma\}$ is the vector of genetic parameters. In this likelihood, F is

the cumulative distribution function of a Poisson random variable with parameter γ, TP(X) is the number of true positive screening results given the swab results X, and FN(X, T) is the number of false negative screening results, given the swab results X and augmented data T. The admission state of patient i, ϕ_i, is 0 if the patient was admitted to the ward in a susceptible state and 1 if they were admitted in a colonised state. The number of colonised patients present on the ward on day t is given by $C(t)$. The admission, colonisation and discharge times of patient i are given by t_i^a, t_i^c, and t_i^d respectively.

Chain Poisson model The augmented likelihood for the Chain Poisson model is

$$\begin{aligned}\pi(X,\Psi|T,\rho,Z)\pi(T|\rho,Z) = z^{\mathrm{TP}(X)}(1-z)^{\mathrm{FN}(X,T)}p^{\sum_i \phi_i}(1-p)^{n-\sum_i \phi_i}\\ \times \prod_{j=2}^{n_{seqs}}\prod_{i=1}^{j}\Bigg[\mathbb{1}_{\mathrm{trans}(i,j)=1}\frac{\theta^{\Psi_{i,j}}\exp(-\theta)}{\Psi_{i,j}!}\\ + \mathbb{1}_{\mathrm{trans}(i,j)>1}\frac{D_{i,j}^{\Psi_{i,j}}\exp(-D_{i,j})}{\Psi_{i,j}!}\\ + \mathbb{1}_{\mathrm{trans}(i,j)=\infty}\frac{\theta_{gl}^{\Psi_{i,j}}\exp(-\theta_{gl})}{\Psi_{i,j}!} + \mathbb{1}_{\mathrm{trans}(i,j)=0}\frac{\theta_i^{\Psi_{i,j}}\exp(-\theta_i)}{\Psi_{i,j}!}\Bigg]\\ \times \prod_{i=1}^{n}\left[\mathbb{1}_{t_i^c=t_i^a} + \mathbb{1}_{t_i^c\neq t_i^a}\exp\Big(-\sum_{t=t_i^a}^{min(t_i^c-1,t_i^d)}\beta C(t)\Big)\right]\\ \times \prod_{\substack{j:t_j^c\neq\infty\\ \phi_j=0}}\frac{(1-\exp(-\beta C(t_j^c)))}{C(t_j^c)}\mathbb{1}_{\{s_j\in C(t)\}},\end{aligned} \tag{4.4.2}$$

where TP(X) is the number of true positive screening results given the swab results X, s_j is the source of patient j's colonisation, and FN(X, T) is the number of false negative screening results, given the swab results X and augmented data T. The parameter vector, ρ, is $\{p, z, \beta, \boldsymbol{\Theta}\}$ and $\boldsymbol{\Theta} = \{\theta, \theta_i, \theta_{gl}\}$ is the vector of genetic parameters. The admission state of patient i, ϕ_i, is 0 if the patient was admitted to the ward in a susceptible state and 1 if they were admitted in a colonised state. The number of colonised patients present on the ward on day t is given by $C(t)$. The admission, colonisation and discharge times of patient i are given by t_i^a, t_i^c, and t_i^d respectively.

Time Dependent Distances model The augmented likelihood for the Time Dependent Distances model is

$$\begin{aligned}\pi(X,\Psi|T,\rho,Z)\pi(T|\rho,Z) = z^{\mathrm{TP}(X)}(1-z)^{\mathrm{FN}(X,T)}p^{\sum_i \phi_i}(1-p)^{n-\sum_i \phi_i}\\ \times \prod_{j=2}^{n_{seqs}}\prod_{i=1}^{j}\left[\mathbb{1}_{\mathrm{trans}(i,j)=1}\frac{(t_{i,j}\theta)^{\Psi_{i,j}}\exp(-(t_{i,j}\theta))}{\Psi_{i,j}!}\right.\\ + \mathbb{1}_{\mathrm{trans}(i,j)>1}\frac{D_{i,j}^{\Psi_{i,j}}\exp(-D_{i,j})}{\Psi_{i,j}!}\\ \left. + \mathbb{1}_{\mathrm{trans}(i,j)=\infty}\frac{\theta_{gl}^{\Psi_{i,j}}\exp(-\theta_{gl})}{\Psi_{i,j}!} + \mathbb{1}_{\mathrm{trans}(i,j)=0}\frac{\theta_{i}^{\Psi_{i,j}}\exp(-\theta_{i})}{\Psi_{i,j}!}\right]\\ \times \prod_{i=1}^{n}\left[\mathbb{1}_{t_i^c=t_i^a} + \mathbb{1}_{t_i^c\neq t_i^a}\exp(-\sum_{t=t_i^a}^{min(t_i^c-1,t_i^d)}\beta C(t))\right]\\ \times \prod_{\substack{j:t_j^c\neq\infty\\ \phi_j=0}}\frac{(1-\exp(-\beta C(t_j^c)))}{C(t_j^c)}\mathbb{1}_{\{s_j\in C(t)\}},\end{aligned} \tag{4.4.3}$$

where $t_{i,j} = |t_i^s - t_j^s|$ and t_i^s and t_j^s are the swab times for sequences i and j and s_j is the source of patient j's colonisation. In this likelihood TP(X) is the number of true positive screening results given the swab results X, and FN(X, T) is the number of false negative screening results, given the swab results X and augmented data T. The parameter vector, ρ, is $\{p, z, \beta, \mathbf{\Theta}\}$ and $\mathbf{\Theta} = \{\theta, \theta_i, \theta_{gl}\}$ is the vector of genetic parameters. The admission state of patient i, ϕ_i, is 0 if the patient was admitted to the ward in a susceptible state and 1 if they were admitted in a colonised state. The number of colonised patients present on the ward on day t is given by $C(t)$. The admission, colonisation and discharge times of patient i are given by t_i^a, t_i^c, and t_i^d respectively.

4.5 An MCMC routine for fitting the model for the spread of MRSA in a hospital ward

We can fit the models introduced in section 4.3 to the data introduced in section 4.2 by using a data-augmented MCMC routine to sample the parameters, ρ, and the transmission dynamics, T, from $\pi(X, \Psi|T, \rho, Z)\pi(T|\rho, Z)\pi(\rho)$, where $\pi(\rho)$ is the joint prior distribution of the parameters. At each iteration our MCMC routine first updates the

parameters of the model, ρ, before updating the structure of the unobserved transmission tree, T. In the next sections we describe our MCMC algorithm.

4.5.1 Parameter updates

Here we describe how each of the parameters in the model is updated by our MCMC algorithm. We show which parameters can be updated using Gibbs steps, and assign each parameter appropriate prior distributions.

4.5.1.1 Genetic parameter updates for each model

We assume that the genetic parameter θ has a $\Gamma(\nu_\theta, \lambda_\theta)$ prior distribution, which means that the probability density function $f(x) \propto x^{\nu_\theta - 1} \exp(-\lambda_\theta x)$ for $x > 0$. Therefore we may derive, up to proportionality, the full conditional distribution of θ, which is given as $\pi(\theta|\rho_{-\theta}, X, T)$ where $\rho_{-\theta}$ is the parameter vector without the component θ. Under the Time Dependent Distances model it follows from likelihood 4.4.3 that

$$\begin{aligned}
\pi(\theta|\rho_{-\theta}, X, T) &\propto \theta^{\nu_\theta - 1} \exp(-\lambda_\theta \theta) \prod_{j=2}^{n_{seqs}} \prod_{i=1}^{j} \mathbb{1}_{\text{trans}(i,j)=1} \frac{(t_{i,j}\theta)^{\Psi_{i,j}}}{\Psi_{i,j}!} \exp(-(t_{i,j}\theta)) \\
&\propto \theta^{\nu_\theta - 1} \exp(-\lambda_\theta \theta) \prod_{j=2}^{n_{seqs}} \prod_{i=1}^{j} \mathbb{1}_{\text{trans}(i,j)=1} \theta^{\Psi_{i,j}} \exp(-(t_{i,j}\theta)) \\
&\propto \theta^{\nu_\theta - 1} \exp(-\lambda_\theta \theta) \theta^{\sum_{(i,j)\in Y} \Psi_{i,j}} \prod_{j=2}^{n_{seqs}} \prod_{i=1}^{j} \mathbb{1}_{\text{trans}(i,j)=1} \exp(-(t_{i,j}\theta)) \\
&\propto \theta^{\sum_{(i,j)\in Y} \Psi_{i,j} + \nu_\theta - 1} \exp(-\lambda_\theta \theta) \exp(-\sum_{(i,j)\in Y} t_{i,j}\theta) \\
&\propto \theta^{\sum_{(i,j)\in Y} \Psi_{i,j} + \nu_\theta - 1} \exp(-\theta(\sum_{(i,j)\in Y} t_{i,j} + \lambda_\theta))
\end{aligned}$$

where Y is the set of (i, j) such that trans(i, j)= 1. It follows that θ may be sampled directly, using a Gibbs step, from the distribution:

$$\Gamma\left(\sum_{(i,j)\in Y} \Psi_{i,j} + \nu_\theta, \sum_{(i,j)\in Y} t_{i,j} + \lambda_\theta\right).$$

Similarly, for both the Chain Poisson model and the Chain Error model the parameter θ assigned $\theta \sim \Gamma(\nu_\theta, \lambda_\theta)$ *a priori* can be sampled from the distribution:

$$\Gamma\left(\sum_{(i,j)\in Y} \Psi_{i,j} + \nu_\theta, \sum_{(i,j)\in Y} N_{\text{trans}(i,j)=1} + \lambda_\theta\right)$$

where Y is the set of (i, j) such that trans(i,j)= 1 and $N_{\text{trans}(i,j)=1}$ is the total number of pairs of sequences for which trans(i,j)= 1.

The parameter θ_{gl} is assumed to have a $\Gamma(\nu_{\theta_{gl}}, \lambda_{\theta_{gl}})$ prior distribution, so under the each of the three genetic distance models θ_{gl} can be sampled from the distribution:

$$\Gamma\left(\sum_{(i,j)\in Y_g} \Psi_{i,j} + \nu_{\theta_{gl}}, \sum_{(i,j)\in Y_g} N_{\text{trans}(i,j)=\infty} + \lambda_{\theta_{gl}}\right)$$

where Y_g is the set of (i, j) such that trans(i,j)= ∞ and $N_{\text{trans}(i,j)=\infty}$ is the total number of pairs of sequences for which trans(i,j)= ∞, so the patients from which the pair of sequences are taken are in different transmission chains.

The parameter θ_i is assumed to have a $\Gamma(\nu_{\theta_i}, \lambda_{\theta_i})$ prior distribution and hence can similarly be sampled from the distribution

$$\Gamma\left(\sum_{(i,j)\in Y_i} \Psi_{i,j} + \nu_{\theta_i}, \sum_{(i,j)\in Y_i} N_{\text{trans}(i,j)=0} + \lambda_{\theta_i}\right)$$

where Y_i is the set of within-host distances, so the set of (i, j) such that trans(i,j)= 0 and $N_{\text{trans}(i,j)=0}$ is the total number of within-host distances. This distribution remains the same for the Chain Poisson model and the Chain Error model, as the way in which we model θ_i does not change.

The parameter γ in the Chain Error model is assumed to have an improper uniform prior distribution on $(0, \infty)$ and is updated using a Metropolis-Hastings random-walk. The step size of the random walk varies according to a Normal distribution with mean 0 and variance σ^2; the acceptance rate is checked every 1000 iterations in order to adjust the variance to maintain an acceptance rate between 0.2 and 0.6.

4.5.1.2 Epidemiological parameter updates

The importation parameter p may be updated using a Gibbs step for each of the three models.We assume that p had a Beta(α_p, β_p) prior distribution which is a Beta distribution with probability density function $f(x) \propto x^{\alpha_p - 1}(1 - x)^{\beta_p - 1}$ for $0 \leq x \leq 1$. Then the full conditional distribution of p, up to proportionality, may be derived as $\pi(p|\rho_{-p}, X, T) \propto p^{\alpha_p + \sum_i \phi_i}(1 - p)^{\beta_p + n - \sum_i \phi_i}$, so it follows that p may be sampled directly from the distribution:

$$\text{Beta}(\alpha_p + \sum_i \phi_i, \beta_i + n - \sum_i \phi_i)$$

where n is the number of patients in the study. The admission states, $\phi = \phi_1, \phi_2, \ldots, \phi_n$, have the value 1 if the patient is colonised before their admission to the ward, and 0 otherwise, so $\sum_i \phi_i$ is the number of patients who were colonised on admission, and $n - \sum_i \phi_i$ is the number of patients who were susceptible on admission.

Similarly, we assume that the sensitivity parameter z has a Beta(α_z, β_z) prior distribution so the full conditional distribution under each of the three models may be derived up to proportionality as $\pi(z|\rho_{-z}, X, T) \propto z^{\alpha_z + \mathrm{TP}(X)}(1-z)^{\beta_z + \mathrm{FN}(X,T)}$ and it follows that z may be sampled directly, using a Gibbs step, from the distribution:

$$\mathrm{Beta}(\alpha_z + \mathrm{TP}(X), \beta_z + \mathrm{FN}(X, T))$$

where $\mathrm{TP}(X)$ is is the number of true positive screening results given the swab results X, and $\mathrm{FN}(X, T)$ is the number of false negative screening results, given the swab results X and augmented data T.

The transmission parameter, β, is assumed to have an improper uniform prior distribution on $(0, \infty)$ and is updated using a Metropolis-Hastings random-walk under each of the three models. The step size of the random walk varies according to a Normal distribution with mean 0 and variance σ^2; the acceptance rate is checked every 1000 iterations in order to adjust the variance to maintain an acceptance rate between 0.2 and 0.6.

4.5.2 Augmented data updates

In order to estimate the unobserved transmission processes in the epidemic we use a data-augmented MCMC routine. Our MCMC algorithm has four data augmentation steps. At each iteration the algorithm performs one of these steps with equal probability, and running the algorithm for a large number of iterations will give us the posterior probability of possible transmissions between patients. During each step a candidate data set $T^* = \{s^*, t^{c*}, \phi^*, \Psi^{a*}\}$ is proposed. Here we describe each step, and define the proposal ratio, $q_{T,T^*} = \mathrm{P}(T^* \to T)/\,\mathrm{P}(T \to T^*)$, (the ratio of the probability of making the reverse move to the probability of making this move) for the model described in section 4.3. Detailed explanations of these proposal ratios can be found in appendix A.

- **Add colonisation.** In this move we select uniformly at random a currently uncolonised patient, i, and propose a colonisation for them. If there are no un-

colonised patients to choose from then no move is made. The number of uncolonised patients to choose from consists of all patients who are not observed positive, n_{sus}, minus those who currently have a colonisation added, n_{add}. With probability w the chosen patient is proposed to be colonised before admission to the ward, so they are an importation. With probability $1 - w$ the patient is proposed to be colonised by another infectious patient on the ward. In this case we draw a day of colonisation, t_i^{c*}, from $\{t_i^a, \ldots, t_i^d\}$. We select a source of colonisation at random from the set of colonised patients on this day. If there are no available patients to be a source on this day, the move is not made. If the move is possible, whether we are inferring an importation or a colonisation on the ward (acquisition), we then draw a set of genetic distances, $\Psi^a_{i,1}, ..., \Psi^a_{i,n_{seqs}+n_{add}}$, from the patient to every other sequence from every colonised patient (those observed positive in the data, and those currently added by the algorithm). These distances are drawn according to the relevant probability distributions depending on the genetic distance model and whether the two patients from whom the sequences are collected are in the same transmission chain and adjacent to each other, in the same chain but separated by more than one transmission event, or in separate chains. The relevant probability distributions for the Chain Error model are given in equations 4.3.1 and 4.3.2 on page 76, for the Chain Poisson model are given in equations 4.3.3 and 4.3.4 on page 76, and for the Time Dependent Distances model are given in equations 4.3.5 and 4.3.6 on page 77.

If n_{add_0} is the number of patients who currently have no offspring (we define the offspring of a patient to be those patients who are inferred to have this patient as the source of their colonisation) as well as a colonisation time added by the algorithm, the proposal ratio is for adding an importation is

$$q_{T,T^*} = \frac{n_{sus} - n_{add}}{w(1 + n_{add_0})Y_{add}},$$

and the proposal ratio of adding an acquisition is

$$q_{T,T^*} = \frac{C(t_i^{c*})(n_{sus} - n_{add})(t_i^d - t_i^a + 1)}{(1 - w)(n_{add_0} + 1)Y_{add}},$$

where

$$Y_{add} = \prod_{j=1}^{n_{seqs}+n_{noseqs}+n_{add}} \mathrm{P}\left[\Psi_{i*,j} = \Psi^{a*}_{i*,j} | \Theta\right]$$

where Θ is the vector of genetics parameters. The number of patients who have a positive swab result but no genetic sequence taken is given by n_{noseq}, the number of sequences in collected is given by n_{seqs} and the number of sequences currently added by the algorithm is given by n_{add}.

- **Remove colonisation.** In this move we select uniformly at random one of the n_{add_0} currently added colonised patients who are not inferred to be the source of any other colonisations. We can not make this move if no such individuals exist. If we are removing a patient who is assumed to be an importation the proposal ratio is

$$q_{T,T^*} = \frac{n_{add_0} \cdot w \cdot \Upsilon_{rm}}{n_{sus} - n_{add} + 1},$$

and if we are removing a patient assumed to have acquired MRSA during their stay in the ward the proposal ratio is

$$q_{T,T^*} = \frac{\Upsilon_{rm} \cdot n_{add_0} \cdot (1-w)}{(t_i^d - t_i^a + 1)(n_{sus} - n_{add} + 1)(C(t_i^c) - 1)},$$

where

$$\Upsilon_{rm} = \prod_{j:\, i \neq j} \mathrm{P}\left[\Psi_{i,j} = \Psi_{i,j}^a | \Theta\right].$$

- **Move a colonisation time.** In this move we pick a patient uniformly at random from the number of colonised patients, $n_{seqs} + n_{noseq} + n_{add}$, and move their colonisation time. As when we added a colonisation time, we propose that the patient was positive on admission with probability w. With probability $1 - w$ the patient acquired the pathogen whilst on the ward so we sample a colonisation time uniformly at random from the days in the interval $\{t_i^a, t_{i+1}^a, \ldots, f_i\}$ where f_i is the latest day on which this patient could have been colonised. This is either the day of the patient's first positive swab result, or the day of the 'birth' of the patient's first offspring, whichever is smaller, or the day of discharge for those patients who have neither positive swabs nor offspring. We then uniformly at random draw a source for the patient from the set of colonised patients on the chosen day of colonisation. If there are no positive patients no move is made.

  If the patient we choose to move is an acquisition on day t_i^c and we reassign them as an acquisition on day t_i^{c*} the proposal ratio is

$$q_{T,T^*} = \frac{C(t_i^{c*})}{C(t_i^c)}.$$

  If the patient that we choose to move is an acquisition on day t_i^c and we reassign them as an importation on day t_i^a the proposal ratio is

$$q_{T,T^*} = \frac{1-w}{w(f_i - t_i^a + 1)C(t_i^c)}.$$

If the patient that we choose to move is an importation on day t_i^a and we reassign them as an acquisition on day t_i^{c*} the proposal ratio is

$$q_{T,T^*} = \frac{w \cdot (f_i - t_i^a + 1) \cdot C(t_i^{c*})}{1 - w}.$$

If the patient that we choose to move is an importation on day t_i^a and we reassign them as an importation on day t_i^a the proposal ratio is $q_{T,T^*} = 1$.

- **Change genetic distances.** In this move we pick uniformly at random one of the $n_{noseq} + n_{add}$ patients who either have a positive swab but no sequence, or have an added colonisation time, and change the genetic distances between their sequence (or one of their sequences picked uniformly at random if they have more than one) and each other sequence. If no such patients exist no move is made. We draw a new set of distances between this sequence and each other sequence from each colonised patient according to the relevant probability distributions. The relevant probability distributions for the Chain Error model are given in equations 4.3.1 and 4.3.2 on page 76, for the Chain Poisson model are given in equations 4.3.3 and 4.3.4 on page 76, and for the Time Dependent Distances model are given in equations 4.3.5 and 4.3.6 on page 77. The proposal ratio for this move is

$$q_{T,T^*} = \frac{\prod_{j:\, i \neq j} \mathrm{P}\left[\Psi_{i,j} = \Psi_{i,j}^a | \Theta\right]}{\prod_{j:\, i \neq j} \mathrm{P}\left[\Psi_{i,j} = \Psi_{i,j}^{a*} | \Theta\right]}.$$

Acceptance probability

For each of the augmented data updates described above we accept the proposed augmented data set with probability

$$\min\left(1, \frac{\pi(X, \Psi | T^*, \rho)\pi(T^* | \rho)}{\pi(X, \Psi | T, \rho)\pi(T | \rho)} q_{T,T^*}\right).$$

4.5.3 Improvements to augmented data steps in the Worby et al. model

The augmented data updates used by Worby et al. in the MCMC algorithm for fitting their models are not greatly different from those stated here for our model. However, there are two changes that we made to the augmented data steps to improve the mixing of the algorithm that also improve the Worby et al. method. When we ran

the Worby et al. algorithm without these steps we found that the model was prone to get stuck in certain regions of the likelihood and so not explore the space well. Introducing these changes helped with that problem. We now explain what these changes are.

4.5.3.1 Proposal distributions for new genetic distances

The Transmission Chain Diversity model and the Importation Structure model MCMC algorithms draw genetic distances for sequences from patients for whom a colonisation is added by the algorithm, or those who are positive but without a sequenced isolate, from probability mass functions $m(\cdot)$ and $m_G(\cdot)$, for strains in the same group and different groups respectively. These distributions are completely defined pre-analysis, meaning that any parameters are static throughout the algorithm. We adapted this to allow for the distributions to take the same parameter values already being inferred in our MCMC algorithm. This improved mixing of the algorithm as we were more likely to draw distances that we would accept.

4.5.3.2 Changing genetic distances for added patients

In the Worby et al. algorithm the step in which the genetic distances are changed only alters the distances of those patients for whom we had a positive swab result but no genetic sequence. In our MCMC algorithm we propose instead to change the distances of sequences from any patient for whom a colonisation time has been added by the algorithm as well as those whom we know are positive with no sequence. In theory the Worby et al. algorithm can explore different genetic distances for added colonisations by deleting added colonisations and re-adding the same patient but with different distances. However, we found that mixing was highly improved when we were able to update the distances for sequences from these patients in a separate step to the addition and deletion steps.

4.6 Additional MCMC update steps for the Thai data

The MCMC algorithm with the augmented data steps described in 4.5.2 was used to fit each of the three models to the data from the hospital in Thailand. Parameters p and z were assigned $\mathrm{U}(0,1)$ prior distributions, parameters θ, θ_i and θ_{gl} were assigned $\Gamma(1, 10^{-6})$ prior distributions and parameters β and γ were assigned improper uniform prior distributions on $(0, \infty)$. The parameters were initially given

values based on the results of Worby [69]. We initialised the infection times by giving each patient who had received a positive swab a colonisation time of the day before their first positive swab. If a patient's first positive swab was on their day of admission they were assigned as an importation. For patients who were not importations and had a positive test we drew a source uniformly at random from the set of other colonised patients on the ward on the day of colonisation. If no source was available we reassigned that patient as an importation. We initialised the missing sequences by drawing a genetic distance between each patient who had a positive swab but no sequence and each other patient sequence from a Poisson distribution with mean 30.

The Chain Poisson model and the Time Dependent Distances model showed good mixing, however the Chain Error model, with its extra parameter and more complex distribution, showed evidence of the chain getting stuck at certain configurations of the transmission tree and not making any further moves. Therefore we created additional augmented data steps to enable the chain to move away from these positions. An example of the resulting traceplots from the improved MCMC algorithm for each model on Ward 1 can be found in appendix B.

4.6.1 Additional data augmentation

- **Updating θ and θ_{gl} alongside the genetic distances**
  From the output we detected that it was the likelihood of the genetic distances that was preventing the algorithm from moving. Therefore we proposed to update both the distance parameters and the genetic distances for one patient at the same time. In this move we propose a new value for θ and θ_{gl} using a Gaussian random-walk: the step size of the random walk varies according to a Normal distribution with mean 0 and variance σ^2, and we check the acceptance rate every 1000 iterations in order to adjust the variance to maintain an acceptance rate between 0.2 and 0.6. Then we pick a patient at random from the patients either with a positive swab but no sequence, of which there are a total of n_{noseq}, or with an added colonisation time, of which there are a total of n_{add}, and change the genetic distances from their sequence to each other sequence from other colonised patients using the proposed values θ^* and θ^*_{gl}. If no such patients exist no move is made and the parameters retain their previous values. We draw a new set of distances between this patient's sequence and each other colonised patient's sequences according to the relevant probability distributions. The relevant probability distributions for the Chain Error model

are given in equations 4.3.1 and 4.3.2 on page 76, for the Chain Poisson model are given in equations 4.3.3 and 4.3.4 on page 76, and for the Time Dependent Distances model are given in equations 4.3.5 and 4.3.6 on page 77. The proposal ratio for this move is

$$q_{T,T^*} = \frac{\prod\limits_{j:\, i \neq j} \mathrm{P}\left[\, \Psi_{i,j} = \Psi^{a}_{i,j} | \Theta \right].}{\prod\limits_{j:\, i \neq j} \mathrm{P}\left[\, \Psi_{i,j} = \Psi^{a*}_{i,j} | \Theta^* \right].}$$

- **Updating one distance at a time using a random-walk**
  In the steps which update the genetic distances described in section 4.5.2, a patient with distances added by the algorithm is selected and we propose to update all of the distances from their sequence to each other sequence from every colonised patient by drawing each distance from a probability distribution. An alternative is to update the genetic distances one at a time via a random-walk. Therefore we pick uniformly at random from the patients either with a positive swab but no sequence, of which there are a total of n_{noseq}, or with an added colonisation time, of which there are a total of n_{add}, and then uniformly at randomly select one of the distances between their sequence and one other sequence to update. We update this distance by either adding, with probability $\frac{1}{2}$, or subtracting, with probability $\frac{1}{2}$, from the current distance. The integer to be added or subtracted is drawn from a Poisson distribution with parameter ϖ which is specified before we run the algorithm. We allow for this parameter to be scaled up or down if the algorithm is either accepting too many ($> 80\%$) or not enough ($< 20\%$) proposals. Here the probability of the move and the reverse move are equal, so the proposal ratio is 1.

- **Swap a source with one of their offspring**
  In this move we pick a pair of patients who are connected by a direct transmission event in the current configuration of the transmission tree and swap their places in the chain. The idea behind this move is that the chain gets stuck in configurations of the transmission tree because the jump to an entirely new tree is too large, but it is easier to find smaller steps that will be accepted.

  To execute this move a patient, j, is picked uniformly at random from the total number of patients who are colonised during their stay on the ward ($n_{acq} = n_{seqs} + n_{noseq} + n_{add} - n_{imp}$, where n_{imp} is the number of importations of the pathogen to the ward). We then find this patient's source of colonisation, i, and this is the pair of patients which we swap. If the source patient colonises

another patient before the one we have picked there is no move made. If the colonisation time of the source patient is before the time at which the selected offspring is admitted to the ward there is no move made. If there is a move to be made, we record the time of colonisation of j, t_{swap} before setting $t_j = t_i$ and the source of colonisation $s_j = s_i$. Then we set $t_i = t_{swap}$ and the source of colonisation $s_i = j$. Here the probability of the move and the reverse move are equal, so the proposal ratio is 1.

- **Change a source without changing the time of colonisation**
  In this move we pick one colonised patient and change their source of colonisation whilst keeping their colonisation time the same. Again, the idea is that a smaller move should be more easily accepted, allowing the algorithm to move away from the configurations of the transmission tree that it was getting stuck in.

  Once we have chosen a patient, i, uniformly at random from the total number of patients who are colonised during their stay on the ward ($n_{acq} = n_{seqs} + n_{noseq} + n_{add} - n_{imp}$, where n_{imp} is the number of importations of the pathogen to the ward), we then pick uniformly at random from the set of other colonised patients present on that day to obtain a new source, j, for the patient. We set $s_i = j$, and there is no need to change the source of any other patients who go on to be colonised by this one as we have not changed the colonisation time, so it is still available to colonise them. Here the probability of the move and the reverse move are equal, so the proposal ratio is 1.

4.7 Simulation study

In order to assess the performance of our MCMC algorithm we performed a simulation study where we simulated a number of outbreaks of MRSA according to our model, and then fitted the model using the MCMC routine to check that the parameters and transmission tree were recovered. The parameters were again assigned uninformative prior distributions: $\mathrm{U}(0,1)$ for p and z, $\Gamma(1,10^{-6})$ for θ, θ_i and θ_{gl} and improper uniform distributions on $(0,\infty)$ for β and γ. The initialisation was performed using the same method described in section 4.6. Here we describe the method for simulating data from our models, and then assess the output from the MCMC algorithm on simulations with different values for the parameters.

4.7.1 Simulation method

In order to simulate an outbreak of MRSA on a hospital ward we specify the number of patients in the study, n, the length of the study, L, and the average length of stay for patients on the ward, A. We assume that the ward is initially empty. We set a test frequency, κ, to allow for tests to be taken from colonised patients. This means tests are taken every κ days from all patients who are presents on the ward on that day, so the set of test days t^t can be generated independently of the patient stays.

For each of the n patients we draw a date of admission to the ward uniformly at random from time 0 to time L. We draw their length of stay from a Poisson distribution with parameter A. Each of these patients is independently admitted colonised with probability p.

Patients who are not admitted colonised either remain susceptible for their whole stay on the ward, or become colonised through contact with another colonised patient. A susceptible patient, i, avoids colonisation on day t with probability $\mathrm{P}(\mathrm{avoid}(t)) = \exp(-\beta C(t))$, where $C(t)$ is the number of colonised patients present on the ward on day t, so the number of importation patients who have arrived on or before day t and are discharged after day t, plus the number of patients who acquire the pathogen before day t and are discharged after day t. If patient i does not avoid colonisation on day t then they acquire the pathogen and $t_i^c = t$. A source of colonisation is drawn for this patient's colonisation uniformly at random from the $C(t_i^c)$ patients available to colonise them.

For each patient, i, who is colonised, either before admission to the ward or during their stay on the ward, we generate a test result for each of the test days, t^t, that patient i was present on the ward for. When positive patients are tested the result of their test is positive with probability z, and negative with probability $1 - z$, independently of all other tests. When negative patients are tested their tests are always negative, so in effect we only need to simulate test results for patients who are colonised on the ward. We assume that for each positive swab result a patient receives a genetic sequence is observed. We also assume that a patient i who is colonised but never receives a positive test result has an unobserved sequence on their day of colonisation t_i^c which has a genetic distance to each sequence (observed or unobserved) on day t_i^c or earlier. We assume each sequence taken on day t has a genetic distance to each sequence (observed or unobserved) on day t or earlier.

We draw these genetic distances from the distributions specified by the model, according to the relative positions on the transmission tree of the patients who have sequences i and j. For sequences from patients who are in distinct transmission chains, all models draw the genetic distance from a Poisson(θ_{gl}) distribution. For sequences from patients who share a direct transmission event, the Chain Error model and Chain Poisson model draw the genetic distance from a Poisson(θ) distribution and the Time Dependent Distances model draws the genetic distance from a Poisson$(t_{i,j}\theta)$ distribution. For sequences from patients who are in the same transmission chain but are separated by more than one transmission event the Chain Poisson model and the Time Dependent Distances model draw the genetic distance from a Poisson$(D_{i,j})$ distribution where $D_{i,j}$ is the sum of the underlying distances in the transmission chain, so $D_{i,j} = \sum_{r=0}^{k-1} \Psi_{Q_{p_r}, Q_{p_{r+1}}}$ where $p_0 = H_i$, $p_k = H_j$. The Chain Error model draws this distance by adding or subtracting from $D_{i,j}$, with probability 0.5, an error term drawn from a Poisson distribution with parameter $k\gamma$ which has been truncated at the value $D_{i,j}$. The genetic distance between two sequences taken from the same patient is drawn from a Poisson distribution with parameter θ_i under each of the three models.

4.7.2 Results of the simulation study

In order to assess the performance of the MCMC algorithm we discuss the quality of the parameter estimation and the network reconstruction from our simulation study. For each of the three models, the Chain Error model, the Chain Poisson model and the Time Dependent Distances model, we simulated a number of data sets with different values of the parameters for 100 patient admissions over 100 days with an average length of stay of 7 days. We set tests to be taken from the positive patients every 3 days.

To investigate the MCMC algorithm's ability to recover the parameters we varied these parameters one at a time. For each parameter, ρ_i, we fixed the other parameters, ρ_{-i}, to 'sensible' values which would allow us to see the impact of varying ρ_i. Our choices of 'sensible' parameters were informed by the results of the Worby et al. model [69].

Each parameter ρ_i was varied over a range of values which included extreme values in order to assess the performance of the algorithm in these cases:

- For the importation parameter p we fixed the other parameters ($z = 0.7, \beta = 0.01, \theta = 40$ for the Chain Error model and the Chain Poisson model or 0.1 for the Time Dependent Distance model, $\theta_{gl} = 300$) and varied p between 0 and 1 in increments of 0.1.

- For the test sensitivity parameter z we fixed the other parameters ($p = 0.2, z = 0.01, \theta = 40$ or 0.1, $\theta_{gl} = 300$) and varied z between 0 and 1 in increments of 0.1.

- For the transmission parameter β we fixed the other parameters ($p = 0.6, z = 0.7, \theta = 40$ or 0.1, $\theta_{gl} = 300$) and varied β between 0.005 and 0.05 in increments of 0.005.

- For the global genetic parameter θ_{gl} we fixed the other parameters ($p = 0.2, z = 0.7, \beta = 0.01, \theta = 40$ or 0.1) and varied θ_{gl} between 40 and 400 in increments of 40.

- For the chain genetic parameter θ for the Chain Error model and Chain Poisson model we fixed the other parameters ($p = 0.2, z = 0.7, \beta = 0.1, \theta_{gl} = 300$) and varied θ between 10 and 55 in increments of 5.

- For the chain genetic parameter θ for the Time Dependent Distances model we fixed the other parameters ($p = 0.2, z = 0.7, \beta = 0.1, \theta_{gl} = 300$) and varied θ between 0 and 1 in increments of 0.1.

Therefore each of the 5 parameters was varied to create 10 sets of parameters each (50 sets in total). For each of these 10 sets of parameters we simulated 10 outbreaks of MRSA.

4.7.2.1 Parameter estimation

The idea behind creating data sets with varying values of each parameter was to assess the performance of the MCMC algorithm across a range of conceivable values and to find areas where it might be limited. For the investigation into each parameter we ran the MCMC routine for 50,000 iterations on each simulated dataset and plotted each resulting posterior estimate for the parameter of interest as a boxplot on the same graph as for the other 99 simulations. An example for varying θ for the Chain Poisson model is shown in figure 4.6. From this graph it is easy to see whether we have recovered the fact that θ is increasing but it is not visually obvious exactly how well we have recovered the specific values for θ so for this we separate the boxplots into individual graphs for each value of θ and plot them over a line which shows the true

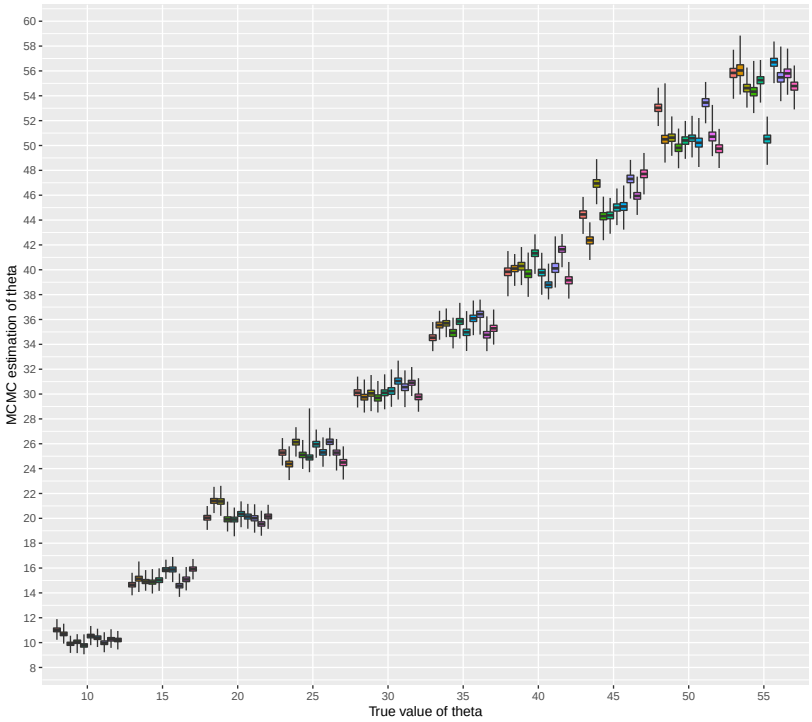


Figure 4.6: The posterior estimates of parameter θ from fitting the Chain Poisson model to 100 simulated datasets from the same model with input value for θ which varies from 10 to 55, with 10 simulations for each increase.

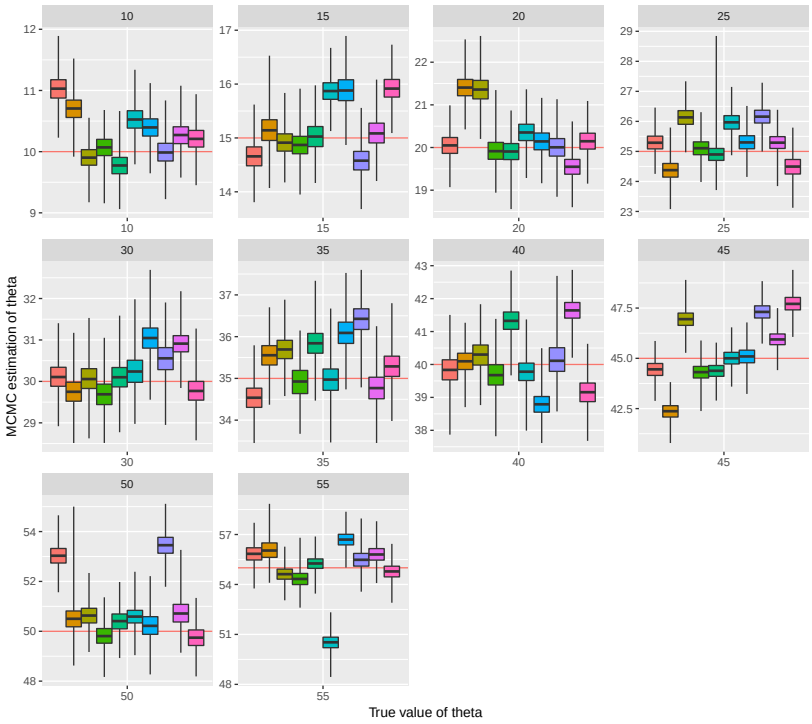


Figure 4.7: The posterior estimates of parameter θ from fitting the Chain Poisson model to 100 simulated datasets from the same model with input value for θ which varies from 10 to 55, with 10 simulations for each increase.

value, as in figure 4.7. In this example is is clear that the algorithm has recovered the increase of θ and that it estimates the specific value for θ well, although as θ increases so does the spread of the estimate around the true value. The graphs for the other parameters can be found in appendix C, but we will discuss what they show us here.

Importation parameter p

From the graphs C.1, C.2, C.11, C.12, C.21 and C.22 we can see that the algorithm estimates the importation parameter p well over the range 0.1 to 1 for each of the three models. When $p = 1$ the algorithm slightly underestimates p, which is understandable because in this case every patient is an importation, but the algorithm has to move around the tree space which means exploring a lot of space where there are transmissions between patients.

Sensitivity parameter z

From the graphs in figures C.3, C.4, C.13, C.14, C.23 and C.24 we can see that the algorithm estimates the sensitivity parameter z fairly well over the range 0.1 to 1, although for the lower end (0.1- 0.3) the success is more variable, especially for the Chain Error model. This is not surprising, as when the sensitivity parameter is low there will be a lot more uncertainty about the tree structure and the course of the epidemic so it will be harder to estimate the parameters. We can also see that the algorithm begins to underestimate the value of z more often once it gets close to 1, but the estimates are still generally within 0.1 of the true value.

Transmission parameter β

From the graphs in figures C.5 and C.6 we can see that for the Chain Error model, the estimates of β do capture the increases in its value, but there are a few in the lower values (0.005- 0.001) which have very wide ranges, suggesting a lot of uncertainty around their estimation. Also, for the higher values (0.035- 0.05) the boxplots start to fall below the line of the true value, and this is even more clear in the graphs for the Chain Poisson model (C.15, C.16) and the Time Dependent Distances model (C.25, C.26), although these do not have the same problem with the lower values. This may be due to the fact that the difference in the final size of the epidemics created with parameter β between 0.035 and 0.05 is not as significant as the difference in final size when β varies between 0.005 and 0.03.

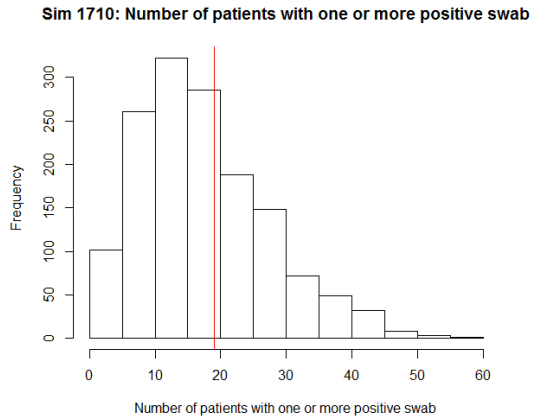


Figure 4.8: A histogram comparing the number of patients observed positive in a simulated dataset to the number observed positive in 100 simulations with the parameters estimated by MCMC algorithm for the original dataset

Chain genetic parameter θ

The graphs in figures 4.6, 4.7, C.7 and C.8 show that θ is well estimated by the algorithm for the Chain Error model and the Chain Poisson model across the range 10 to 55. For the Time Dependent Distance model (C.27, C.28) the parameter has a slightly different definition due to the factor $t_{i,j}$ in the distribution. The graphs show that the algorithm captures the general increase of θ as we vary it, but the precise estimation is very variable, and becomes more variable as the parameter value increases. This is understandable as the variation in the times between patients' samples and their colonisation which are tied up in the estimation of θ are themselves quite varied.

Global genetic parameter θ_{gl}

The graphs in figures C.9, C.10, C.19, C.20, C.29 and C.30 show that θ_{gl} is well estimated by the algorithm for each of the three models across the range 40 to 400.

As a further test of our parameter estimations we simulated a dataset, ran the MCMC algorithm to obtain parameter estimates and then simulated 100 datasets from those parameter estimates and plotted some summary statistics of those 100 datasets compared to the original 'true' simulation. Figures 4.8 and 4.9 show the difference in the number of patients with one or more positive swabs on a 'true' simulated ward compared to 1000 simulations with the parameter estimates made by our algorithm for the 'true' ward. It is clear that there can be a lot of variation in simulations despite them using the same parameter values. However, as the 'truth' is well within the credible bounds it seems that the estimation of our parameters is reasonable.

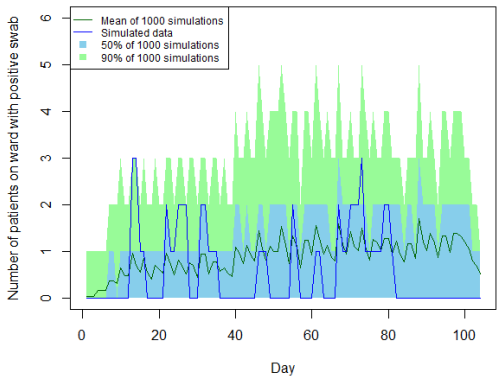


Figure 4.9: Comparing the number of patients with positive swab(s) on the simulated ward over time with the number of patients with positive swab(s) over time on 1000 wards simulated with the parameter estimates from the MCMC algorithm run on the original simulated dataset

4.7.2.2 Transmission tree estimation

We used the same set of simulations that were used for investigating parameter estimation in section 4.7.2.1 to investigate the strengths and weaknesses of the algorithm in recovering the transmission tree for simulations created with a range of values for each parameter.

In order to visualise how many of the transmission events were correctly estimated by the algorithm, for each simulation we took the output of the MCMC algorithm and found the most likely source for each patient's colonisation by finding which source was assigned to them for the largest number of iterations (the burn in time was excluded) and compared this to their true source. We produced separate plots for each parameter in each version of the model. We plotted a boxplot for each value of the parameter being investigated which shows the proportion of sources correctly identified for each of the 10 simulations with that value. An example is given here in figure 4.10 which shows the proportion of sources correctly identified for varying values of the sensitivity parameter z for the Chain Error model. The plots for the other parameters and models are found in appendix D but we discuss what they show here.

Effect of varying importation parameter p

From the graphs in figures D.1, D.6 and D.11 we can see that for each of the models the source estimation gets steadily better as the value of the importation parameter p increases. This is because it is easier for the algorithm to correctly identify an impor-

tation than it is for it to correctly identify the source of an acquisition from the population of colonised patients on the ward. However, the algorithm only gets fewer than 50% of the sources right for $p = 0.1$ and once $p > 0.3$ the source estimation is always near to 75%.

Effect of varying sensitivity parameter z

From the graphs in figures 4.10, D.7 and D.12 we can see that the algorithm for each variation of the model performs as we would expect with regards to transmission tree estimation under variations of the test sensitivity in that the higher the sensitivity, the better we recover the tree. Clearly this is because as the sensitivity increases we have more correct information about when the patients were colonised.

Effect of varying transmission parameter β

From the graphs in figures D.3, D.8 and D.13 for each version of the model we can see that the proportion of sources correctly identified decreases as β increases. This is because as β increases the transmission tree becomes more complicated and harder to recover fully. However, for the Chain Poisson model most of the simulations were still recovered with over 75% accuracy, and for the other two models most were recovered with over 50% accuracy. Also, these plots only take into account the most common source given to the patient by the algorithm, so it could be that the second most common source is the correct one.

Effect of varying genetic parameters θ **and** θ_{gl}

From the graphs in figures D.4, D.9 we can see that variation in the value of θ does not seem to affect the recovery of the transmission tree for the Chain Error model or the Chain Poisson model, and the tree was consistently recovered quite well. For the Time Dependent Distances model (D.14) there does not appear to be much variation in the proportion of sources recovered correctly until $\theta > 0.8$. This could be because once θ gets this large the distribution gives similar draws for the chain genetic distances to those which are drawn for the global genetic distances, making the transmission tree harder to recover. From the graphs in figures D.5, D.10 and D.15 we see that for all the models the proportion of sources recovered correctly increases as the value of θ_{gl} gets larger. This is because as it gets larger the distinction between genetic distances drawn for unrelated patients and those drawn for related patients becomes more obvious, making it easier to discover which patients are in the same chains of transmission.

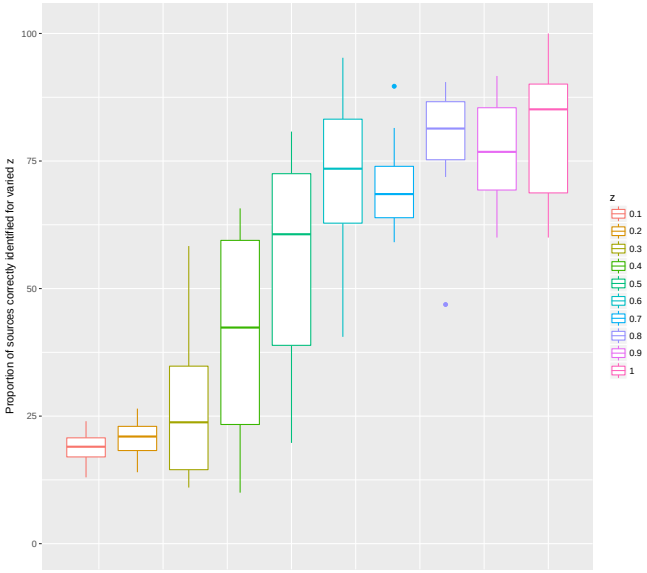


Figure 4.10: Boxplots to show the proportion of colonisation sources for patients recovered correctly for simulations with varied values for parameter z for the Chain Error model.

4.8 Analysing the Thai hospital data

Having tested the algorithm on simulated data, we analysed the available data from an outbreak of MRSA on two hospital wards in Thailand from 2008. We analysed the data under each of our three models in order to estimate the transmission tree, which would suggest who colonised whom on each ward, and when these transmissions of the pathogen took place. For each model on each ward we ran the MCMC algorithm for 200,000 iterations, with 10 augmented data steps taking place during each iteration. The prior distributions for the parameters were as follows:

$$\begin{aligned} p &\sim \mathrm{U}(0,1), \\ z &\sim \mathrm{U}(0,1), \\ \theta &\sim \Gamma(1,10^{-6}), \\ \theta_i &\sim \Gamma(1,10^{-6}), \\ \theta_{gl} &\sim \Gamma(1,10^{-6}), \end{aligned}$$

and the parameters β and γ (in the Chain Error model) were given improper uniform prior distributions on $(0,\infty)$.

We initialised the infection times by giving each patient who had received a positive swab a colonisation time of the day before their first positive swab. If a patient's first positive swab was on their day of admission they were assigned as an impor-

tation. For patients who were not importations and had a positive test we drew a source uniformly at random from the set of other colonised patients on the ward on the day of colonisation. If no source was available we reassigned that patient as an importation. We initialised the missing sequences by drawing a genetic distance between each patient who had a positive swab but no sequence and each other patient sequence from a Poisson distribution with mean 30. The parameters were initially given values based on the results of Worby [69].

We checked for convergence by inspecting the traceplots of the output of the parameter estimates. We checked that we had not converged to a local mode by altering the initial state of the chain a number of times and running the MCMC algorithm again to ensure that the posterior estimates were the same.

4.8.1 Results from the Chain Error model on each ward

We performed analysis under the Chain Error model on each ward separately. The posterior mean estimates of the parameters with 95% equitailed credible intervals are given in tables 4.1 and 4.2. The model assumes that sequences from patients who share a direct transmission event have their genetic distance drawn from a Poisson distribution with parameter θ, and that those from patients who are in distinct chains of transmission have their genetic distance drawn from a Poisson distribution with parameter θ_{gl}. The analysis for ward 1, the surgery ward, suggests that genetic sequences from immediately linked patients are expected to differ by 40 SNPs with 95% credible interval of $(38, 41)$ SNPs. Sequences from unlinked patients are expected to differ by 381 $(379, 383)$ SNPs. The analysis for ward 2, the paediatric ward, suggests that genetic sequences from linked patients are expected to differ by 52 $(50, 54)$ SNPs, and that sequences from unlinked patients are expected to differ by 339 $(337, 341)$ SNPs.

For ward 1 the model estimated that 5% $(2\%, 9\%)$ of patients who were admitted to the ward were already colonised with MRSA and for ward 2 the model estimated that 7% $(3\%, 12\%)$ of patients were colonised before admission to the ward. It was estimated that the test sensitivity for the surgery ward was 72% $(59\%, 83\%)$, and the test sensitivity for the paediatric ward was 79% $(68\%, 84\%)$. For ward 1 it was estimated that the transmission rate was equivalent to 1.3 $(0.7, 2.1)$ patients colonised by each infectious patient per 100 days on the ward. For ward 2 it was estimated that the transmission rate was equivalent to 1.0 $(0.6, 1.4)$ patients colonised by each infectious

Ward 1						
Model	p	z	β	θ	θ_{gl}	θ_i
Chain Error model	0.048 (0.02,0.09)	0.717 (0.59,0.83)	0.013 (0.007,0.021)	39.596 (38.08,41.13)	380.886 (378.81,383.19)	37.202 (36.31,38.11)
Chain Poisson model	0.049 (0.019,0.092)	0.705 (0.58,0.81)	0.012 (0.007,0.019)	40.243 (39.12,41.40)	380.558 (378.92,382.20)	37.202 (36.31,38.11)
Time Dependent Distances model	0.056 (0.025,0.10)	0.695 (0.58,0.80)	0.012 (0.007,0.019)	0.148 (0.144,0.152)	380.55 (378.92,382.18)	37.20 (36.31,38.10)

Table 4.1: Posterior mean estimates of the model parameters for each of the three models on ward 1, with 95% equitailed credible intervals.

Ward 2						
Model	p	z	β	θ	θ_{gl}	θ_i
Chain Error model	0.067 (0.028,0.12)	0.786 (0.68,0.84)	0.010 (0.006,0.014)	52.319 (50.31,54.34)	338.979 (337.08,341.16)	7.996 (6.33,9.84)
Chain Poisson model	0.033 (0.007,0.076)	0.813 (0.71,0.90)	0.013 (0.008,0.019)	61.699 (59.25,68.44)	212.021 (209.511,214.525)	8.009 (6.32,9.87)
Time Dependent Distances model	0.019 (0.002,0.052)	0.837 (0.75,0.91)	0.014 (0.009,0.019)	0.294 (0.27,0.33)	176.081 (174.71,177.34)	7.990 (6.37,9.81)

Table 4.2: Posterior mean estimates of the model parameters for each of the three models on ward 2, with 95% equitailed credible intervals.

patient per 100 days on the ward.

Figure 4.11 shows the estimated transmission network for the surgical ward, and for the paediatric ward. We can see that on the first ward 4 transmission events are attributed to patient T126 with high probability. This patient stayed on the ward for 71 days, which is 51 days longer than the patient who stayed for the second longest period, and T126 was observed to have a positive swab the day after their admission. However, the number of colonisations from patient T126 is still higher than expected given the estimate of the transmission parameter β. In the second ward we can see that 4 transmission events are attributed to patient T12, over the course of their 3 separate stays in the ward, with high probability. Combining each of this patient's stay times on the ward shows that they were admitted for a total of 103 days, which is 10 days longer than any other patient. However, the transmissibility of patient T12 is still far higher than expected given the estimate of the transmission parameter β.

4.8.2 Results from the Chain Poisson model for each ward

We used the Chain Poisson model to analyse the data from each ward from the hospital. Tables 4.1 and 4.2 give the posterior mean estimates of the parameters with 95% equitailed credible intervals for each ward. This model again assumes that sequences from patients who share a direct transmission event have their genetic distance drawn from a Poisson distribution with parameter θ, and that those from patients who are in distinct chains of transmission have their genetic distance drawn from a Poisson distribution with parameter θ_{gl}. The analysis for ward 1 suggests that genetic sequences from immediately linked patients are expected to differ by 40 $(39, 41)$ SNPs, and that sequences from unlinked patients are expected to differ by 381 $(379, 382)$ SNPs. The analysis for ward 2 suggests that genetic sequences from linked patients are expected to differ by 62 $(59, 68)$ SNPs, and that sequences from unlinked patients are expected to differ by 212 $(210, 215)$ SNPs.

Under the Chain Poisson model we estimate that the proportion of colonised admissions is slightly higher for ward 1 than for ward 2, with the estimate for patients admitted colonised to ward 1 being 5% $(1\%, 9\%)$ and the estimate for patients admitted colonised to ward 2 being 3% $(1\%, 8\%)$. It was estimated that the test sensitivity for the surgery ward was 71% $(58\%, 81\%)$, and the test sensitivity for the paediatric ward was 81% $(71\%, 90\%)$. The estimates for the transmission parameter, β, under this model are similar for each ward, with an estimated transmission rate equivalent

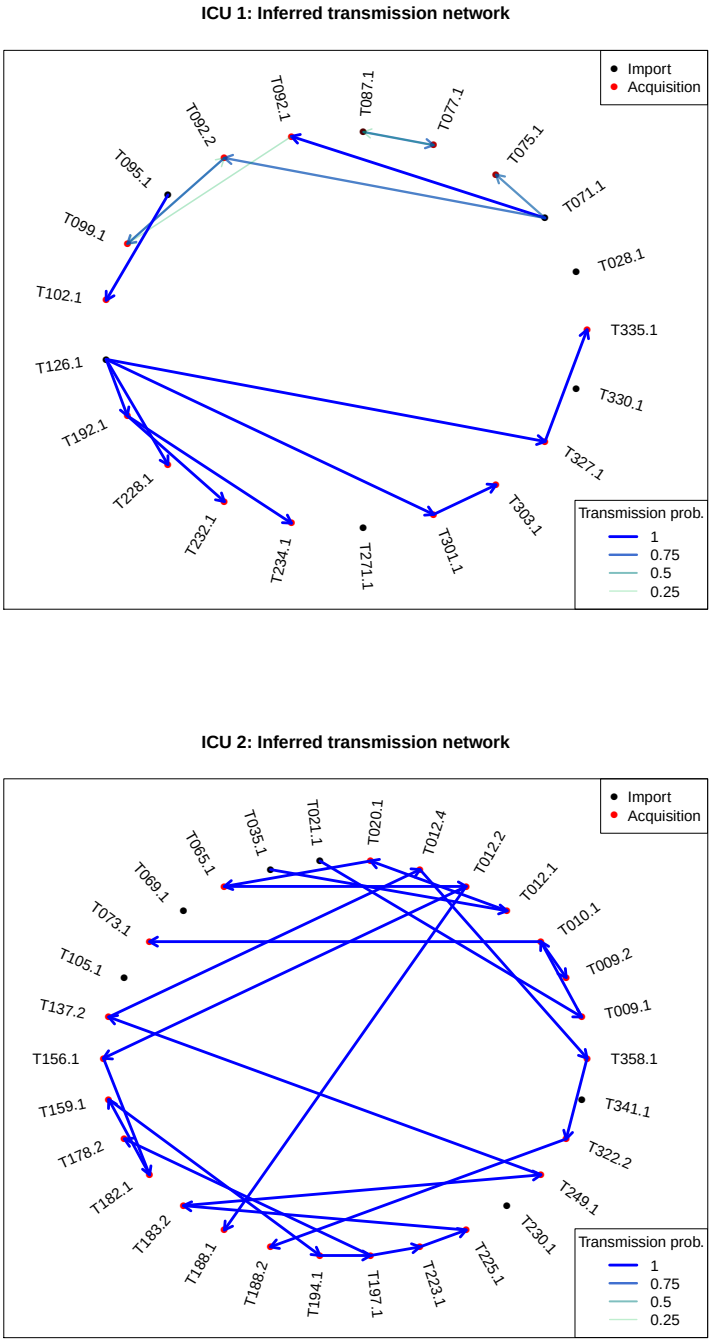


Figure 4.11: The inferred transmission trees for each ward of the Thai data given by the MCMC algorithm output for the Chain Error model. The colour of the nodes represents the probability that the patient was an importation. The arrows represent inferred transmission events between patients. The colour of these represents how likely they were to have taken place.

to 1.2 $(0.7, 1.9)$ patients colonised by each infectious patient per 100 days on ward 1, and 1.3 $(0.8, 1.9)$ patients colonised by each infectious patient per 100 days on ward 2.

Figure 4.12 shows the estimated transmission network for each ward under the Chain Poisson model. Again we estimate that the patients who were present longest on each ward (T126 on ward 1 and T12 on ward 2) were the source of a disproportionally large number of acquisitions, with T126 colonising 4 patients over 71 days with high probability, and T12 colonising 6 patients over 103 days with high probability.

4.8.3 Results from the Time Dependent Distances model for each ward

The Time Dependent Distances model was used to analyse the data from each ward. Tables 4.1 on page 102 and 4.2 on page 102 give the posterior mean estimates of the parameters with 95% equitailed credible intervals for each ward. This model uses the parameter θ in a different way to the other two models, as genetic distances between sequences from two patients who share a transmission event are drawn from a Poisson distribution with parameter $t_{i,j}\theta$ where $t_{i,j}$ is a measure of the time between the sampling of the two sequence, so it is harder to interpret. However, the parameter θ_{gl} for the Poisson distribution from which genetic distances are drawn for sequences from patients in independent chains of transmission has the same meaning. It is estimated that the genetic sequences between two patients in independent chains are expected to differ by 381 $(379, 382)$ SNPs in ward 1, and by 176 $(175, 177)$ SNPs in ward 2.

Under this model we estimate that 6% $(3\%, 10\%)$ of patients admitted to ward 1 were already colonised, and only 2% $(0.2\%, 5\%)$ of patients admitted to ward 2 were already colonised. It was estimated that the test sensitivity for the surgery ward was 70% $(58\%, 80\%)$, and the test sensitivity for the paediatric ward was 84% $(75\%, 91\%)$. Under this Time Dependent Distances model the estimate for the transmission parameter, β, on ward 1 is equivalent to 1.2 $(0.7, 1.9)$ patients colonised by each infectious patient per 100 days on the ward. On ward 2 the estimate for β is equivalent to 1.4 $(0.9, 1.9)$ patients colonised by each infectious patient per 100 days on the ward.

Figure 4.13 shows the estimated transmission network under the Time Dependent Distances model for each ward. As with the other two models we estimate that the patients who stayed for the longest time on each ward were the source of a large

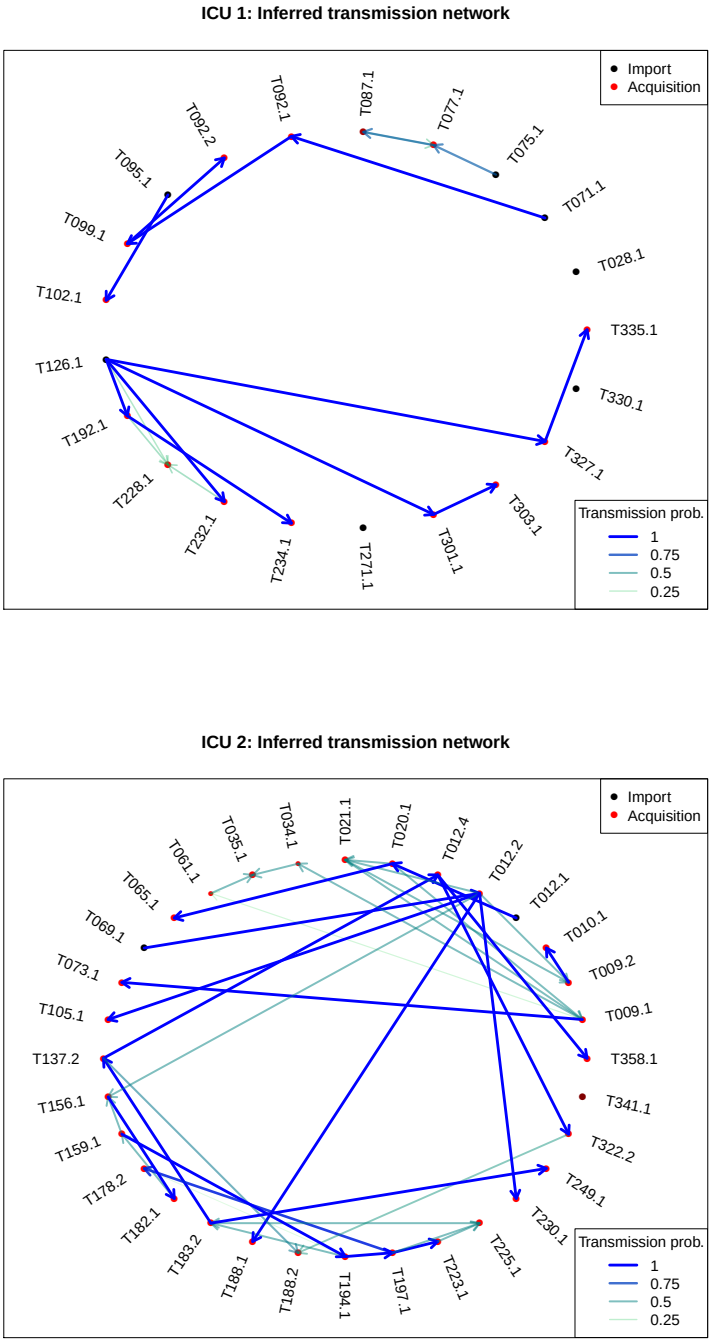


Figure 4.12: The inferred transmission trees for each ward of the Thai data given by the MCMC algorithm output for the Chain Poisson model. The colour of the nodes represents the probability that the patient was an importation. The arrows represent inferred transmission events between patients. The colour of these represents how likely they were to have taken place.

number of acquisitions with high probability, with T126 on ward 1 estimated to be the source of colonisation for 5 patients over 71 days, and T12 on ward 2 estimated to be the source of colonisation for 6 patients over 103 days.

4.8.4 Comparison of results from each model

For ward 1, each of our three models give similar estimates for each of the comparable parameters. The importation parameter p has posterior mean 0.05 or 0.06, the sensitivity has posterior mean 0.70 to 0.72, and the transmission rate has posterior mean 0.012 or 0.013 for the three models. We can compare these estimates to those obtained under the Worby et al. models, as the basic epidemic model is comparable to our models. Analysis of the ward 1 data under the Worby et al. models also estimates the importation parameter $p = 0.06\ (0.03, 0.11)$, and the estimate of the sensitivity $z = 0.73\ (0.62, 0.84)$ or $0.75\ (0.64, 0.86)$ (for the two different variations of the Worby et al. model) is well within our credible interval. The transmission parameter is estimated to be $\beta = 0.01\ (0.005, 0.02)$ which is again within our credible interval. The estimates for epidemiological parameters of all the models are very similar.

However, we observe more variation in the estimates of the parameters for ward 2. The Chain Error model gives the posterior mean of the importation parameter as $0.07\ (0.05, 0.12)$, which is similar to what was estimated by all three models for ward 1. The other two models estimate p to be much lower than this, with posterior mean either $0.03\ (0.01, 0.08)$ or $0.02\ (0.002, 0.05)$. The figures show that the Chain Error model infers 6 importations with high probability, whereas the Chain Poisson model only infers 2, and the Time Dependent Distances model infers only 1. Importantly, all three posterior means fall within the 95% equitailed credible intervals estimated by the models. Interestingly, all of our posterior means are lower than those from the Worby et al. models which are $0.08\ (0.04, 0.14)$ and $0.16\ (0.09, 0.25)$, and fall outside of each other's credible intervals. The patients which are being inferred as importations by the Worby et al. models but not in our models tend to be those who are observed to be positive for MRSA within a couple of days of their arrival on the ward. If such patients have small genetic distances with other patients on the ward our model more often classifies them as having acquired the disease shortly after arrival, whereas the Worby et al. models tend to classify them as an importation.

The estimates for the transmission parameter under each of our three models are more similar, with the Chain Error model posterior mean of $\beta = 0.010\ (0.006, 0.014)$,

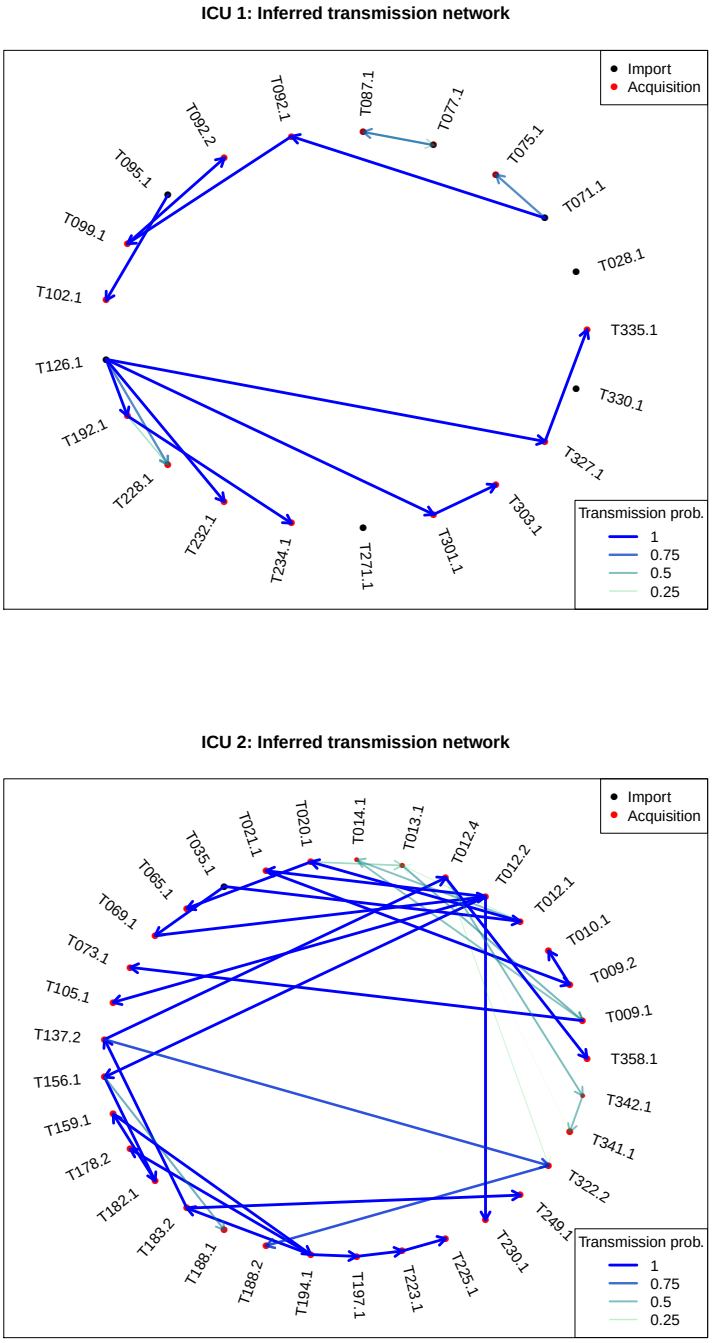


Figure 4.13: The inferred transmission trees for ward 2 of the Thai data given by the MCMC algorithm output for the Time Dependent Distances model. The colour of the nodes represents the probability that the patient was an importation. The arrows represent inferred transmission events between patients. The colour of these represents how likely they were to have taken place.

the Chain Poisson model posterior mean of $\beta = 0.013$ $(0.008, 0.019)$ and the Time Dependent Distances model posterior mean of $\beta = 0.014$ $(0.009, 0.019)$. The posterior means from the Worby et al. models are $\beta = 0.0077$ $(0.004, 0.01)$ and $\beta = 0.010$ $(0.006, 0.015)$. It follows that the Worby et al. models would estimate a lower transmission rate than our models as they estimated a higher importation rate than we did. Our models and the Worby et al. models produce posterior means for the test sensitivity on ward 2 that are all in a similar region. From our models we get $z = 0.79$ $(0.68, 0.84)$, $z = 0.81$ $(0.71, 0.90)$ and $z = 0.84$ $(0.75, 0.91)$, and from the Worby et al. models we get $z = 0.83$ $(0.75, 0.90)$ and $z = 0.85$ $(0.77, 0.91)$.

We can compare the global genetic parameter and the within-host genetic parameter for our three models, although the genetic parameters in the Worby et al. model have different meaning so can not be compared. Analysis of ward 1 under each of the three models gives the same posterior mean of $\theta_{gl} = 381$ and $\theta_i = 37$. The three models also give the same posterior mean of $\theta_i = 8$ for ward 2. There is much more variation in the estimates for the global genetic parameter from analysis of ward 2 under the three models. The Chain Error model gives a posterior mean of $\theta_{gl} = 339$ $(337, 341)$ whereas the Chain Poisson model gives a posterior mean of $\theta_{gl} = 212$ $(210, 215)$ and the Time Dependent Distances model gives a posterior mean of $\theta_{gl} = 176$ $(175, 177)$. Similarly, comparing the posterior mean for θ for the Chain Error model and the Chain Poisson model, which give it the same meaning, we see that the posterior means are equal for ward 1 ($\theta = 40$), whereas for ward 2 we get posterior mean $\theta = 52$ $(50, 54)$ under the Chain Error model, and $\theta = 62$ $(59, 68)$ under the Chain Poisson model.

Figures 4.11, 4.12 and 4.13 allow us to compare the estimated transmission tree for each ward under the three different models. For the first ward we can clearly see that the estimated transmission trees under the Chain Poisson model and the Time Dependent model are almost identical, and the transmission tree under the Chain Error model is also similar. The Chain Error model estimates a chain of transmission that goes directly from T071.1 to T092.2 and on to T099.1, whereas the other two models have this transmission chain starting with T071.1 to T092.1, then to T099.1 and finally to T092.2. There is also a slight difference between the models in the exact route of transmission estimates for the group of patients colonised directly and indirectly by patient T126.1, however all of the models infer this patient as the source of colonisation for a much larger number of other patients (posterior mean outdegree for patient T126.1 was 6.1, 6.3 and 7.2 for each of the models) than the transmission parameter

estimate would suggest. For this ward we estimated a posterior mean outdegree of 2.0 from each patient in the tree under each of the three models. We note that the transmission trees estimated under the Worby et al. models for this ward have a very similar shape, although ours appear to have greater resolution.

For the second ward there is slightly more variation between our models in the estimation of the transmission tree, although again we can immediately see from the figures that the broad shape of the trees remains the same. Here again we have all three models estimating one patient (T12) to be the source of colonisation for a larger number of other patients than we would expect given the estimation of the transmission parameter. The posterior mean outdegree for patient T12 (including each of the patient's three admissions to the ward) is 5.7 under the Chain Error model, 6.9 under the Chain Poisson model, and 7.9 under the Time Dependent Distances model. The posterior mean outdegree per patient for each model was 1.8 for both the Chain Poisson model and the Time Dependent Distances model, and 1.5 for the Chain Error model, which reflects the slightly lower estimate of the transmission parameter for this model compared to the other two. Compared to the Worby et al. models estimated trees, ours show a greater resolution and more transmission events estimated at higher probability levels. The Worby et al. models also estimate that patient T12 was the source of more colonisations than anyone else.

4.9 Model assessment

In section 5.8 the results of analysing MRSA data from two separate wards in a Thai hospital under our three different models were presented. The results from the different models were compared in section 4.8.4 and it was found that in some cases the results were distinctly different under different models. Therefore, model assessment is required in order to determine the goodness-of-fit of each of the models and to assess whether one model can be distinguished which fits better than the rest.

4.9.1 Epidemic model assessment

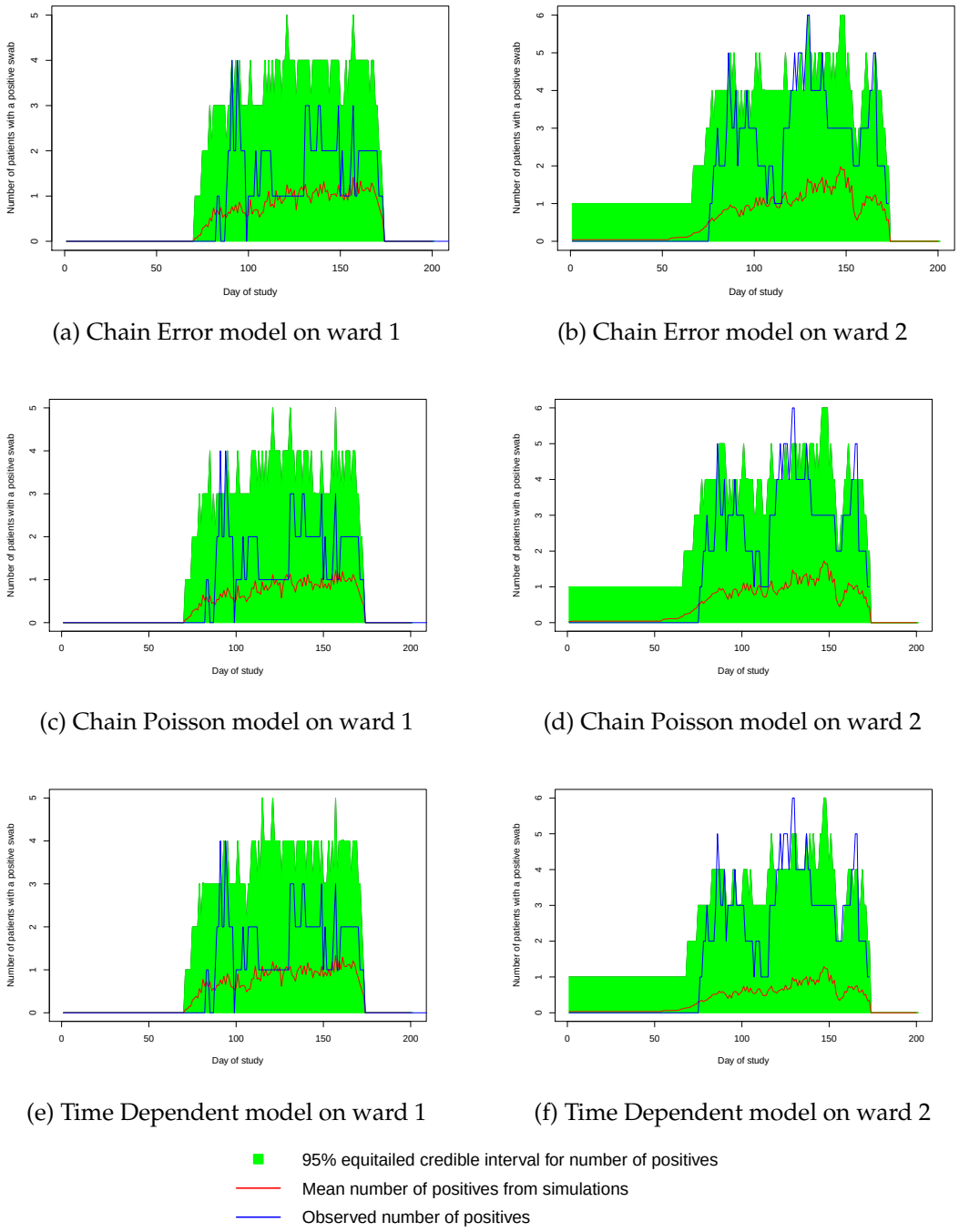
Posterior predictive checks were used in order to check the goodness-of-fit of the three models to the data from each of the wards. The number of patients ever to have a positive swab taken over the course of the epidemic was used as a summary statistic for the epidemic data. The number of patients present on the ward each day with a positive swab taken on that, or a previous, day was taken as a summary set of

	Chain Error model	Chain Poisson model	Time Dependent model
Ward 1	(7,53)	(2,33)	(3,33)
Ward 2	(5,28)	(1,32)	(1,27)

Table 4.3: 95% highest density regions for the number of patients to have a positive swab for each ward under each of our models. The observed value for ward 1 was 22 and for ward 2 was 30. Red HDRs indicate that the observed value falls outside the region, and green HDRs indicate that the observed value falls inside.

statistics. We refer to these as positive patients per day. For each model 1000 sets of data were simulated, with fixed patient admission and discharge times from the data (these do not form part of our model framework), using values of the parameters which were taken from the posterior densities given by the output of the MCMC algorithm. The values of the summary statistics were recorded for each simulation. The observed number of patients with a positive swab on the real ward, n_{sw}, was compared to the distribution of the number of patients with a positive swab in the simulated wards, $\widetilde{n}_{sw}$. Table 4.3 gives the 95% highest density region for the posterior predictive distributions for each ward under each model. The red HDRs indicate that the posterior predictive p-value, $\mathrm{P}(n_{sw} > \widetilde{n}_{sw})$, is extreme and falls outside the HDR. In these cases there is evidence against the fit of the model. The posterior predictive p-value for both the Chain Error model and the Time Dependent model indicated a lack of fit to the data for ward 2, whereas there was no evidence against the fit of any of the models for ward 1, or against the Chain Poisson model for ward 2.

Figure 4.14 displays the results of recording the positive patients per day for the 1000 simulations. The observed positive patients present on day i, n^i_{sw}, was compared to the distribution of the positive patients present on the same day in the simulated wards, $\widetilde{n}^i_{sw}$. The green area in figure 4.14 gives the 95% highest density region for the posterior predictive distributions for each ward under each model and the blue line gives the observed values from the dataset. The posterior predictive p-value, $\mathrm{P}(n^i_{sw} > \widetilde{n}^i_{sw})$, for each day of the study can be examined. If the blue line departs from within the green area then the p-value is extreme. It can be noted that for ward 1 only the Chain Error model displays no extreme p-values, although the other two models only have one out of 173 study days. For ward 2 each model displays some extreme p-values, with 10 out of 173 study days under both the Chain Error model and the Time Dependent model, and only 5 under the Chain Poisson model. These results for ward 2 agree with the posterior predictive check of the total number of patients with a positive swab which suggested that the Chain Poisson model was the



(a) Chain Error model on ward 1

(b) Chain Error model on ward 2

(c) Chain Poisson model on ward 1

(d) Chain Poisson model on ward 2

(e) Time Dependent model on ward 1

(f) Time Dependent model on ward 2

Figure 4.14: Posterior predictive checking for the number of patients present on the ward who have had a positive swab on that day or a previous day. The green area gives the 95% highest density region of the posterior predictive distribution, the red line shows the mean of the distribution and the blue line shows the observed values from the data.

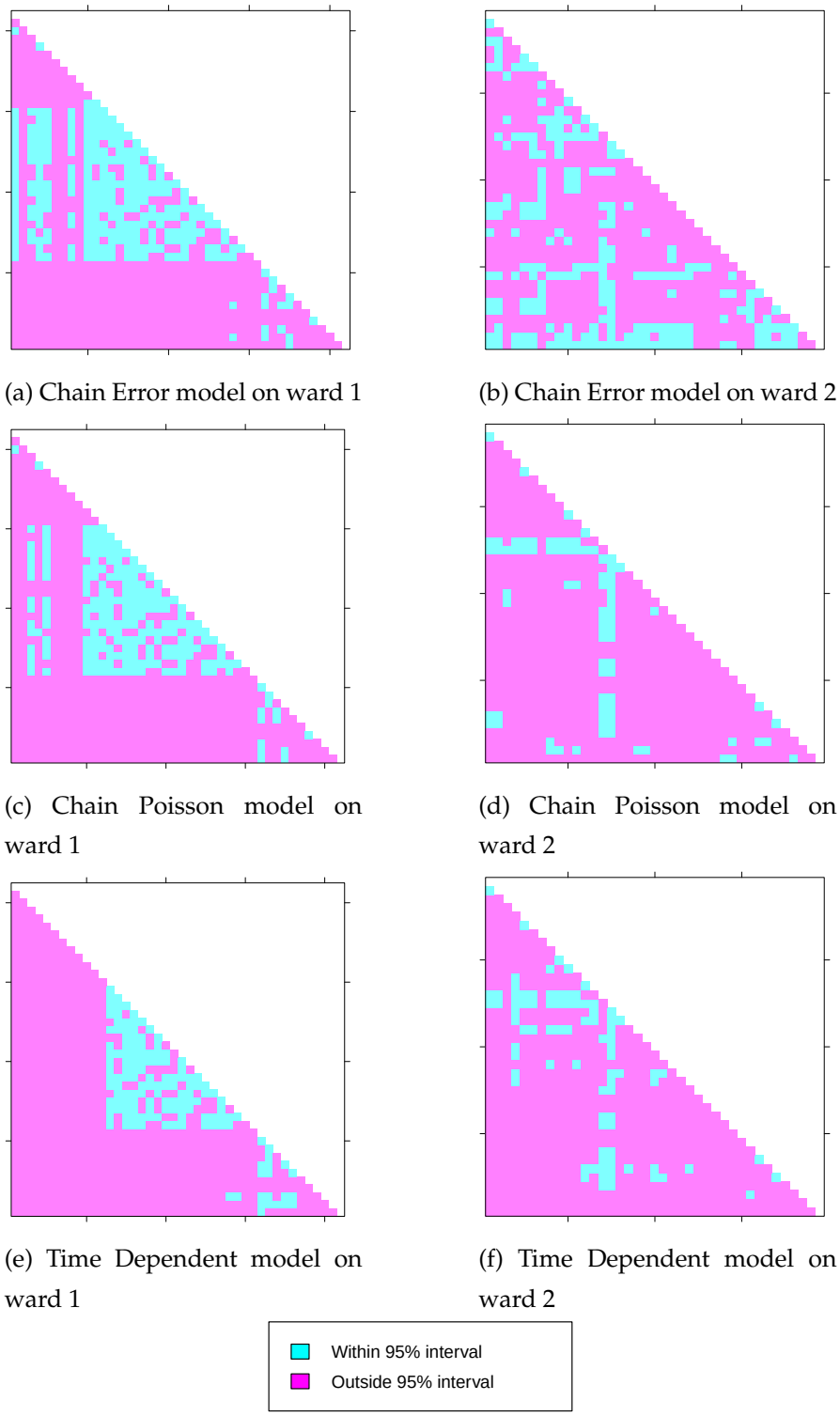
best fit. For ward 1 this assessment of the positive swabs per day has given more information than the assessment of the total number of patients with a positive swab and has indicated that the Chain Error model is the best fit.

So far only the epidemic parts of the models have been assessed for goodness-of-fit. In section 4.10 the novel method of genetic model assessment which was introduced in section 3.3 is applied.

4.10 Genetic model assessment

In order to assess the fit of each different model to the genetic data a total of 1000 genetic distance matrices, $\widetilde{\mathbf{\Psi}}$, were simulated using values for the genetic parameters, the times of colonisation and the sources of colonisation drawn from the posterior densities given by the output of the MCMC algorithm. The observed genetic distance between two sequences from patients who were both sampled during the course of the epidemic is given by $\Psi_{i,j}$, so the posterior predictive p-value can be defined as $\mathrm{P}(\Psi_{i,j} < \widetilde{\Psi}_{i,j})$. An extreme p-value that falls outside the 95% highest density region indicates that that particular genetic distance, $\Psi_{i,j}$, was poorly fitted by the model used. In order to assess the model for the genetic distance matrix as a whole, the percentage of these genetic distances which give extreme p-values is recorded, and also a binary matrix is plotted to visualise this percentage across the genetic distance matrix. Figure 4.15 displays these binary matrices for each ward under each model. It is immediately clear that none of the genetic models do well on the data for these wards. The Chain Error model has a posterior predictive matrix score of 31.36% for ward 1 and 29.15% for ward 2. The Chain Poisson model has a posterior predictive matrix score of 24.85% for ward 1 and 10.80% for ward 2. The Time Dependent model has a posterior predictive matrix score of 16.49% for ward 1 and 11.47% for ward 2. The blocks in the centres of the matrices representing ward 1 for each model where the majority of the distances are well fitted correspond to a set of within-host distances from patient T126 who had 19 genetic sequences sampled at regular intervals throughout their long stay. These blocks of blue show that the within-host genetic data are well captured by the model.

In order to investigate the poor fit of these models to the data the matrix of observed genetic distances was compared to the mode matrix of genetic distances from the set of simulations using histograms such as those in figure 4.16. The red bars represent the observed genetic distances for each ward, and the green bars represent the mode



(a) Chain Error model on ward 1

(b) Chain Error model on ward 2

(c) Chain Poisson model on ward 1

(d) Chain Poisson model on ward 2

(e) Time Dependent model on ward 1

(f) Time Dependent model on ward 2

Figure 4.15: Posterior predictive checking for the genetic distance matrices from each ward under each model. Blue cells indicate that the observed genetic distance falls within the 95% highest density region given by the posterior predictive distribution and pink cells indicate that the observed genetic distance falls outside that HDR. The layout of the matrices corresponds to the layout of the original genetic distance matrices, so the x-axis, left to right, runs from sequence 1 to sequence n_{seqs}, whereas the y-axis, from bottom to top, runs backwards from sequence n_{seqs} to sequence 1.

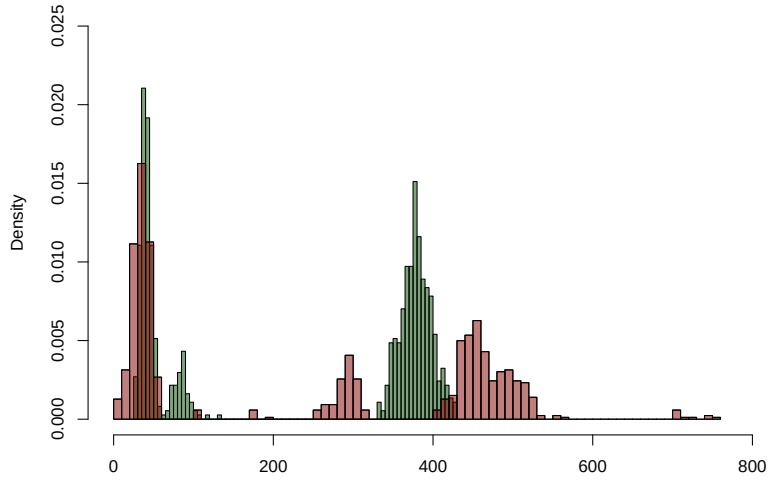
set (there were two mode sets for ward 2) from the simulations produced using the Chain Poisson model. By 'mode matrix' we mean the most commonly simulated matrix among the 1000 simulations. There will not necessarily always be a mode matrix because they may all be different, but because the transmission tree is fixed it is fairly likely. For each of our posterior predictive sets of simulations we found that a mode matrix existed. This figure shows that the sets of observed genetic distances for both wards are trimodal, and that the simulated genetic distance matrices are failing to imitate this because of the way in which genetic distances are assumed to be drawn in the model. The other two models experience the same problem; the variances of the distributions for drawing the genetic distances, which are Poisson distributions, appear to be too small. Therefore, an alternative distribution to the Poisson distribution needs to be considered. In section 4.11 the Geometric distribution and Negative Binomial distribution are explored as alternatives.

4.11 Alternative distributions for the basis of the genetic models

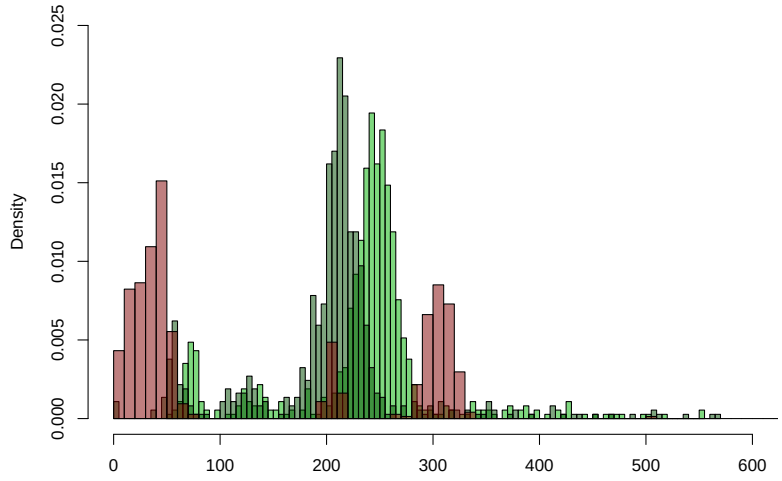
The model assessment carried out in section 4.10 indicated that the three models introduced in chapter 2 performed poorly in terms of fit for the genetic MRSA data from the two Thai hospital wards, despite fitting the epidemic data well in most cases. The problem appeared to be the restriction of the variance of the Poisson distribution, so in this section we will adapt the models by replacing the Poisson distribution with a Geometric distribution, which has a larger variance, and with a Negative Binomial distribution, which allows the mean and variance to be specified independently of one another.

4.11.1 Geometric distributions for the genetic distance model

The Chain Error and Chain Poisson models introduced in chapter 2 use the Poisson distribution with parameters θ, θ_{gl} and θ_i to model genetic distances between pairs of sequences from patients who are directly next to each other in a chain, sequences from patients who are in separate chains, and within-host genetic distances respectively. The Time Dependent model uses the same Poisson distributions for sequences from patients in separate chains and within-host distances, and a Poisson distribution with parameter $t_{i,j}\theta$ for sequences from patients directly next to each other in a chain, where $t_{i,j}$ is a measure of the time separating the sampling of the two sequences.



(a) Ward 1



(b) Ward 2

Figure 4.16: Histograms of the observed genetic distances from each ward, in red, and the mode matrix of genetic distances from 1000 posterior predictive simulations, in green.

These Poisson distributions can be replaced with Geometric distributions in order to allow for a larger variance. Geometric distributions have been used to model genetic distances from sequences which are sampled over small time intervals such as those in this study as the genetic distances are not expected to be large. The Geometric versions of the three models are described here.

4.11.1.1 Geometric Chain Error model

The Geometric Chain Error model assumes that genetic distances between pairs of sequences, i, j, from patients who are separated by $k \leq 1$ or $k = \infty$ transmission events are drawn from the following distributions:

$$\mathrm{P}(\Psi_{i,j} = x) = \begin{cases} (1-\varphi_i)^x \varphi_i & \text{if } k = 0 \\ (1-\varphi)^x \varphi & \text{if } k = 1 \\ (1-\varphi_{gl})^x \varphi_{gl} & \text{if } k = \infty. \end{cases} \tag{4.11.1}$$

As in the original Chain Error model, the conditional probability distribution for genetic distances for pairs of sequences which are in the same chain but separated by more than one transmission event is:

$$\mathrm{P}(\Psi_{i,j} = x | D_{i,j}) = \frac{(k\gamma)^{|D_{i,j}-x|}}{|D_{i,j}-x|! \left(\sum_{l=0}^{D_{i,j}} (k\gamma)^l / l!\right)} \left(\frac{1}{2}\right)^{\mathbb{1}_{\{x \neq D_{i,j}\}}} \mathbb{1}_{\{x \leq 2D_{i,j}\}} \quad \text{if } k > 1, \tag{4.11.2}$$

where $D_{i,j}$ is the sum of the consecutive distances between the isolates from patients that compose the transmission chain between H_i and H_j which are the host patients of sequences i and j.

4.11.1.2 Geometric Chain Poisson model

The Geometric Chain Poisson model assumes that the genetic distances between pairs of sequences, i, j, from patients who are separated by $k \leq 1$ or $k = \infty$ transmission events are drawn from the following distributions:

$$\mathrm{P}(\Psi_{i,j} = x) = \begin{cases} (1-\varphi_i)^x \varphi_i & \text{if } k = 0 \\ (1-\varphi)^x \varphi & \text{if } k = 1 \\ (1-\varphi_{gl})^x \varphi_{gl} & \text{if } k = \infty. \end{cases} \tag{4.11.3}$$

As in the original Chain Poisson model the conditional probability distribution for genetic distances for pairs of sequences which are in the same chain but separated by more than one transmission event is:

$$\mathrm{P}(\Psi_{i,j} = x | D_{i,j}) = (D_{i,j}^x / x!) \exp(-D_{i,j}) \quad \text{if } k > 1, \tag{4.11.4}$$

where $D_{i,j}$ is the sum of the consecutive distances between the sequences taken from patients that compose the transmission chain between H_i and H_j .

4.11.1.3 Geometric Time Dependent model

The Geometric Time Dependent model assumes that the genetic distances between pairs of sequences, i, j, from patients who are separated by $k \leq 1$ or $k = \infty$ transmission events are drawn from the following distributions:

$$\mathrm{P}(\Psi_{i,j} = x) = \begin{cases} (1-\varphi_i)^x \varphi_i & \text{if } k = 0 \\ -\exp(-t_{i,j}\varphi x)(1-\exp(-t_{i,j}\varphi)) & \text{if } k = 1 \\ (1-\varphi_{gl})^x \varphi_{gl} & \text{if } k = \infty, \end{cases} \tag{4.11.5}$$

where $t_{i,j}$ is a measure of the time between the sampling of the two sequences. We will use $t_{i,j} = |t_j^s - t_i^s|$, where t_i^s is the sampling time of sequence i. As before, the conditional probability distribution for genetic distances for pairs of sequences which are in the same chain but separated by more than one transmission event is:

$$\mathrm{P}(\Psi_{i,j} = x | D_{i,j}) = (D_{i,j}^x / x!) \exp(-D_{i,j}) \quad \text{if } k > 1, \tag{4.11.6}$$

where $D_{i,j}$ is the sum of the consecutive distances between the sequences from patients that compose the transmission chain between H_i and H_j.

A Geometric distribution with parameter $1 - \exp(-t_{i,j}\varphi)$ is used for the distances between sequences from patients who share a transmission event because the idea of this version of the model is that these distances are influenced by the time between the sampling of the sequences, but the parameter $t_{i,j}\varphi$ is unsuitable for the Geometric distribution as it can be bigger than 1. It is known that the Exponential distribution is a limiting form of the Geometric distribution when the parameter is small, as $t_{i,j}\varphi$ will be. To derive our parameter, define $Y \sim \mathrm{Exp}(\lambda)$, where $\lambda = t_{i,j}\varphi$. Define $Z = \lfloor Y \rfloor$, and n as a non-negative integer. Therefore,

$$\mathrm{P}(Z \geq n) = \mathrm{P}(\lfloor Y \rfloor \geq n) = \mathrm{P}(Y \geq n) = \mathrm{e}^{-\lambda n}.$$

If $\mathrm{P}(Z = n)$ is considered as $\mathrm{P}(Z \geq n) - \mathrm{P}(Z \geq n+1)$ then clearly

$$\mathrm{P}(Z = n) = \mathrm{e}^{-\lambda n} - \mathrm{e}^{-\lambda(n+1)} = \mathrm{e}^{-\lambda n}(1 - \mathrm{e}^{-\lambda}),$$

which implies that Z has a Geometric distribution with parameter $1 - \mathrm{e}^{-\lambda}$.

4.11.2 Negative Binomial distributions for the genetic distance model

Another option for replacing the Poisson distributions in the models for the genetic distances is the Negative Binomial distribution. This distribution allows for the mean and variance to be specified independently, and the variance must always be larger than the mean which makes it suitable for use here where the data seem unsuitable for the Poisson distribution which has variance equal to the mean. Therefore, for ease, the mean, μ, and variance, σ^2, parameterisation of the Negative Binomial distribution is used. As the within-host distances do not affect the structure of the tree, and were found to be modelled well with a single parameter, within-host distances are assumed to be drawn from a Geometric distribution in order not to introduce unnecessary parameters. We present the Negative Binomial versions of the Chain Error model and Chain Poisson model here. We judged that adapting the Time Dependent Distances model with the Negative Binomial distribution would be over-complex.

4.11.2.1 Negative Binomial Chain Error model

The Negative Binomial Chain Error model assumes that the genetic distances between a pair of sequences, i, j, from patients who are separated by $k \leq 1$ or $k = infty$ transmission events is drawn from one of the following distributions:

$$\mathrm{P}(\Psi_{i,j} = x) = \begin{cases} (1-\varphi_i)^x \varphi_i & \text{if } k = 0 \\ \binom{x - 1 + \mu^2(\sigma^2-\mu)^{-1}}{x} \left(\frac{\sigma^2-\mu}{\sigma^2}\right)^x \left(\frac{\mu}{\sigma^2}\right)^{\mu^2(\sigma^2-\mu)^{-1}} & \text{if } k = 1 \\ \binom{x - 1 + \mu_{gl}^2(\sigma_{gl}^2-\mu_{gl})^{-1}}{x} \left(\frac{\sigma_{gl}^2-\mu_{gl}}{\sigma_{gl}^2}\right)^x \left(\frac{\mu_{gl}}{\sigma_{gl}^2}\right)^{\mu_{gl}^2(\sigma_{gl}^2-\mu_{gl})^{-1}} & \text{if } k = \infty. \end{cases} \tag{4.11.7}$$

As before the conditional probability distribution for genetic distances for pairs of sequences which are in the same chain but separated by more than one transmission event is:

$$\mathrm{P}(\Psi_{i,j} = x | D_{i,j}) = \frac{(k\gamma)^{|D_{i,j}-x|}}{|D_{i,j}-x|! \left(\sum_{l=0}^{D_{i,j}} (k\gamma)^l / l!\right)} \left(\frac{1}{2}\right)^{\mathbb{1}_{\{x \neq D_{i,j}\}}} \mathbb{1}_{\{x \leq 2D_{i,j}\}} \quad \text{if } k > 1, \tag{4.11.8}$$

where $D_{i,j}$ is the sum of the consecutive distances between the sequences from patients that compose the transmission chain between H_i and H_j.

4.11.2.2 Negative Binomial Chain Poisson model

The Negative Binomial Chain Poisson model assumes that the genetic distances between a pair of sequences, i, j, from patients who are separated by $k \leq 1$ or $k = \infty$ transmission events is drawn from one of the following distributions:

$$\mathrm{P}(\Psi_{i,j} = x) = \begin{cases} (1-\varphi_i)^x \varphi_i & \text{if } k = 0 \\ \binom{x - 1 + \mu^2(\sigma^2 - \mu)^{-1}}{x} \left(\frac{\sigma^2 - \mu}{\sigma^2}\right)^x \left(\frac{\mu}{\sigma^2}\right)^{\mu^2(\sigma^2-\mu)^{-1}} & \text{if } k = 1 \\ \binom{x - 1 + \mu_{gl}^2(\sigma_{gl}^2 - \mu_{gl})^{-1}}{x} \left(\frac{\sigma_{gl}^2 - \mu_{gl}}{\sigma_{gl}^2}\right)^x \left(\frac{\mu_{gl}}{\sigma_{gl}^2}\right)^{\mu_{gl}^2(\sigma_{gl}^2-\mu_{gl})^{-1}} & \text{if } k = \infty. \end{cases} \tag{4.11.9}$$

As before the conditional probability distribution for genetic distances for pairs of sequences which are in the same chain but separated by more than one transmission event is:

$$\mathrm{P}(\Psi_{i,j} = x | D_{i,j}) = (D_{i,j}^x / x!) \exp(-D_{i,j}) \quad \text{if } k > 1, \tag{4.11.10}$$

where $D_{i,j}$ is the sum of the consecutive distances between the sequences from patients that compose the transmission chain between H_i and H_j.

4.12 Analysing the Thai hospital data with the Geometric and Negative Binomial models

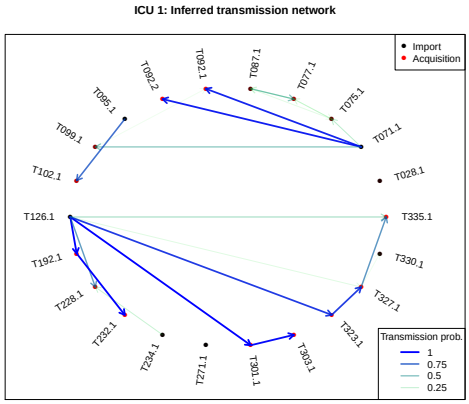
In section 4.11 the models which were introduced in chapter 2 were adjusted to allow for the distribution for the genetic distances to have greater variance than in the original models which used Poisson distributions. The results from analysing the data from the two wards in the Thai hospital under these new models which use the Geometric and Negative Binomial distributions are presented here. The same data-augmented MCMC algorithm (see 4.5.2) was used. For each model on each ward the algorithm was run for 100,000 iterations, with 10 augmented data steps during each iteration. For the Geometric versions of the models the parameters p, z, φ, φ_i and φ_{gl} were given Beta$(1,1)$ prior distributions and the parameters β and γ (in the Chain Error model) were given improper uniform prior distributions on $(0, \infty)$. For the Negative Binomial versions of the models the parameters p, z and φ_i were given Beta$(1,1)$ prior distributions and the parameters β, γ (in the Chain Error model), μ, σ, μ_{gl} and σ_{gl} were given improper uniform prior distributions on $(0, \infty)$. All of the genetics parameters for both new versions of the models were updated using Gaussian random-walk Metropolis-Hastings steps in the MCMC algorithm. The step size

of the random walk varied according to a Normal distribution with mean 0 and variance σ^2 and the acceptance rate was checked every 1000 iterations in order to adjust the variance to maintain an acceptance rate between 0.2 and 0.6.

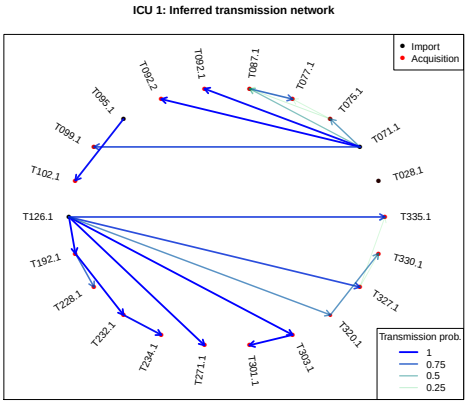
4.12.1 Results for the adjusted models on ward 1

Tables 4.4 and 4.5 give the posterior mean estimates for the values of the model parameters for ward 1 with 95% equitailed credible intervals. Under the Negative Binomial versions of the Chain Error and Chain Poisson models we get similar estimates for the importation probability p. The NB Chain Error model estimated that 4% of patients admitted to the ward 1 were already colonised and the NB Chain Poisson model estimates that 3% were already colonised. The Geometric versions of the model show slightly more variation, with the posterior mean for p ranging from 0.02 to 0.06. Interestingly, we can see from figure 4.17 that all of these models infer 4 importations of the disease with high probability, except the Geometric Chain Error model which infers 8 importations, corresponding to its higher posterior mean of $p = 0.06$. All of the models give similar estimates for the test sensitivity, z, for which the mean estimates range from 0.68 to 0.72. The posterior means for the transmission parameter, β, range between 0.013 and 0.017 under the Geometric models, although both Negative Binomial models estimate posterior mean $\beta = 0.016$. These different posterior means for β all fall well within each other's credible intervals.

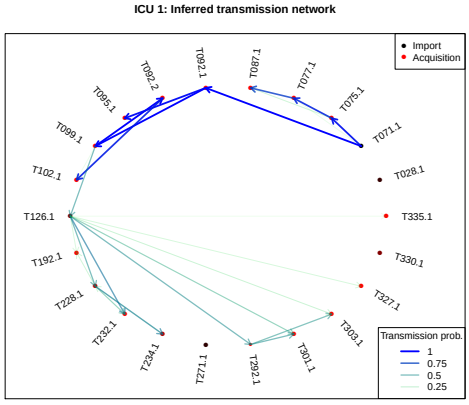
The only directly comparable genetic parameter between all five of the models is the within-host genetic parameter φ_i. This is estimated as having posterior mean 0.026 by all of the models and the expected genetic distance between within-host sequences is 38. The estimated expected genetic distance between two patients who share a transmission event varies significantly between the different models. Under the Geometric Chain Error model the expected genetic distance is estimated as 48, under the Geometric Chain Poisson model it is estimated as 94, and under the Negative Binomial versions of these models it is estimated as 121 and 131 respectively. The Negative Binomial models also estimate the variance of distribution for these genetic distances to be much higher than the variances under the Geometric models. The estimated expected genetic distance between two patients who are in different transmission chains does not vary so much under the different models. The Geometric models all give expected genetic distances of between 367 and 369, and the Negative Binomial models give the expected genetic distances as 384 or 386.



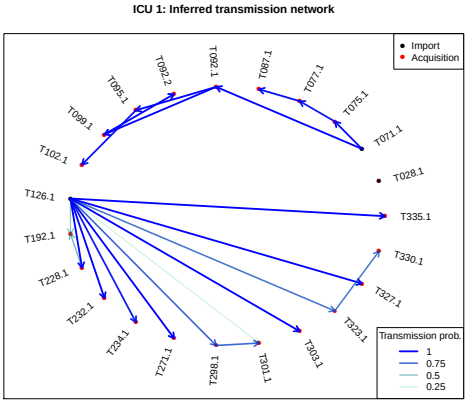
(a) Geometric Chain Error model



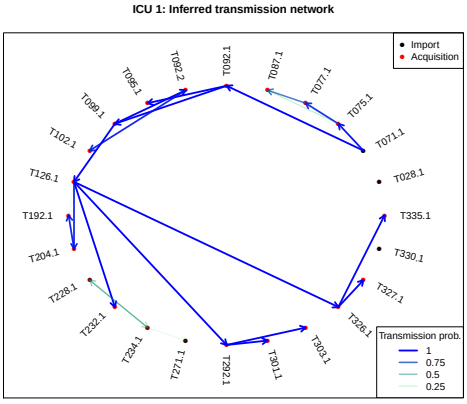
(b) Negative Binomial Chain Error model



(c) Geometric Chain Poisson model



(d) Negative Binomial Chain Poisson model



(e) Geometric Time Dependent model

Figure 4.17: Posterior transmission trees for ward 1 of the Thai data under each of the Geometric and Negative Binomial models. The colour of the nodes represents the probability that the patient was an import. The arrows represent inferred transmission events between patients. The colour of these represents how likely these were to have taken place.

Ward 1						
Model	p	z	β	φ	φ_{gl}	φ_i
Chain Error model	0.06 (0.024,0.112)	0.684 (0.555,0.802)	0.015 (0.008,0.025)	0.022 (0.014,0.031)	0.003 (0.002,0.003)	0.026 (0.023,0.03)
Chain Poisson model	0.034 (0.01,0.071)	0.688 (0.567,0.80)	0.017 (0.009,0.027)	0.011 (0.006,0.02)	0.003 (0.003,0.003)	0.026 (0.023,0.03)
Time Dependent Distances model	0.02 (0.004,0.05)	0.703 (0.587,0.809)	0.013 (0.008,0.018)	2.81×10^{-5} (2.11×10^{-5}, 4.12×10^{-5})	0.003 (0.003,0.003)	0.026 (0.023,0.03)

Table 4.4: Posterior mean estimates of the model parameters for each of the three models with Geometric distributions in the genetic models on ward 1, with 95% equitailed credible intervals.

Figure 4.17 shows the inferred transmission trees from each of the models. Although some transmission events occur in all of these trees, the shape of the whole tree varies from model to model. Four of the models infer patient T126, who stayed on the ward for 71 days, to be a 'superspreader' of the disease to varying degrees. The Geometric Chain Error model and the Geometric Time Dependent model infer T126 as the source of colonisation of 3 and 4 other patients, respectively. The NB Chain Error model and the NB Chain Poisson model infer T126 as the source of colonisation of a greater number of patients- 6 and 8 respectively. The Geometric Chain Poisson model, however, infers T126 as a source of colonisation only with a low probability. All models appear to infer one or two chains of transmission between the top half of the patients, and one chain between the bottom half, and only the Geometric Time Dependent model joins these two chains together with high probability.

4.12.2 Results for the adjusted models on ward 2

Tables 4.6 and 4.7 give the posterior mean estimates for the values of the model parameters for ward 2 with 95% equitailed credible intervals. The posterior means for the importation parameter p under the Geometric models are either 0.10 or 0.11, and the posterior means vary slightly more under the Negative Binomial models with $p = 0.084$ and $p = 0.12$. Figure 4.18 shows that the number of patients who were inferred with high probability to have been colonised before arrival on the ward was 9 for all models except the Geometric Chain Error model for which is was 8, and the NB

Ward 1				
Model	p	z	β	φ_i
Chain Error model	0.038 (0.013,0.076)	0.718 (0.60,0.83)	0.016 (0.009,0.024)	0.026 (0.023,0.03)
Chain Poisson model	0.030 (0.008,0.066)	0.711 (0.59,0.82)	0.016 (0.009,0.024)	0.026 (0.023,0.03)
	μ	σ	μ_{gl}	σ_{gl}
Chain Error model	120.532 (96.70,155.52)	140.155 (107.04,189.13)	386.05 (365.56,406.75)	218.938 (199.79,241.91)
Chain Poisson model	131.497 (106.70,157.05)	154.305 (118.28,190.15)	383.617 (364.07,403.96)	213.344 (195.45,232.59)

Table 4.5: Posterior mean estimates of the model parameters for each of the three models with Negative Binomial distributions in the genetic models on ward 1, with 95% equitailed credible intervals.

Chain Poisson model, for which is was 11. The posterior mean of the test sensitivity, z, did not vary significantly under the different models. The posterior means range from 78% to 84% and all fall within the 95% credible intervals. Similarly, the posterior means for the transmission parameter, β, were either 0.011 or 0.012 under all models.

The directly comparable within-host genetic parameter, φ_i, was estimated to have posterior mean 0.12 under each of the models and the expected within-host genetic distances was 9. The expected genetic distance between two patients who share a transmission event was estimated to be 66 by the Geometric Chain Error model, and 42 by the Geometric Chain Poisson model. The NB Chain Error model estimated this expected genetic distance to be 55, but the NB Chain Poisson model gave the estimate as 183 although the variance of the distribution was also much larger, with a standard deviation of 304. The expected genetic distance between two patients who are in different transmission chains was estimated to be in a similar region by all of the models, with the Geometric models estimating it between 222 and 224 and the Negative Binomial models estimating it as 208 or 218.

Figure 4.18 shows the inferred transmission trees from each of the models. Similarly

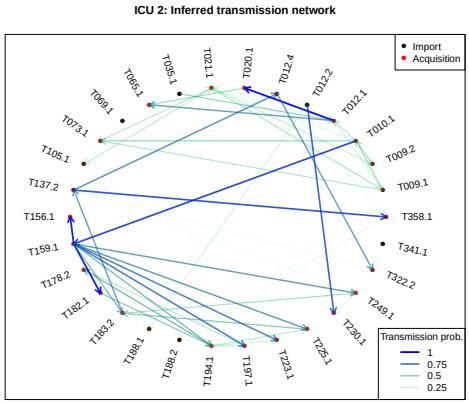
Ward 2						
Model	p	z	β	φ	φ_{gl}	φ_i
Chain Error model	0.10 (0.04,0.19)	0.781 (0.681,0.881)	0.011 (0.006,0.016)	0.017 (0.009,0.024)	0.005 (0.004,0.005)	0.12 (0.062,0.192)
Chain Poisson model	0.113 (0.053,0.188)	0.836 (0.742,0.912)	0.011 (0.007,0.018)	0.028 (0.013,0.049)	0.005 (0.004,0.005)	0.12 (0.062,0.194)
Time Dependent Distances model	0.105 (0.052,0.174)	0.839 (0.755,0.909)	0.012 (0.007,0.018)	0.00013 (9.08×10^{-5}, 1.81×10^{-4})	0.004 (0.004,0.005)	0.119 (0.062,0.193)

Table 4.6: Posterior mean estimates of the model parameters for each of the three models with Geometric distributions in the genetic models on ward 2, with 95% equitailed credible intervals.

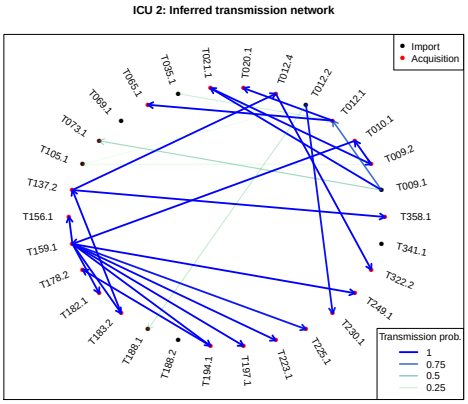
to ward 1, although some transmission events occur with high probability in all of the trees, the broad structure of the trees appears to differ somewhat between models. The Geometric Chain Error model and Geometric Chain Poisson model both infer many different transmission events with lower probability, whereas the Geometric Time Dependent model and the Negative Binomial models all infer more transmission events with higher probability, giving us higher resolution trees. Each model, except the Geometric Chain Poisson model, infers one or more patients with a higher mean outdegree than would be expected given the estimates of β, but these 'super-spreaders' are not the same patients under all the models. The Geometric Chain Error model infers T12 as the source of 4 colonisations with high probability over the course of T12's 3 stays on the ward. The Geometric Time Dependent model also infers T12 as the source of 4 colonisations with high probability, and this model infers T194 to be the source of 3 colonisations. The NB Chain Error model infers T159 as the source of 8 colonisations with high probability, and the NB Chain Poisson model also infers T159 as the source of 3 colonisations, and T10 as the source of 3 colonisations.

4.12.3 Comparison of results with results from Tong et al.

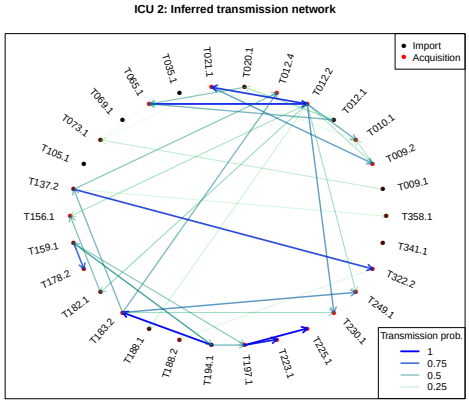
Tong et al. [75], who performed the original study, use a 'clustering' approach in order to infer clades for the sequences in the data. Although these clades do not correspond to a transmission tree it is still of interest to compare our results. We can assume that patients who did colonise each other would have sequences belonging



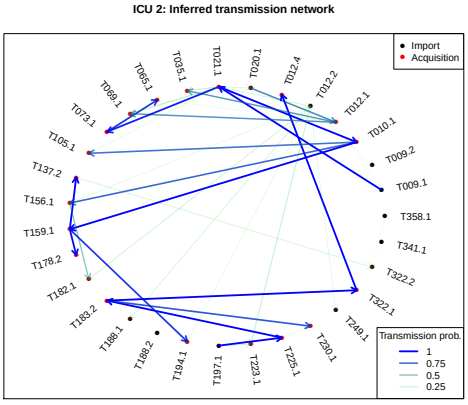
(a) Geometric Chain Error model



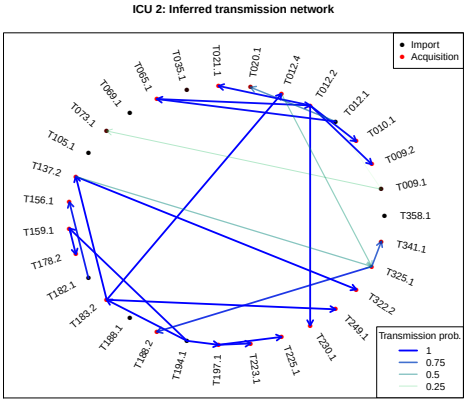
(b) Negative Binomial Chain Error model



(c) Geometric Chain Poisson model



(d) Negative Binomial Chain Poisson model



(e) Geometric Time Dependent model

Figure 4.18: Posterior transmission trees for ward 2 of the Thai data under each of the Geometric and Negative Binomial models. The colour of the nodes represents the probability that the patient was an import. The arrows represent inferred transmission events between patients. The colour of these represents how likely these were to have taken place.

Ward 2				
Model	p	z	β	φ_i
Chain Error model	0.084 (0.036,0.15)	0.827 (0.74,0.90)	0.012 (0.007,0.018)	0.12 (0.062,0.019)
Chain Poisson model	0.12 (0.057,0.20)	0.789 (0.70,0.87)	0.011 (0.006,0.017)	0.12 (0.061,0.019)
	μ	σ	μ_{gl}	σ_{gl}
Chain Error model	55.171 (41.56,76.21)	50.10 (33.06,78.51)	217.702 (196.70,240.31)	258.86 (229.78,288.83)
Chain Poisson model	183.186 (95.38,365.18)	304.44 (134.70,607.11)	208.021 (189.98,228.94)	241.06 (217.30,268.77)

Table 4.7: Posterior mean estimates of the model parameters for each of the three models with Negative Binomial distributions in the genetic models on ward 2, with 95% equitailed credible intervals.

to the same clade. Tong et al. suggest that sequences from the same clade will have a genetic distance < 60 SNPs. For ward 2 the majority of our models (all except the Chain Poisson model and the NB Chain Error model) estimate the expected genetic distance between patients who share a transmission event to be < 60. For ward 1, all three of the original models, with Poisson distributions, and the Geometric Chain Error model estimate that the expected genetic distance between patients who share a transmission event is < 60 whereas the other models give much larger estimates (91, 121, 131). This may suggest that the Geometric Chain Error model is a better fit (as we have seen that the original models are not), or that the simple approach for setting the SNP threshold between clades given by Tong et al. does not work as well for this ward.

Tong et al. found that the mean genetic distance between clades was $140 - 373$ SNPs. We can compare this to our estimates for the expected pairwise genetic distance between the sequences of two patients who are in separate transmission chains. For ward 2 all of the estimates for this expected distance from each of the different models fell within the range given by Tong et al. For ward 1 only the three Geometric versions of the models gave estimates of this expected genetic distance that fell within

the range $140 - 373$, with all of the other models giving larger estimates.

The majority of our models (all except the Geometric Chain Poisson model) infer patient T126 on ward 1 to be a 'superspreader' of the pathogen, meaning that they colonise more susceptible patients that would be expected. Tong et al.'s analysis also suggests that this patient was the source of many transmissions of the pathogen. They found that this patient was continuously colonised and that their pathogen sequences over time were all from the same clade, which was a clade not present on the ward prior to this patient's admission and which became a widespread clade on the ward whilst patient T126 was present. In the same way Tong et al.'s analysis found patient T12 to be a source for many colonisations of susceptible patients on ward 2. All three of our original models and the Geometric Chain Error model and the Geometric Time Dependent model also similarly found patient T12 to colonise more patients than would be expected.

4.13 Model assessment for the Geometric and Negative Binomial versions of the models

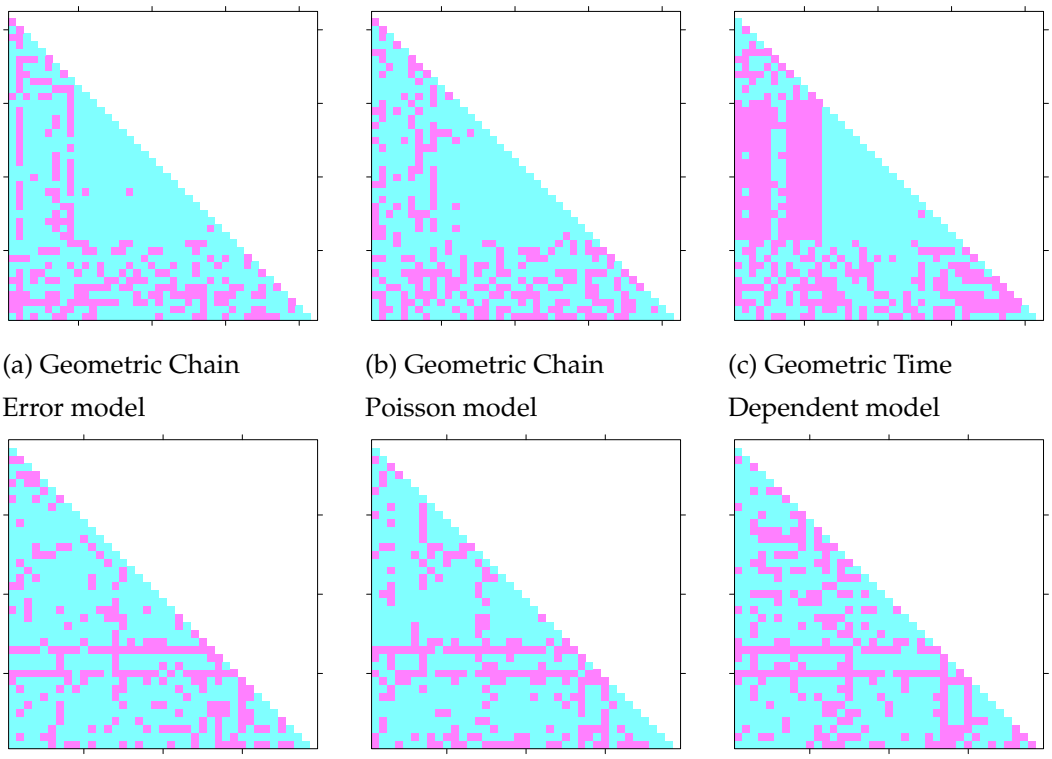
In order to assess whether the new version of the models, which use the Geometric and Negative Binomial distributions fit the data better than the original models which used the Poisson distributions, posterior predictive checks were carried out. For each model on each ward a total of 1000 simulations were generated, with fixed patient admission and discharge times from the data (these do not form part of our model framework), using values for the parameters which were drawn from the posterior densities given by the output of the MCMC algorithm. The number of patients to ever have a positive swab, and the number of patients present on the ward each day with a positive swab, were recorded for each simulation to give an approximation to the posterior predictive distributions of these summary statistics. Table 4.8 gives the 95% highest density regions from the posterior predictive distributions of the number of patients to ever have a positive swab under each model. No extreme p-values were found, giving no evidence against the fit of any of the models. The NB Chain Poisson Error model gave an extreme p-value for the number of positive patients on 8 out of the 173 days of the study for ward 2. All the other models gave a maximum of 2 extreme p-values out of 173 days for each ward. For ward 1 the Geometric Chain Error model and NB Chain Error models both gave no extreme p-values, and for ward 2 the Geometric Chain Poisson, Geometric Time Dependent and NB Chain Poisson models

	Geometric Chain Error model	Geometric Chain Poisson model	Geometric Time Dependent model	NB Chain Error model	NB Chain Poisson model
Ward 1	(2,53)	(1,34)	(1,22)	(7,56)	(1,31)
Ward 2	(4,39)	(7,49)	(7,43)	(6,33)	(7,44)

Table 4.8: 95% highest density regions for the number of patients to have a positive swab for each ward under each of the adjusted models. The observed value for ward 1 was 22 and for ward 2 was 30. Green HDRs indicate that the observed value falls within the HDR.

all gave no extreme p-values.

In order to assess the goodness-of-fit of each of these models to the genetic data from the two wards 1000 genetic distance matrices, $\widetilde{\mathbf{\Psi}}$, were simulated using values drawn from the posterior densities for the genetic parameters, and the times and sources of transmission events. These simulated genetic distance matrices allow for the posterior predictive distribution to be approximated for each individual genetic distance between two patients who both had sequences collected during the course of the epidemic. The percentage of the observed distances that fall within their 95% highest density region is recorded as a goodness-of-fit score, and a binary matrix is plotted with 0 for distances outside their HDR, and 1 for distances inside their HDR. These matrices are given in figure 4.19 for the Geometric versions of the models, and in figure 4.20 for the Negative Binomial versions of the models. When these matrices are compared to the matrices produced under the original models (figure 4.15) it is immediately clear that these models are a much better fit to the genetic data. The Geometric Time Dependent model does not fit as well as the other models, although it is still better than the original models, with a posterior predictive score of 53.19% for ward 1 and 65.59% for ward 2. The rest of the models all perform to a similar standard across both wards. The Geometric Chain Error model has a posterior predictive score of 73.29% for ward 1 and 74.76% for ward 2 whilst the Geometric Chain Poisson model has a posterior predictive score of 71.43% for ward 1 and 72.47% for ward 2. The NB Chain Error model has a posterior predictive score of 76.54% for ward 1 and 68.02% for ward 2 whilst the NB Chain Poisson model has a posterior predictive score of 76.77% for ward 1 and 71.79% for ward 2. Therefore it appears that each of these models fits the data well, although the models with Geometric distributions (except the Time Dependent) appear to fit the ward 1 data slightly better, and the models with Negative Binomial distributions appear to fit the ward 2 data slightly better.



(a) Geometric Chain Error model

(b) Geometric Chain Poisson model

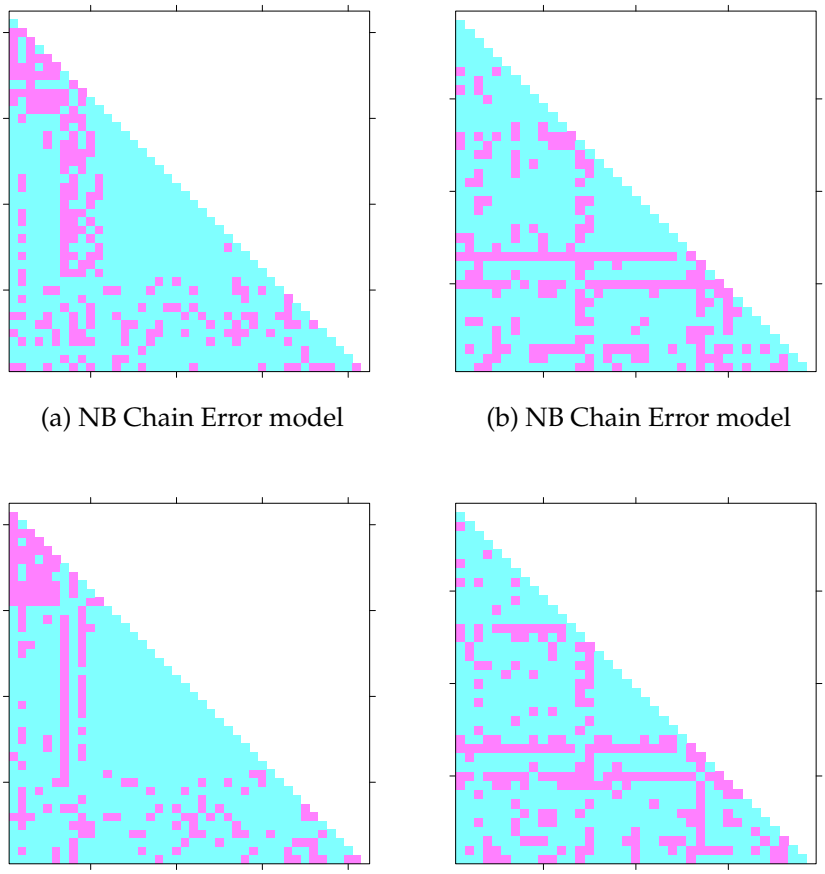
(c) Geometric Time Dependent model

(d) Geometric Chain Error model

(e) Geometric Chain Poisson model

(f) Geometric Time Dependent model

Figure 4.19: Posterior predictive checking to assess the fit of the Geometric versions of the models to the genetic distance matrices from each ward. Blue cells indicate that the observed genetic distance falls within the 95% highest density region given by the posterior predictive distribution and pink cells indicate that the observed genetic distance falls outside the HDR. The layout of the matrices corresponds to the layout of the original genetic distance matrices, so the x-axis, left to right, runs from sequence 1 to sequence n_{seqs}, whereas the y-axis, from bottom to top, runs backwards from sequence n_{seqs} to sequence 1.



(a) NB Chain Error model (b) NB Chain Error model

(c) NB Chain Poisson model (d) NB Chain Poisson model

Figure 4.20: Posterior predictive checking to assess the fit of the Negative Binomial versions of the models to the genetic distance matrices from each ward. Blue cells indicate that the observed genetic distance falls within the 95% highest density region given by the posterior predictive distribution and pink cells indicate that the observed genetic distance falls outside the HDR. The layout of the matrices corresponds to the layout of the original genetic distance matrices, so the x-axis, left to right, runs from sequence 1 to sequence n_{seqs}, whereas the y-axis, from bottom to top, runs backwards from sequence n_{seqs} to sequence 1.

4.14 Discussion

In this chapter we have taken three novel models for the genetic distances between pathogen isolate samples and incorporated them within a model for the spread of the pathogen on a hospital ward. These novel methods introduce dependency between the genetic distances of patients who make up a transmission chain, rather than assuming that the genetic distances are independent, as most previous models have done. The epidemic model is a discrete-time model with a varying population (as patients enter and leave the ward) and the potential for multiple introductions of the disease. We have designed an MCMC algorithm to fit this model to data available from an outbreak of MRSA in a hospital in Thailand in 2008. The algorithm allows for unobserved colonisation times, and missing genetic sequences. The model can be used to harness the information available within whole-genome sequence data in order to reconstruct the pathways of transmission within the wards of the hospital.

A strength of our models is that because we simply use the genetic distances between genetic sequences rather than the sequences themselves, we are able to easily simulate data from our models. Simulating full sequences requires a model which fully specifies which nucleotides in the sequences mutate and is thus more complex and time-consuming. In order to assess the performance of our MCMC algorithm we simulated data using a range of values for each of the parameters. We showed that the algorithm performed well in most situations, and that it recovered both the parameters and the transmission tree well for a range of the parameters which covered what we might reasonably expect from an epidemic such as the one we wished to study.

Assessment of the goodness-of-fit of the models to the data using our novel method for genetic model assessment led to adjustments being proposed to the underlying distributions in the models for the genetic distances. The analysis of the data was performed using these new versions of the models. The model assessment for these models suggested a much better fit to the data from both hospital wards.

The analysis of the outbreak of MRSA in ward 1, under each of these models, revealed fairly similar estimates for the values of the parameters, and fairly similar, high resolution estimations of the transmission tree. However, for ward 2 the models differed more in their estimation of the model parameters and of the transmission tree, although the broad structure remained the same. For ward 1 all of the models

picked out the patient who had the longest stay on the ward as the source of a disproportionally large number of colonisations. For ward 2 all of the models also inferred one or two patients to be the source of a greater number of colonisations than expected, although the identities of these 'super-spreaders' varied between the models. It would be of interest to further investigate these 'super-spreaders' and why they appear to be more infective than other patients on the wards.

4.14.1 Limitations and further work

We note that our model assumes that the state (colonised or susceptible) in which patients are admitted to the ward are independent. However, there are some patients who are admitted more than once to each ward, and therefore it is likely (due to the long carriage time of MRSA) that if these patients were colonised when they were discharged from the ward, then they would still be colonised when they were readmitted, especially as some of them were readmitted only a day after being discharged. It would be of interest to investigate this by setting any patient who leaves the ward colonised to stay colonised if they are readmitted.

We assume homogeneity of susceptibility and infectivity. Relaxing these assumptions may allow us to better investigate those patients who seem to be the source of more colonisation that we would expect. It may also be of interest to explicitly model within-host diversity of the pathogen over time. We currently assume that the sequence taken is representative of the whole population within the host, and that diversity occurs as a result of transmission.

Notation used in Chapter 4

Model description

L	Length of study
n	Number of patients admitted to the ward over the course of the study
n_t	Number of patients present on the ward at time t
n_{pos}	Number of colonised patients over the course of the study
n_{seqs}	Number of genetic sequences in the dataset
t_i^a	Admission time of patient i
t_i^d	Discharge time of patient i
t_i^c	Colonisation time of patient i
ν_i	Number of screening tests received by patient i
t_i^t	Screening time of patient i
	Patients may have multiple tests: $t_i^t = t_{i,1}^t, \ldots, t_{i,\nu_i}^t$
X	Vector of all test results
X_i	Test results for patient i
	Patients may have multiple tests: $X_i = X_{i,1}, \ldots, X_{i,\nu_i}$
ζ_i	Number of sequences sampled from patient i
t_i^s	Sampling time of sequence i
Ψ	The set of all observed genetic distances
$\Psi_{i,j}$	Genetic distance between sequence i and sequence j
H_i	Patient from which sequence i is taken
Q_i	Set of sequences taken from patient i so $Q_i = \{Q_{i,1}, \ldots, Q_{i,\zeta_i}\}$
$D_{i,j}$	Sum of consecutive distances in transmission chain between patients H_i and H_j
	$D_{i,j} = \sum_{r=0}^{k-1} \Psi_{Q_{p_r,1},Q_{p_{r+1},1}}$ where $p_0 = H_i$, $p_k = H_j$
$t_{i,j}$	Difference between the sampling times of sequence i and sequence j
$C(t)$	Number of colonised patients on the ward on day t

Model parameters

ρ	Vector of parameters $\rho = \{p, z, \beta, \Theta\}$
p	Importation parameter
z	Sensitivity parameter
β	Transmission parameter
$\boldsymbol{\Theta}$	Vector of genetic parameters
θ	Genetic diversity parameter for sequences from linked patients
θ_i	Genetic diversity parameter for within host sequences
θ_{gl}	Genetic diversity parameter for sequences from unrelated patients
γ	Genetic chain error parameter for Chain Error model

Notation used in Chapter 4

Model inference and likelihood	
Z	Observed admission, discharge and screening times
T	Unobserved transmission dynamics $T = \{t^c, \phi, s, \Psi^a\}$
Ψ^a	Unobserved genetic distances
s	Sources for each transmission event $s = (s_1, \ldots, s_{n_{acq}})$
ϕ_i	Admission state of patient i
	$\phi_i = 1$ if i is colonised before admission, $\phi_i = 0$ otherwise
n_{acq}	Number of patients who are inferred colonised after admission
TP(X)	Number of true positive test results given X
FN(X, T)	Number of false negative test results given swab results given X and T
trans(i, j)	Number of transmission events between the hosts of sequences i and j
MCMC algorithm description	
n_{sus}	Number of patients with no positive swabs
n_{noseq}	Number of patients with a positive swab but no sequence
n_{add}	Number of the n_{sus} patients with a colonisation added at current iteration
n_{add_0}	Number of patients with colonisation added but no offspring
n_{imp}	Number of patients who are inferred as importations
f_i	Latest possible colonisation day for patient i
Y	Set of (i, j) such that trans$(i, j) = 1$
Y_i	Set of (i, j) such that trans$(i, j) = 0$
Y_g	Set of (i, j) such that trans$(i, j) = \infty$
$N_{\text{trans}(i,j)=1}$	Number of (i, j) such that trans$(i, j) = 1$
$N_{\text{trans}(i,j)=0}$	Number of (i, j) such that trans$(i, j) = 0$
$N_{\text{trans}(i,j)=\infty}$	Number of (i, j) such that trans$(i, j) = \infty$
q_{T,T^*}	Proposal ratio
	$q_{T,T^*} = P(T^* \to T)/P(T \to T^*)$
Simulation method	
A	Average length of stay for a patient on the ward
κ	Test frequency
n_{sw}	Number of patients ever to have a positive swab
Alternative genetic model parameters	
φ	Geometric model genetic diversity parameter for transmitted sequences
φ_i	Geometric model genetic diversity parameter for within-host sequences
φ_{gl}	Geometric model genetic diversity parameter for unrelated sequences
μ, σ	Negative binomial genetic parameters for transmitted sequences
μ_{gl}, σ_{gl}	Negative binomial genetic parameters for unrelated sequences

Chapter 5

Analysis of an epidemic of avian influenza in the Netherlands

5.1 Motivation

In this chapter we analyse data from an outbreak of highly-pathogenic avian influenza in the Netherlands. Highly-pathogenic zoonotic diseases such as this are important to study as they are often characterised by fast transmission and large losses of commercial animals. In the last two decades, a number of outbreaks of these types of pathogens, including avian influenza [64, 65], swine influenza [66, 67], and foot-and-mouth disease [53, 68], have occurred in different countries, with widespread economic impact and concern for public health. Therefore, control measures have an important role to play in lessening the effects of such epidemics, and the better we can understand the dynamics of outbreaks, the better we can design the control measures.

In order to analyse an avian influenza outbreak we will investigate whether the models for genetic distances between isolates which were introduced in chapter 2 can be used within a different model which would be appropriate for non-nosocomial data. This model would need to differ significantly from the model used for the hospital data as populations are likely to be much larger and the impact of individuals coming and going from the population is less dramatic overall, so there is less scope for multiple introductions of the disease. The benefit of using our models in this scenario in comparison to other models, such as that proposed by Bataille et al. [72] (see section 5.3.2), is that instead of including a complex microevolution model for how the genetic mutations occurred in each position on the genome, we simply model the total number of differences between the genomes: the genetic distance. This means that

our model is far simpler and includes less pathogen-specific information, meaning that it can easily be applied in many different scenarios.

5.2 Introduction

The aim of this chapter is to analyse data from an outbreak of avian influenza in the Netherlands in 2003. We will use, and adapt, methods presented in chapter 2 in order to develop a model for reconstructing the path of transmission in this epidemic. This will show that our new models can be fitted to non-hospital datasets, demonstrating their adaptability and flexibility. First, in section 5.3, we will introduce the data which were collected during an outbreak of avian influenza in the Netherlands in 2003 and in section 5.3.2 we will discuss how other research has analysed it, and what types of models have previously been fitted to it. In sections 5.4 we shall introduce models applicable to these data. In section 5.6 we will describe how we implemented these new models in an MCMC algorithm. We will describe, in section 5.7, how we simulated data from this model, and the results we attained from an in depth simulation study. Our results are presented in section 5.8 and we assess the goodness-of-fit of the models in section 5.9. We discuss the value and challenges of this new model in section 5.10. A table of the notation used in this chapter can be found on pages 170-171.

5.3 Avian influenza outbreak in the Netherlands

In this section we will introduce the data collected from farms in the Netherlands affected by the 2003 epidemic of avian influenza. These data were available from Ypma et al. [19], Bataille et al. [72] and Boender et al. [77], with the sequences available on the GISAID database. This was an outbreak over a nine-week period of highly pathogenic avian influenza (HPAI) of type H7N7 that infected at least the 241 commercial farms that we have in our dataset. Control measures put in place led to the culling of 30 million birds [64]. HPAI within poultry has a high transmission rate along with a high death rate and is thought to stem from flocks of wild birds that are infected with low pathogenic avian influenza. EU regulations state that outbreaks of HPAI must be controlled through the culling of infected flocks. In the case of the epidemic in the Netherlands, other control measures were also used, such as a ban on movement of poultry, and the culling of uninfected flocks within a certain radius of infected farms. A timeline of control measure events during the outbreak is shown in

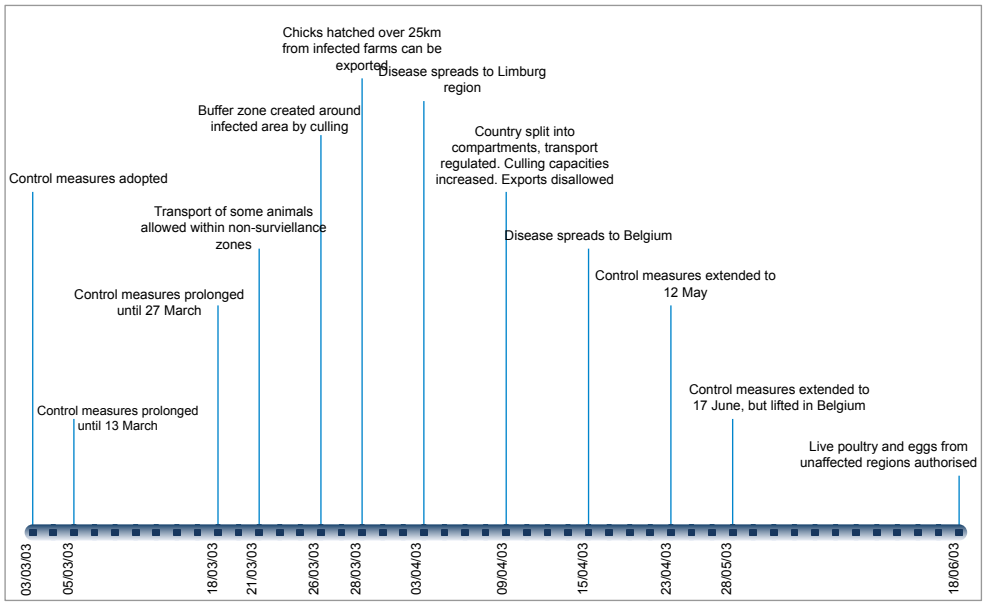


Figure 5.1: A timeline illustrating the timing of key events in the 2003 H7N7 epidemic in the Netherlands.

figure 5.1.

5.3.1 Available data

The data available to us are the geographical coordinates of 241 infected farms, along with details about the type and number of poultry on the farm. We also have the coordinates of 3958 farms which remained susceptible throughout the epidemic, and 1161 farms which were preemptively culled before they were infected, along with the dates of the culling of these farms. Figure 5.2 is a map of the area of the Netherlands with the locations of the farms plotted and coloured according to whether they were infected and culled, culled preemptivity, or remained susceptible throughout the epidemic. Table 5.1 summarises the infected farm data in terms of size and type of farms. We have the date of culling for all 241 farms, and a date on which the pathogen was sampled on 182 of the farms, along with a consensus genetic sequence for the farm from a sample of 5 infected birds taken at this time. For our purposes we only need the total number of positions on the genome where two sequences differ, so we condensed the genetic sequence data into a genetic distance matrix. Figure 5.3a shows a histogram of this genetic distance matrix, and figure 5.3b shows the same information as a heatmap in order to visualise more easily the spread of distances between isolates. From these we can see that all of the genetic distances between infected farms are fairly small, and fairly similar.

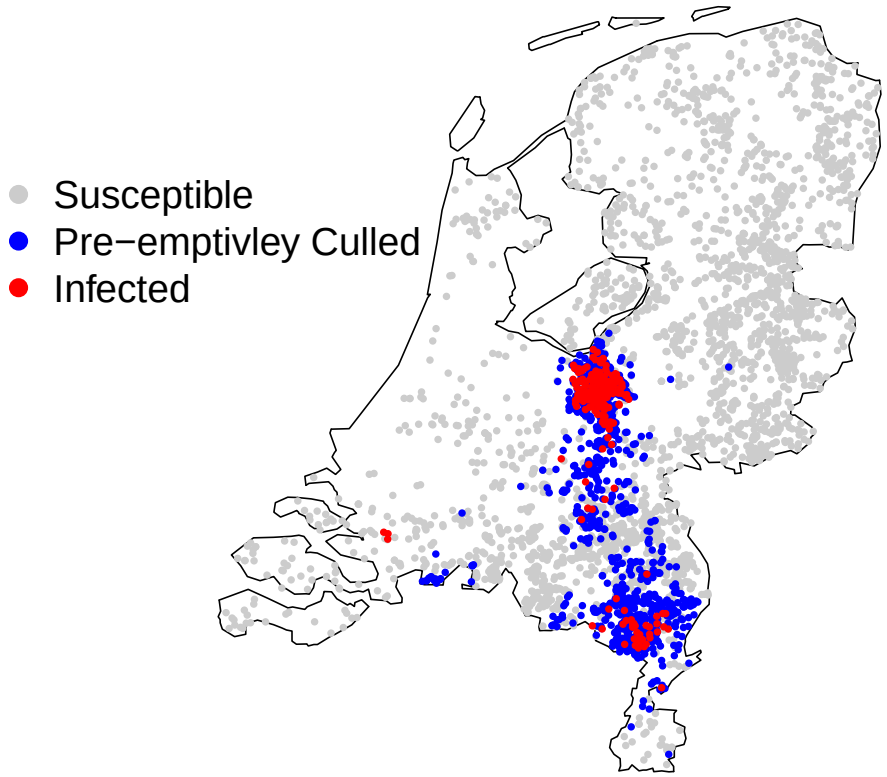
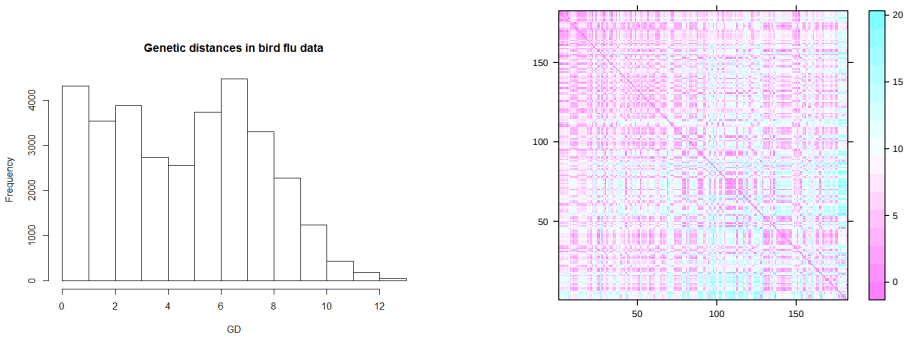


Figure 5.2: A map of the area of the Netherlands in which the epidemic was observed. The locations of the farms in the data are plotted and coloured according to whether they were culled due to infection, culled preemptively, or remained susceptible.



(a) Histogram (b) Heatmap

Figure 5.3: A histogram and a heatmap showing the genetic distances between sequences in the Netherlands data.

Farm Size (no. of birds)	*Layer chickens*	*Broiler chickens*	*Hobby chickens*	*Turkeys*	*Ducks*
$1 \to 100$	0	0	9	0	0
$101 \to 1,000$	3	0	0	0	0
$1,001 \to 5,000$	34	9	0	6	1
$5,001 \to 10,000$	35	16	0	2	1
$10,001 \to 20,000$	42	13	0	6	0
$20,001 \to 30,000$	18	1	0	0	0
$30,001 \to 40,000$	7	0	0	0	0
$40,001 \to 100,000$	12	0	0	1	1
$100,001+$	3	0	0	0	0
Unknown	12	0	5	4	0
Total	**166**	**39**	**14**	**19**	**3**

Table 5.1: The number of infected farms of each size and type in the Netherland data.

5.3.2 Models in the literature

There have been many attempts [19, 64, 65, 72, 78–82] to analyse the data from the H7N7 outbreak in the Netherlands in 2003. Here we will summarise those which model the inter-farm transmission dynamics, as this is what our model will focus on.

Risk maps for the spread of highly pathogenic avian influenza in poultry [82]

In 2007 Boender et al. [82] proposed a method for a spatial analysis of the 2003 avian influenza outbreak in the Netherlands which uses the estimated infection times of farms from Stegeman et al.'s 2004 analysis [64] to produce risk maps for the spread of avian influenza in this setting. The main aim of this analysis was to illustrate how the risk of a spreading epidemic of avian influenza differed throughout the Netherlands, and to estimate the key parameters (transmission and infectious period parameters) governing this spread in the 2003 epidemic. This method identified two high-risk areas within the Netherlands and an estimate of the spatial range over which avian influenza transmits.

The infection times were estimated by assuming that each farm was latently infected for 2 days prior to the date on which it was first reported to be infected, and a discrete time SEIR (susceptible, exposed, infectious, removed) model was used. The risk maps were drawn by plotting each farm in the Netherlands and colouring them ac-

cording to an estimate of a local reproduction number which is equal to the expected number of secondary infections caused by a specific infected farm. The calculation of this reproduction number was informed by the infection status and location of farms, as they were assumed identical otherwise. The risk maps produced show that there were two areas of the Netherlands where farms are dense enough that an outbreak is possible, and that the probability that a farm will be infected by a specific infective farm within a 2km radius is $1\% - 2\%$, compared to less than 0.05% probability that it will be infected by a farm over 10km away. The results fit the data reasonably well, as 162 of the 241 infected in the outbreak were in the areas which were identified as high-risk areas by this method, and of those in lower-risk areas, none started further epidemics in these areas. However, this method does not allow for more in-depth analyses of who-infected-whom in the outbreak in order to answer questions about how the disease spread, and how effective the control measures were.

Evolutionary analysis of inter-farm transmission dynamics in a highly pathogenic avian influenza epidemic [72]

Bataille et al. [72] were the first to use the available genetic data from the epidemic in order to try to reconstruct the route of transmission of the outbreak of avian influenza. Clusters of infection were identified and likely long-distance transmission events were found.

Bataille et al. introduced the genetic data from 72% of the infected farms in the Netherlands avian influenza outbreak. The sequences taken comprised full-length sequences of the H7-hemagglutinin (HA), N7-neuraminidase (NA) and basic polymerase 2 (PB2) gene segments. It was noted that the virus had a high level of genetic diversity, making it suitable for use to determine pathways of transmission. Phylogenies were created using BEAST for each of the different gene segments separately, using a relaxed uncorrelated exponential molecular clock model, which revealed distinct clusters. In order to use the NETWORK program the three gene segments were concatenated, and the program was used to make a median joining phylogenetic network. This also displayed four distinct clusters, with genetic distances within clusters on average being 3-4, and between clusters being on average 11-20. The phylogenetic network suggested 28 likely inter-farm transmission events, with 25 of these between farms close to each other (within a 14km radius). The remaining transmission events were across larger distances and included the transmission that spread the epidemic to the second poultry-dense area of the Netherlands where the outbreak continued.

The identification of that transmission event showed the advantages gleaned by using genetic data, but more insight into the dynamics of the disease spread and how it was affected by control measures would be gained by including the epidemiological data available alongside this useful genetic data.

Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data [19]

Ypma et al. [19] aimed to use both the genetic and epidemiological data to reconstruct the transmission tree of the outbreak of avian influenza by estimating the probability that each infected farm was infected by each other infected farm. By considering each transmission event which was estimated to have a probability greater than 0.5, an estimated transmission tree was constructed.

This method used a Bayesian approach for integrating genetic and epidemiological data to investigate the epidemic. Three types of data were used in this analysis: location, culling date, and genetic sequences. These types of data were assumed to be independent, and a likelihood function was constructed to give the likelihood of the event that farm A infected farm B, and to give the likelihood of the parameters: the rate of decline of infectiousness of a farm after culling, the parameters of the best-fitting distance kernel (scale and shape), and $\mathbf{p} = (p_{ts}, p_{tv}, p_{del})$, the average number of transmissions, transversions and deletions expected in the genetic data between A and B. An MCMC algorithm was used to sample from all possible transmission trees and parameters in a Bayesian framework. The likelihood of observing the genetic distance between farm A and farm B when there are x unobserved transmissions between A and B, given their RNA sequences, uses the probability:

$$\mathrm{p}(d_{ts}, d_{tv}, d_{del} | x, N, \mathbf{p}) = \frac{x^{d_{ts}} \left(p_{ts} N\right)^{d_{ts}}}{\left(1 - (p_{ts} N)\right)^{d_{ts} - xN}} \frac{x^{d_{tv}} \left(p_{tv} N\right)^{d_{tv}}}{\left(1 - (p_{tv} N)\right)^{d_{tv} - xN}} p_{del}^{\mathbb{1}_{del}} \left(1 - p_{del}\right)^{1 - \mathbb{1}_{del}} \tag{5.3.1}$$

where N is the total number of nucleotides that can mutate, d_{ts} and d_{tv} are the observed numbers of transitions and transversions between A and B, and $\mathbb{1}_{del}$ is the indicator function: 1 if a deletion occurred, otherwise 0. This formula assumes that the probability of mutation is so small that the possibility of a nucleotide mutating twice can be neglected. The probabilities of all possible transmission events are attained by averaging over the posterior density over the sample space. By dividing the number of infections caused by each farm by the length of its infectious period an estimate of infectiousness was obtained for each farm in each of the sample trees. The analysis

was rerun excluding geographical or genetic data to see how much information is contained in each type of data respectively. This showed that geographical data are not enough to predict transmission links, although it does give more certainty about transmission pathways when added to the genetic data. Increased accuracy and resolution was shown when using all the data types in the analysis.

This model shows the increased clarity that can be gained in analyses by using both epidemiological and genetic data. However, this model only incorporates the infected farms, meaning that there is no mechanism for estimating the transmission rate and other dynamics of the disease.

5.4 Developing a model for the spread of avian influenza

In chapter 2 we introduced three new models for a genetic distance matrix, which fit into a discrete-time SIR (susceptible, infectious, removed) model. In this chapter we will use a continuous-time SEIR (susceptible, exposed, infectious, removed) model where the individuals are farms and each farm which was exposed and infectious, i, has one genetic sequence, Q_i, which contributes distances to the genetic distance matrix. We will use the models introduced in chapter 2 for modelling the genetic distance matrix. The distances are assumed to be drawn from a probability distribution, the parameters of which depend on the relative positions of the two farms from which the sequences were taken in the transmission tree. Here we recap the three different models. The genetic distance between sequences i and j is given by $\Psi_{i,j}$ and the number of transmission events between F_i and F_j, the farms from which sequences i and j were taken, is given by k. Therefore if F_i directly infects F_j, or vice versa, $k = 1$; if F_i and F_j are in the same transmission chain but separated by more than one transmission event $k > 1$; and if F_i and F_j are not in the same transmission chain $k = \infty$. Since we are only modelling one sequence per farm we do not include the $k = 0$ part of the genetic models in chapter 2, although this could easily be reintroduced if more data were available.

The Chain Error model

In the Chain Error model the genetic distances for sequences from two farms separated by only one transmission event are drawn from a Poisson distribution with parameter θ and those which are from farms in distinct transmission chains have their genetic distance drawn from a Poisson distribution with parameter θ_{gl}. Sequences

from farms which are in the same transmission chain but are separated by more than one transmission event have a genetic distance which is equal to the sum of the genetic distances which make up the underlying transmission chain, with a truncated Poisson distributed error term. Thus, if k is the number of transmission events separating the farms from which sequences i and j were taken, where $k = \infty$ means that they are in different chains, then

$$\mathrm{P}(\Psi_{i,j} = x) = \begin{cases} (\theta_{gl}^x / x!) \exp(-\theta_{gl}) & \text{if } k = \infty \\ (\theta^x / x!) \exp(-\theta) & \text{if } k = 1, \end{cases} \tag{5.4.1}$$

and the conditional probability distribution for genetic distances for pairs of sequences which are in the same chain but separated by more than one transmission event is:

$$\mathrm{P}(\Psi_{i,j} = x | D_{i,j}) = \frac{(k\alpha)^{|D_{i,j}-x|}}{|D_{i,j}-x|! \left(\sum_{l=0}^{D_{i,j}} (k\alpha)^l / l!\right)} \left(\frac{1}{2}\right)^{\mathbb{1}_{\{x \neq D_{i,j}\}}} \mathbb{1}_{\{x \leq 2D_{i,j}\}} \quad \text{if } k > 1 \tag{5.4.2}$$

where $D_{i,j} = \sum_{r=0}^{k-1} \Psi_{Q_{p_r}, Q_{p_{r+1}}}$ where $p_0 = F_i$, $p_k = F_j$, so $D_{i,j}$ is the sum of the consecutive distances between the sequences from farms that compose the transmission chain between F_i and F_j.

The Chain Poisson model

In the Chain Poisson model the genetic distances for sequences from pairs of farms separated by one transmission event are drawn from a Poisson distribution with parameter θ, and those which are from farms separated by more than one transmission event in the same transmission chain have their genetic distance drawn from a Poisson distribution with parameter $D_{i,j} = \sum_{r=0}^{k-1} \Psi_{Q_{p_r}, Q_{p_{r+1}}}$ where $p_0 = F_i$, $p_k = F_j$, which is equal to the sum of the genetic distances between sequences from the farms which make up the underlying transmission chain. Sequences from farms which are in distinct transmission chains have their genetic distances drawn from a Poisson distribution with parameter θ_{gl}. Thus,

$$\mathrm{P}(\Psi_{i,j} = x) = \begin{cases} (\theta_{gl}^x / x!) \exp(-\theta_{gl}) & \text{if } k = \infty \\ (\theta^x / x!) \exp(-\theta) & \text{if } k = 1. \end{cases} \tag{5.4.3}$$

The conditional probability distribution for genetic distances for pairs of sequences which are in the same chain but separated by more than one transmission event is:

$$\mathrm{P}(\Psi_{i,j} = x | D_{i,j}) = (D_{i,j}^x / x!) \exp(-D_{i,j}) \quad \text{if } k > 1. \tag{5.4.4}$$

The Time Dependent Distances model

In the Time Dependent Distances model the genetic distances for sequences from

pairs of farms which are separated by one transmission event are drawn from a Poisson distribution with parameter $t_{i,j}\theta$ where $t_{i,j}$ is a measure of the time that separates the sampling of the i and j. In the case of the avian influenza data we have the sampling times for the sequences, so $t_{i,j} = S_i - S_j$ where S_i and S_j are the sampling times of sequences i and j respectively. The genetic distances for pairs of sequences from farms which are separated by more than one transmission event in the same transmission chain are again drawn from a Poisson distribution with parameter $D_{i,j}$, which is equal to the sum of the genetic distances between sequences from the farms which make up the underlying transmission chain. Sequences from farms which are in distinct transmission chains have their genetic distances drawn from a Poisson distribution with parameter θ_{gl}. Thus,

$$\mathrm{P}(\Psi_{i,j} = x) = \begin{cases} (\theta_{gl}^x / x!)\exp(-\theta_{gl}) & \text{if } k = \infty \\ (t_{i,j}\theta^x / x!)\exp(-t_{i,j}\theta) & \text{if } k = 1. \end{cases} \tag{5.4.5}$$

The conditional probability distribution for genetic distances for pairs of sequences which are in the same chain but separated by more than one transmission event is:

$$\mathrm{P}(\Psi_{i,j} = x | D_{i,j}) = (D_{i,j}^x / x!)\exp(-D_{i,j}) \quad \text{if } k > 1. \tag{5.4.6}$$

The basic idea behind all of these models is the same: that the genetic distance between sequences from two farms which are closely linked in the transmission tree is likely to be significantly smaller than that between sequences from two farms which are not closely linked. In chapter 2 we presented these models for genetic distances in the context of a discrete-time model for nosocomial infections in the situation where we had a small ward of patients, who were admitted and discharged at different times. In order to apply the models to wider situations here we present them within a continuous-time framework to model a closed population (which may be significantly larger than a nosocomial population) with spatial aspects.

5.4.1 Continuous-time model for the spread of avian influenza

The genetic distance model is one part of the whole stochastic model which describes, in continuous time, the spread of a pathogen within a population of (initially) susceptible individuals. These individuals could be patients, single animals, or farms. In this case we will be talking specifically about poultry farms. The model describes the dynamics of the spread of the pathogen on the level of individual farms. The model can be used to construct a transmission tree to show the spread from farm to farm

throughout the area. Although the model described in chapter 2 was an SIR model, here we use an SEIR (susceptible, exposed, infectious, removed) model. This means that each farm carries the pathogen (after an exposure event) for a fixed latent period before it becomes infectious and can pass the infection on to other farms. We use a fixed-length latent period of 1 day, as this value was also used by Ypma et al. [19].

The model assumes that there are a total of N farms in the region. Unlike the model for nosocomial pathogens, we assume that there was a single introduction of the pathogen into the population, and that all infected farms were identified, recorded and culled. All uncolonised farms are assumed equally susceptible, and all colonised farms equally infective. However, the infection pressure from one specific farm to another depends on the geographical distance between the two. The infection pressure is considered to be the amount of the pathogen that is available in the environment to infect a farm. The contribution from infected farm i to the infection pressure on susceptible farm j is given by $\beta_{i,j} = \beta_0 \, \mathrm{e}^{\left(-\delta d_{i,j}\right)}$ where $d_{i,j}$ is the geographic distance between farm i and farm j, so the total infectious pressure on j when it becomes exposed is given by

$$P_j = \sum_{i \in y_j} \beta_0 \mathrm{e}^{\left(-\delta d_{i,j}\right)}$$

where $y_j = \{i : I_i < E_j < R_i\}$, where I_i is the infection time of farm i and E_j is the exposure time of farm j, and R_i is the removal time of farm i. Here, β_0 is our basic transmission rate parameter, while δ is the parameter governing the transmission kernel which dictates how much the geographic distances between farms affect the rate of transmission. We use this infection kernel as we assume that the transmission rate will decay exponentially with the distance between farms. Exponential decay kernels have previously been used for infections between farms [83]. We assume that farms are removed by culling from the population at a rate γ. Therefore our parameter vector is $\rho = (\beta_0, \delta, \gamma, \Theta)$ where Θ is the vector of genetic parameters.

5.5 Inference of parameters of the model for the spread of avian influenza

We can infer the parameters of the model introduced in section 5.4.1 for the dataset discussed in section 5.3.1 which contains swab collection times and culling times for farms, the geographical location of farms, and the differences between the genetic sequences taken from the farms' pathogen isolates.

As we want to use our model to infer the transmission tree of the outbreak we include $T = \{\mathbf{I}, s, \Psi^a\}$. This is the vector of unobserved data consisting of unobserved infection times, sources and genetic distances. There are a total of n_I unobserved infection times, which is one for each infected farm, and they are $\mathbf{I} = (I_1, I_2, \ldots, I_{n_I})$. There are also n_I unobserved sources as each infected farm, i, has the pathogen transmitted to them by a farm, j, which is infected earlier, so j is the source of i. The first farm to acquire the pathogen has source -1 to show that their pathogen came from outside the population of farms. The unobserved sources are given by $s = (s_1, s_2, \ldots, s_{n_I})$. The unobserved genetic distances Ψ^a are:

$$\Psi^a = \left(\Psi^a_{(n_{seq}+1),1}, \Psi^a_{(n_{seq}+1),2}, \ldots, \Psi^a_{(n_{seq}+1),n_{seq}}, \Psi^a_{(n_{seq}+2),1}, \ldots, \Psi^a_{(n_{seq}+n_{noseq}),(n_{seq}+n_{noseq}-1)}\right).$$

We define n_{noseq} to be the number of farms which were observed to be infected but did not have pathogen isolates sequenced, so genetic distances from their sequence are unobserved, and n_{seq} to be the number of farms which have a genetic sequence in the data.

5.5.0.1 Continuous-time model likelihood

We use the model for the spread of avian influenza to derive the likelihood of observing the culling times, X, and the genetic distances, Ψ, given the model parameters, ρ. The model likelihood that we are interested in, $\pi(X, \Psi|\rho)$, is intractable, so we augment the parameter space with unobserved data, T, which gives the augmented likelihood $\pi(X, \Psi|\rho) = \sum_T \pi(X, \Psi, T|\rho, Z) = \sum_T \pi(X, \Psi|T, \rho, Z)\pi(T|\rho, Z)$, where Z is the vector of observed dynamics that we condition on consisting of farm swab times as well as the number of susceptible farms in the region before the epidemic began. Although we can not evaluate this sum due to its complexity and high-dimensionality, we can use a data-augmented MCMC algorithm to sample ρ and T from $\pi(X, \Psi|T, \rho, Z)\pi(T|\rho, Z)$.

We define $L_1 = \pi(X, \Psi|T, \rho, Z)$ as the likelihood of observing the distance matrix and removal times given the unobserved dynamics and parameters, and $L_2 = \pi(T|\rho, Z)$ as the likelihood of the unobserved data given the parameters. The likelihood for each of the three models is given below. In all of these likelihoods N is the total number of farms in the region, $S = \sum_{i=1}^{n_I} \sum_{j=1}^{N} \beta_{ij} \left(R_i \wedge I_j - I_i \wedge I_j\right)$, where $A \wedge B$ means the minimum of A and B, and l is the initial infective. The source of infection for farm j is given by $s(j)$, so if farm k infected farm j then $s(j) = k$. The total number of culled (removed) farms at the end of the epidemic is given by n_R, and n_I is the total number

of farms who are ever infected during the course of the epidemic. These numbers are known from the data. The number of transmission events separating the two farms from which sequences i and j were taken in the transmission tree is denoted by $\text{trans}(i,j)$. Therefore if $\text{trans}(i,j) = 1$ the two sequences belong to farms F_i and F_j where F_i which is directly infected by farm F_j (or vice versa). If $\text{trans}(i,j) > 1$ the two sequences i and j belong to farms which are part of the same transmission chain but are not directly linked through a single transmission, and if $\text{trans}(i,j) = \infty$ the two sequences belong to farms which are not in the same transmission chain. The sum of the consecutive distances between the isolates from farms that compose the transmission chain between F_i and F_j is given by $D_{i,j}$.

Chain Error model likelihood

For the Chain Error model the likelihood is as follows:

$$\begin{aligned}
&\pi(X,\Psi|T,\rho,Z)\pi(T|\rho,Z) = \\
&\gamma^{n_R}\exp\left\{-\sum_{i=1}^{n_R}\gamma\left(R_i-I_i\right)\right\}\times\prod_{j=2}^{n_I}\prod_{i=1}^{j}\left[\mathbb{1}_{\text{trans}(i,j)=1}\frac{\theta^{\Psi_{i,j}}\exp(-\theta)}{\Psi_{i,j}!}\right. \\
&+\mathbb{1}_{\text{trans}(i,j)>1}\frac{(k\alpha)^{|D_{i,j}-\Psi_{i,j}|}}{|D_{i,j}-\Psi_{i,j}|!\left(\sum_{l=0}^{D_{i,j}}\frac{(k\alpha)^l}{l!}\right)}\left(\frac{1}{2}\right)^{\mathbb{1}_{\{\Psi_{i,j}\neq D_{i,j}\}}}\mathbb{1}_{\{\Psi_{i,j}\leq 2D_{i,j}\}} \\
&\left.+\mathbb{1}_{\text{trans}(i,j)=\infty}\frac{\theta_{gl}^{\Psi_{i,j}}\exp(-\theta_{gl})}{\Psi_{i,j}!}\right]\times\prod_{j=1,j\neq l}^{n_I}\left(\beta_{s(j),j}\right)\times\exp\left\{-S\right\},
\end{aligned} \tag{5.5.1}$$

where l is the initial infective.

Chain Poisson model likelihood

For the Chain Poisson model the likelihood is as follows:

$$\begin{aligned}
&\pi(X,\Psi|T,\rho,Z)\pi(T|\rho,Z) = \\
&\gamma^{n_R}\exp\left\{-\sum_{i=1}^{n_R}\gamma\left(R_i-I_i\right)\right\}\times\prod_{j=2}^{n_I}\prod_{i=1}^{j}\left[\mathbb{1}_{\text{trans}(i,j)=1}\frac{\theta^{\Psi_{i,j}}\exp(-\theta)}{\Psi_{i,j}!}\right. \\
&\left.+\mathbb{1}_{\text{trans}(i,j)>1}\frac{D_{i,j}^{\Psi_{i,j}}\exp(-D_{i,j})}{\Psi_{i,j}!}+\mathbb{1}_{\text{trans}(i,j)=\infty}\frac{\theta_{gl}^{\Psi_{i,j}}\exp(-\theta_{gl})}{\Psi_{i,j}!}\right] \\
&\times\prod_{j=1,j\neq l}^{n_I}\left(\beta_{s(j),j}\right)\times\exp\left\{-S\right\},
\end{aligned} \tag{5.5.2}$$

where l is the initial infective.

Time Dependent Distances model likelihood For the Time Dependent Distances model the likelihood is as follows:

$$\begin{aligned}
&\pi(X,\Psi|T,\rho,Z)\pi(T|\rho,Z) = \\
&\gamma^{n_R}\exp\left\{-\sum_{i=1}^{n_R}\gamma\left(R_i-I_i\right)\right\}\times\prod_{j=2}^{n_I}\prod_{i=1}^{j}\left[\mathbb{1}_{\text{trans}(i,j)=1}\frac{t_{i,j}\theta^{\Psi_{i,j}}\exp(-t_{i,j}\theta)}{\Psi_{i,j}!}\right. \\
&\left.+\mathbb{1}_{\text{trans}(i,j)>1}\frac{D_{i,j}^{\Psi_{i,j}}\exp(-D_{i,j})}{\Psi_{i,j}!}+\mathbb{1}_{\text{trans}(i,j)=\infty}\frac{\theta_{gl}^{\Psi_{i,j}}\exp(-\theta_{gl})}{\Psi_{i,j}!}\right] \\
&\times\prod_{j=1,j\neq l}^{n_I}\left(\beta_{s(j),j}\right)\times\exp\left\{-S\right\},
\end{aligned} \tag{5.5.3}$$

where l is the initial infective.

The first term in equations 5.5.1, 5.5.2 and 5.5.3, which is $\gamma^{n_R}\exp\left\{-\sum_{i=1}^{n_R}\gamma\left(R_i-I_i\right)\right\}$, gives the likelihood of the removal times. The infectious period of farm i is given by $R_i - I_i$ (the culling time minus the infection time). The double product term in equations 5.5.1, 5.5.2 and 5.5.3 gives the likelihood of observing the genetic distances between farms given the transmission tree. The first product goes from $j = 2$ to $j = n_I$ and the second goes from $i = 1$ to $i = j$ in order to ensure that we include each pair of sequences (i, j) once and only once, and that we do not include the distance from any sequence to itself (this is why j begins at 2). As there is one sequence for each farm (whether observed or unobserved) the first product goes to n_I. If we did not keep track of the sources of infection for each farm we would expect the last term in equations 5.5.1, 5.5.2 and 5.5.3 to be a double product of $\beta_{i,j}$ over both i and j, but as we do keep track of the sources this is simplified.

5.6 An MCMC algorithm for fitting the model for the spread of avian influenza

In order to fit our models to the data our objective is to sample from the posterior distribution of the parameters, ρ, and the unobserved transmission dynamics, T, given the removal times, X, and the genetic distances, Ψ. Hence we aim to sample from $\pi(X,\Psi|T,\rho,Z)\pi(T|\rho,Z)\pi(\rho)$, where $\pi(\rho)$ is the prior distribution of the parameters. To do this we use a data-augmented MCMC algorithm which, at each iteration, updates the parameters, ρ, and then updates the state of the unobserved transmission tree, T, including the infection times and sources of infection for farms observed to be infected. The vector T also includes unobserved genetic distances, Ψ^a, between all

other sequences and those unobserved sequences from farms which we know to be positive, but lack genetic sequencing data from. These genetic distances are necessary for calculating the probability of transmission from these farms to other farms, and we draw them from the appropriate distributions specified by the model. The MCMC algorithm is now described in detail.

5.6.1 Parameter updates

In this section we describe how the MCMC algorithm updates each of the model parameters, and assign each parameter a prior distribution.

5.6.1.1 Epidemiological parameter updates

We assume that the transmission parameter β_0 has a $\Gamma(\nu_{\beta_0}, \lambda_{\beta_0})$ distribution *a priori*. Therefore the full conditional distribution of β_0, which is $\pi\left(\beta_0|\rho_{-\beta_0}, X, T\right)$ where $\rho_{-\beta_0}$ is the vector of parameters without the component β_0, may be derived, up to proportionality, as:

$$\begin{aligned}
\pi\left(\beta_0|\rho_{-\beta_0}, X, T\right) &\propto {\beta_0}^{\nu_{\beta_0}-1} \exp(-\lambda_{\beta_0}\beta_0) \prod_{j=1, j\neq l}^{n_I} \left(\sum_{i\in y_j} \beta_{i,j}\right) \exp(-S) \\
&\propto {\beta_0}^{\nu_{\beta_0}-1} \exp(-\lambda_{\beta_0}\beta_0) \prod_{j=1, j\neq l}^{n_I} \left(\sum_{i\in y_j} \beta_0 \exp(\delta d_{i,j})\right) \exp(-S) \\
&\propto {\beta_0}^{\nu_{\beta_0}-1} \exp(-\lambda_{\beta_0}\beta_0) {\beta_0}^{n_I-1} \prod_{j=1, j\neq l}^{n_I} \left(\sum_{i\in y_j} \exp(\delta d_{i,j})\right) \exp\left(-\beta_0 A\right) \\
&\propto {\beta_0}^{n_I+\nu_{\beta_0}-1-1} \exp\left(-\beta_0(A+\lambda_{\beta_0})\right),
\end{aligned}$$

where T is the augmented data, X is the vector of removal times, $\rho_{-\theta}$ is the vector of parameters without the parameter θ and

$$S = \sum_{i=1}^{n_I} \sum_{j=1}^{N} \beta_{ij} \left(R_i \wedge I_j - I_i \wedge I_j\right),$$

and

$$A = \sum_{i=1}^{n_I} \sum_{j=1}^{N} \exp\left(\delta d_{ij}\right) \left(R_i \wedge I_j - I_i \wedge I_j\right).$$

It follows that β_0 may be sampled directly, using a Gibbs step, from the distribution

$$\Gamma\left(n_I + \nu_{\beta_0} - 1, A + \lambda_{\beta_0}\right).$$

Similarly, if γ is assigned prior distribution $\gamma \sim \Gamma(\nu_\gamma, \lambda_\gamma)$ then we may sample γ directly, using a Gibbs step, from:

$$\Gamma\left(n_R + \nu_\gamma, \sum_{i=1}^{n_R}(R_i - I_i) + \lambda_\gamma\right).$$

The spatial parameter, δ, is assigned an improper uniform prior distribution on $(0, \infty)$ and is updated using a Metropolis-Hastings random-walk. The step size of the random walk varies according to a Normal distribution with mean 0 and variance σ^2; the acceptance rate is checked every 1000 iterations in order to adjust the variance to maintain an acceptance rate between 0.2 and 0.6.

5.6.1.2 Genetic parameter updates

We assume that the genetic parameter θ has a $\Gamma(\nu_\theta, \lambda_\theta)$ prior distribution. Under the Chain Error model and the Chain Poisson model therefore, the parameter θ, may be sampled directly, using a Gibbs step from:

$$\Gamma\left(\sum_{\substack{(i,j):\\ \text{trans}(i,j)=1}} \Psi_{i,j} + \nu_\theta - 1, N_{par} + \lambda_\theta\right),$$

where N_{par} is the number of pairs of sequences, (i, j), for which $\text{trans}(i, j) = 1$. Under the Time Dependent Distances model the parameter θ may be sampled using a Gibbs step from:

$$\Gamma\left(\sum_{\substack{(i,j):\\ \text{trans}(i,j)=1}} \Psi_{i,j} + \nu_\theta - 1, \sum_{\substack{(i,j):\\ \text{trans}(i,j)=1}} t_{i,j} + \lambda_\theta\right),$$

where $t_{i,j}$ is the difference in sampling times between farm i and j.

Under all three of the models, θ_{gl}, given a $\theta_{gl} \sim \Gamma(\nu_{\theta_{gl}}, \lambda_{\theta_{gl}})$ distribution *a priori*, can be sampled using a Gibbs step from:

$$\Gamma\left(\sum_{\substack{(i,j):\\ \text{trans}(i,j)=\infty}} \Psi_{i,j} + \nu_{\theta_{gl}} - 1, N_{glo} + \lambda_{\theta_{gl}}\right),$$

where N_{glo} is the number of pairs of sequences, (i, j), for which $\text{trans}(i, j) = \infty$.

The genetic error parameter, α, for the Chain Error model, is assigned an improper prior distribution on $(0, \infty)$ and is updated using a Metropolis-Hastings random-walk. The step size of the random walk varies according to a Normal distribution with mean 0 and variance σ^2; the acceptance rate is checked every 1000 iterations in order to adjust the variance to maintain an acceptance rate between 0.2 and 0.6.

5.6.2 Augmented data updates

At each iteration of the MCMC algorithm it performs one of the following augmented data updates, so running it for a large number of iterations gives us the posterior probability of possible transmission events between farms. During each step a candidate data set $T^* = \{I^*, s^*, \Psi^a*\}$ is proposed. Here we describe each step, and define the proposal ratio, $q_{T,T^*} = P(T^* \to T)/P(T \to T^*)$, which is the ratio of the probability of making the reverse move to the probability of making this move. Details of these proposal ratios can be found in appendix E.

- **Changing genetic distances**
  In order to change the genetic distances for a farm which does not have a sequence in the data we pick uniformly at random one of the n_{noseq} farms which was infected, but did not have a pathogen genome sequenced, and change the genetic distances from their sequence to other farms' sequences. If no such farms exist no move is made. We draw a new set of distances between their sequence and each other sequence from each colonised farm according to the relevant probability distributions for the model depending on whether the two farms are in the same transmission chain and adjacent to each other, in the same chain but separated by more than one transmission event, or in separate chains. The probability distributions from which these genetic distances are drawn can be found in equations 5.4.1 and 5.4.2 for the Chain Error model, equations 5.4.3 and 5.4.4 for the Chain Poisson model, and equations 5.4.5 and 5.4.6 for the Time Dependent Distances model. The proposal ratio for this move is

$$q_{T,T^*} = \frac{\prod_{j:\, i \neq j} \mathrm{P}\left[\Psi_{i,j} = \Psi^a_{i,j} | \Theta\right]}{\prod_{j:\, i \neq j} \mathrm{P}\left[\Psi_{i,j} = \Psi^{a*}_{i,j} | \Theta\right]}.$$

- **Updating an infection time and resampling the sources**
  In the original algorithm described in chapter 4, we allowed an update of the infection times for one infective and resampled the source of that infective. However, in the non-hospital case where we assume that there is not more than one

initial infective this move is not reversible in all possible scenarios. Therefore, if we are to update the infection times of one farm we must resample all of the sources of infection (the whole transmission tree for that set of infection times). For this, we select one farm, i, uniformly at random from those which ever get infected. We then sample a new infection time, I_i^*, for that farm by sampling from an exponential distribution with the current value of parameter γ and subtracting this value from the removal time of the farm. The new exposure time, E_i^*, of this farm is the new infection time minus the fixed latent period. If the new exposure time is after any recorded sampling time, no move is made. We then find the farm with the earliest infection time and set this as the initial infection time, and this farm's source is -1. For each other farm, i, we sample a new source of infection, $s^*(i)$, from set $L = \{j : I_j < E_i^* < R_j\}$, where $j \in L$ is picked as the source for i with probability weight $\frac{\beta_{i,j}}{\sum_{j:I_j<E_i^*<R_j}\beta_{i,j}}$. If there is no possible source of infection for any farm then the move is rejected. We define the proposal ratio, q_{T,T^*}, to be

$$q_{T,T^*} = \frac{\mathrm{e}^{\left(\gamma\left(I_i - I_i^*\right)\right)} \prod_{i=1}^{n_I} \frac{\beta_{i,s(i)}}{\sum_L \beta_{i,j}}}{\prod_{i=1}^{n_I} \frac{\beta_{i,s^*(i)}}{\sum_{L^*} \beta_{i,j}}}.$$

- **Change the infection time and source of one farm**
  In this step we pick one of the infected farms uniformly at random from the set of all infected farms, excluding the initial infective. Excluding the initial infective is another way to avoid the irreversibility of the original move from chapter 4 in this scenario with a closed population. We then find the last time that this farm could have been susceptible, which is dependent on whether the farm has any 'offspring' (whether it goes on the infect any other farms in our current configuration of the transmission tree) and whether it has a sampling date. If the farm i has no offspring and no sampling date it could have been susceptible up to the time at which it was removed, R_i, whereas if it does have offspring or a sampling date, it could have only been susceptible up to the time that its first 'child' was exposed, O_i, or the time that it was sampled, S_i, depending on which occurred first. The new infection time, I_i^*, for farm i is sampled by drawing from an exponential distribution truncated at $(R_i \wedge (O_i \wedge S_i)) - I_l$, where I_l is the infection time of the initial infective (the distribution is truncated so that we can not propose that i has an infection time before the infection time of the current initial infective) with the current value of parameter γ and subtracting this draw from $R_i \wedge (O_i \wedge S_i)$. The new exposure time, E_i^*, of this farm is the new infection time minus the fixed latent period. A new source, $s^*(i)$, is sampled for

farm i from set L where $j : I_j < E_i^* < R_j$, where $j \in L$ is picked with probability weight $\frac{\beta_{i,j}}{\sum_{j:I_j<E_i^*<R_j}\beta_{i,j}}$. If there are no possible sources, no move is made. The proposal ratio for this move is given by

$$q_{T,T^*} = \frac{e^{\left(\gamma\left(I_i - I_i^*\right)\right)}\frac{\beta_{i,s(i)}}{\sum_L \beta_{i,j}}}{\frac{\beta_{i,s^*(i)}}{\sum_{L^*}\beta_{i,j}}}.$$

- **Swap the initial infection**

  The previous move does not allow for updating the initial infective, as to update it in the same way as the other infected farms would prevent the move from being reversible. In order to update the initial infective we propose a move in which we swap the initial infective, i say, with the farm, j say, which has the earliest infection time after the initial infective. This means that I_i becomes I_j and vice versa, swapping their infection and exposure times as well as their sources. Naturally i and j retain their removal and sampling time, as these are specified by the data, and if the sampling time of the initial infective before the move is made is before the exposure time of the second infective, no move is made. Any further farms infected by i or j maintain their sources and infection times. The proposal ratio for this move is 1.

- **Change the time of the initial infection**

  Here we sample a new time, I_i^*, for the initial infective, i, by drawing from an exponential distribution with the current value of parameter γ and subtracting this number from $(O_i \wedge S_i)$, the minimum of the exposure time of the initial infective's first offspring j, O_i, and the sampling time of i, S_i (since this is the last time at which the initial infective could have been susceptible). The new exposure time of the initially infected farm is the new infection time minus the fixed latent period. The proposal ratio for the move is given by

  $$q_{T,T^*} = e^{\left(\gamma\left(I_l - I_l^*\right)\right)}$$

  where I_l^* is the proposed initial infection time of initial infective l, and I_l is the current initial infection time of the initial infective.

Acceptance probability

For each augmented data update we accept the proposed augmented data set with probability

$$\min\left(1, \frac{\pi(X,\Psi|T^*,\rho)\pi(T^*|\rho)}{\pi(X,\Psi|T,\rho)\pi(T|\rho)}q_{T,T^*}\right).$$

5.6.3 A block update of an infection time, the full set of sources and the genetic parameters

In order to improve the mixing of the algorithm we also introduced a block update which updates both the transmission tree and the genetics parameters. This update is described here.

This block update was introduced in order to avoid mixing problems similar to those described in chapter 4 where the likelihood becomes stuck in one region without accepting any updates to the configuration of the transmission tree. Here we select one farm uniformly at random from those which ever get infected. We then sample a new infection time, I_1^*, for that farm by sampling from an exponential distribution with the current value of parameter γ and subtracting this number from the removal time of the farm. The new exposure time, E_i^*, of this farm is the new infection time minus the fixed latent period. If the new exposure time is after any recorded sampling time, no move is made. We then find the farm with the earliest infection time and set this as the initial infection time, and this farm's source is -1. For each other farm, i, we sample a new source of infection, $s^*(i)$, from set L where $j : I_j < E_i^* < R_j$, where $j \in L$ is picked as the source for i with probability weight $\frac{\beta_{i,j}}{\sum_{j:I_j<E_i^*<R_j}\beta_{i,j}}$. If there is no possible source of infection for any farm then the move is rejected. Once we have proposed a new transmission tree we propose new values for the genetic parameters, θ^* and θ_{gl}^*, using the distributions $\Gamma(\mu_\theta, \zeta_\theta)$ and $\Gamma\left(\mu_{\theta_{gl}}, \zeta_{\theta_{gl}}\right)$ where

$$\mu_{\theta_{gl}}^* = \sum_{\substack{(i,j):\\ \text{trans}^*(i,j)=\infty}} \Psi_{i,j} + \nu_{\theta_{gl}} - 1$$

and $\zeta_{\theta_{gl}}^* = N_{glo}^* + \lambda_{\theta_{gl}}$, and μ_θ^* and ζ_θ^* for the different genetic distances models are given in table 5.2.

Model	**Parameter** μ_θ^*	**Parameter** ζ_θ^*
Chain Error model	$\sum_{\substack{(i,j):\\ \text{trans}^*(i,j)\leq 1}} \Psi_{i,j} + \nu_\theta - 1$	$N_{par}^* + \lambda_\theta$
Chain Poisson model	$\sum_{\substack{(i,j):\\ \text{trans}^*(i,j)\leq 1}} \Psi_{i,j} + \nu_\theta - 1$	$N_{par}^* + \lambda_\theta$
Time Dependent Distances model	$\sum_{\substack{(i,j):\\ \text{trans}(i,j)\leq 1}} \Psi_{i,j} + \nu_\theta - 1$	$\sum_{\substack{(i,j):\\ \text{trans}(i,j)\leq 1}} t_{i,j} + \lambda_\theta$

Table 5.2: Parameters for the distribution $\Gamma(\mu_\theta, \zeta_\theta)$, from which we draw our new value for θ.

The proposal ratio, q_{T,T^*}, is:

$$q_{T,T^*} = \frac{\mathrm{e}^{\left(\gamma\left(I_i - I_i^*\right)\right)} \prod_{i=1}^{n_I} \frac{\beta_{i,s(i)}}{\sum_L \beta_{i,j}}}{\prod_{i=1}^{n_I} \frac{\beta_{i,s^*(i)}}{\sum_{L^*} \beta_{i,j}}} \times \frac{\frac{\zeta_\theta^{\mu_\theta}}{\Gamma(\mu_\theta)} \theta^{\mu_\theta - 1} \mathrm{e}^{-\zeta_\theta \theta}}{\frac{\zeta_\theta^{*(\mu_\theta^*)}}{\Gamma(\mu_\theta^*)} \theta^{*(\mu_\theta^* - 1)} \mathrm{e}^{-\zeta_\theta^* \theta^*}} \times \frac{\frac{\zeta_{\theta_{gl}}^{\mu_{\theta_{gl}}}}{\Gamma(\mu_{\theta_{gl}})} \theta_{gl}^{\mu_{\theta_{gl}} - 1} \mathrm{e}^{-\zeta_{\theta_{gl}} \theta_{gl}}}{\frac{\zeta_{\theta_{gl}}^{*(\mu_{\theta_{gl}}^*)}}{\Gamma(\mu_{\theta_{gl}}^*)} {}^{*}\theta_{gl}^{(\mu_{\theta_{gl}}^* - 1)} \mathrm{e}^{-\zeta_{\theta_{gl}}^* \theta_{gl}^*}}.$$

Again, we accept the proposed augmented data set with probability

$$\min\left(1, \frac{\pi(X, \Psi | T^*, \rho)\pi(T^* | \rho)}{\pi(X, \Psi | T, \rho)\pi(T | \rho)} q_{T,T^*}\right).$$

5.7 Simulation study

In order to assess the performance of our MCMC algorithm we performed a simulation study. We simulated a number of epidemics of avian influenza according to each of our three models, and then fitted the model using the MCMC algorithm to investigate whether the parameters and transmission tree were well recovered. The parameters β_0, γ, θ and θ_{gl} were given $\Gamma(1, 10^{-6})$ prior distributions and the parameters δ and α (in the Chain Error model) were given improper uniform priors on $(0, \infty)$.

The infection times were initialised by subtracting 1.0 farm's sampling time. If there was no sampling time then the infection time was set to the farm's culling time minus 2.0. The exposure times were set to these infection times minus the length of the latent period. The infection sources were initialised by choosing, uniformly at random, one of the farms which were infectious at the time of the farm's exposure. If no such farms were available then we subtracted 0.5 from the exposure time until the farm had an exposure time at which there was another infectious farm to be their source. The parameters were initialised using 'sensible' values and we ran the MCMC algorithm a number times from different initialisations to check that these did not influence the results. We checked for convergence using the traceplots of the posterior estimates of the parameters and of the likelihood.

Here we describe the method for simulating the data from our models, and in section 5.7.2 we assess the output from the MCMC algorithm on simulations with different parameter values.

5.7.1 Simulation method

In order to simulate an epidemic of avian influenza, we first define the size of the population, N, which is the number of susceptible farms before the outbreak starts, and the relative geographic locations of these farms. For simplicity we pick the farms to be a random subset of the farms from the Netherlands data since we already have a geographical distance matrix for these farms. We also assign values to each of the parameters in the model, and set a fixed latent period, P. We set a length of time h before culling that a farm will be sampled in order for genetic sequences to be produced.

We begin the simulation by setting the first farm, l, to be exposed to the infection by some outside source (-1) at time 0. They become infectious after the fixed latent period, so they are able to infect other farms from time $0 + P = P$. For each remaining susceptible farm j we draw the time at which they would be exposed by this farm from an exponential distribution with parameter $\beta_0 \exp(-\delta d_{i,j})$ where $d_{l,j}$ is the geographical distance between the two farms. We also draw a culling time for the infected farm l from an exponential distribution with parameter γ. Then we find the minimum of the set of exposure times and the culling time, and set that as the next event that happens. We carry on in this fashion, working out the next event which will happen, which will either be an exposure, an infection after a latent period, or a culling. Every time a new farm is exposed we remove any other exposure times we had for them from other farms, and generate their infection time after the latent period, the time at which they would expose each remaining susceptible farm, and their culling time. Then we find the event with the lowest time from the set of all possible exposure, infection and culling times including the new ones we have generated. Thus we move through time until their are either no susceptible farms left to infect, or no infective farms left to transmit the pathogen.

We set each farm to have been sampled h days before their date of culling. If this time would have been before their exposure, we set their sample time to be halfway between their exposure time and their culling time. Each sample from an infected farm produces a genetic sequence. Instead of simulating the exact sequence we just simulate the genetic distance between each sequence as they are sampled and each previously sampled sequence according to the distributions specified by the model. For sequences from farms in distinct chains of transmission we draw the genetic distance from a Poisson distribution with parameter θ_{gl}. For the Chain Error model and the Chain Poisson model we draw the distances for sequences from two farms which

share a direct transmission event from a Poisson distribution with parameter θ, and for the Time Dependent Distances model we draw these distances from a Poisson distribution with parameter $t_{i,j}\theta$ where $t_{i,j} = S_j - S_i$, the difference between each sequence's sampling time. For genetic distances between sequences from farms who are in the same chain but are separated by more than one transmission event, for the Chain Poisson model and the Time Dependent Distances model we draw from a Poisson distribution with parameter $D_{i,j}$, the sum of the distances between sequences from farms in the underlying transmission chain. For the Chain Error model we draw these genetic distances by adding or subtracting from $D_{i,j}$, with probability 0.5, an error term drawn from a normalised Poisson distribution with parameter $k\alpha$ which has been reflected in the y-axis to be a symmetric distribution, and then has been truncated at the value $D_{i,j}$.

5.7.2 Results of the simulation study

The quality of the parameter estimation and the network reconstruction from our simulation study allows us to investigate the performance of the MCMC algorithm. For each of the three models, the Chain Error model, the Chain Poisson model and the Time Dependent Distances model, we simulated 100 data sets with different values of the parameters. A population of 100 farms was used for each simulation. For the geographical distances between these farms each simulation used a different random 100-farm subset of the Netherland distances. We set samples to be taken from infected farms $h = 2$ days before culling, and the latent period was fixed at $P = 2$.

As the aim of the simulation study is to assess the performance of the MCMC algorithm in estimating varied values of the parameters, we varied them one at a time in sets of simulations. Each parameter, ρ_i, was varied over a range which included unlikely, extreme values as well as values which we would expect the parameter to take in real outbreaks. Whilst we varied ρ_i we kept the other parameters, ρ_{-i}, fixed at values which we deemed to be realistic for an outbreak in a region of 100 farms.

Here we give the variations for each parameter:

- For the transmission parameter β_0 we fixed the other parameters ($\delta = 1, \gamma = 0.3, \theta = 5$ (0.5 for Time Dependent Distances model), $\theta_{gl} = 25$) and varied β_0 between 0.1 and 1 in increments of 0.1.
- For the spatial parameter δ we fixed the other parameters ($\beta_0 = 0.2, \gamma = 0.4, \theta =$

5 (0.5 for Time Dependent Distances model), $\theta_{gl} = 25$) and varied δ between 0.5 and 1.4 in increments of 0.1.

- For removal parameter γ we fixed the other parameters ($\beta_0 = 0.2, \delta = 1, \theta =$ 5 (0.5 for Time Dependent Distances model), $\theta_{gl} = 25$) and varied γ between 0.1 and 1 in increments of 0.1.
- For the global genetic parameter θ_{gl} we fixed the other parameters ($\beta_0 = 0.2, \delta = 1, \gamma = 0.4, \theta = 5$ (0.5 for Time Dependent Distances model)) and varied θ_{gl} between 5 and 50 in increments of 5.
- For the chain genetic parameter θ for the Chain Error model and Chain Poisson model we fixed the other parameters ($\beta_0 = 0.2, \delta = 1, \gamma = 0.4, \theta_{gl} = 25$) and varied θ between 1 and 19 in increments of 2.
- For the chain genetic parameter θ for the Time Dependent Distances model we fixed the other parameters ($\beta_0 = 0.2, \delta = 1, \gamma = 0.4, \theta_{gl} = 25$) and varied θ between 0.5 and 1.4 in increments of 0.1.

Therefore each of the 5 parameters was varied to create 10 sets of parameters (50 sets in total). For each of these 10 sets of parameters we simulated 10 epidemics of avian influenza.

5.7.2.1 Parameter estimation

In creating data sets with varying values of each parameter the aim was to assess the performance of the MCMC algorithm across a range of conceivable values and to find the strengths and weaknesses of the algorithm in estimating the parameters. In order to investigate this the MCMC algorithm was run for 50,000 iterations on each simulated dataset and the resulting posterior estimation for the parameter of interest was plotted as a boxplot on the same axes as the other 99 simulations which varied that particular parameter. An example of varying θ for the Chain Poisson model is shown in figure 5.4. From this graph it is easy to see whether the algorithm has recovered the fact that θ is increasing but it is not immediately clear exactly how well it recovers the specific values for θ, so for this we separate the boxplots into individual graphs for each value of θ and plot them over a line which shows the true value, as in figure 5.5. In this example it is clear that the MCMC algorithm has recovered the increase of θ and that it estimates the specific value of θ well. The graphs for the other parameters and models can be found in appendix F, but we will discuss what they show here.

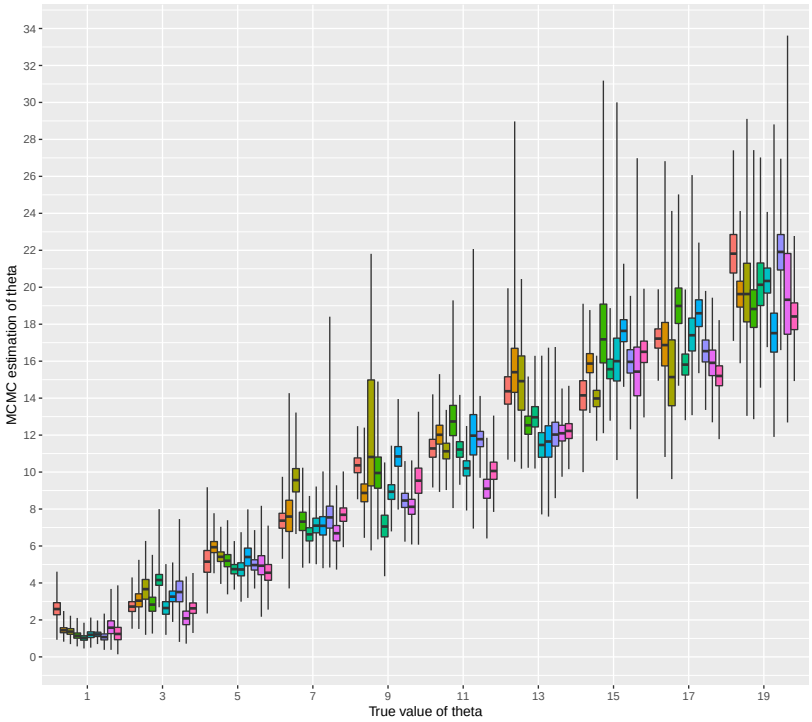


Figure 5.4: The posterior estimates of parameter θ from fitting the Chain Poisson model to 100 simulated datasets from the same model with input value for θ which varies from 5 to 50, with 10 simulations for each increase.

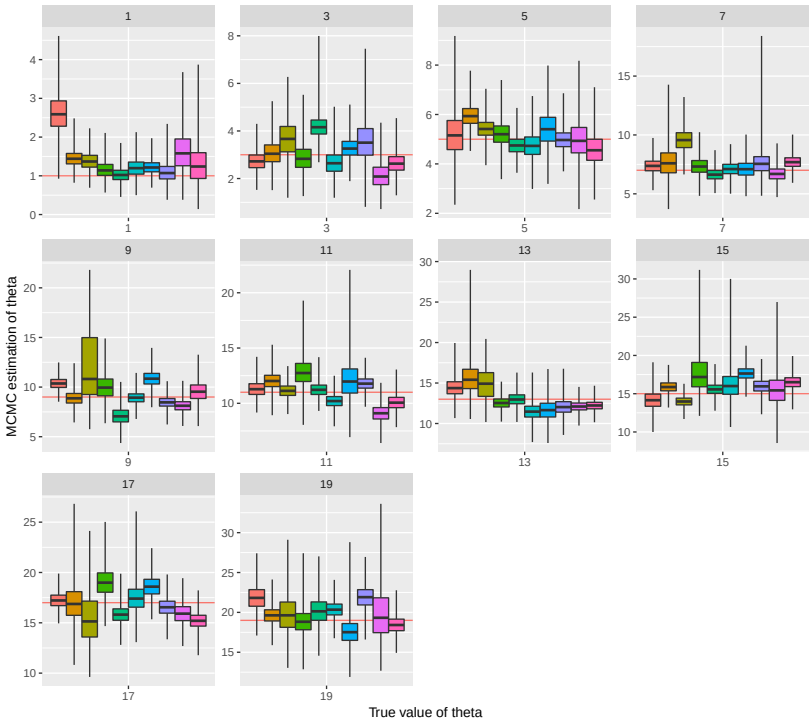


Figure 5.5: The posterior estimates of parameter θ from fitting the Chain Poisson model to 100 simulated datasets from the same model with input value for θ which varies from 5 to 50, with 10 simulations for each increase.

Transmission parameter β_0

The plots (F.1, F.2) for the estimation of the varied transmission parameter β_0 show that for the Chain Error model it is well recovered when $\beta_0 \leq 0.5$, but it is consistently underestimated for $\beta > 0.5$, although the algorithm is still generally recovering that β_0 is increasing between sets of simulations. For the Chain Poisson model (F.11, F.12) and Time Dependent Distances model (F.21, F.22) this underestimation of the parameter is evident even at $\beta_0 = 0.4$. Since we would expect parameteres β_0 and δ to be correlated we also looked at the estimation of δ as we varied β_0. We found that δ was generally well estimated when β_0 was well estimated, although for very small values of β_0 the parameter δ tended to be overestimated.

Spatial parameter δ

The plots (F.3, F.4, F.13, F.14, F.23, F.24) for the estimation of the varied spatial parameter δ show that for all three models the algorithm estimates the parameter well across the range 0.5 to 1.4, although there is slightly more variability in the estimates at the ends of the range.

Removal parameter γ

The plots (F.5, F.6, F.15, F.16, F.25, F.26) for the estimation of the varied removal parameter γ show that for all three models the estimation gets better and more consistent as γ increases. For the Chain Poisson model the estimates generally look good across the range, whereas for the Chain Error model there is quite a lot of variability in the success of the estimation when $\gamma \leq 0.6$. The Time Dependent Distances model consistently underestimates γ.

Genetic parameters θ and θ_{gl}

The plots for the estimation of the varied chain genetic parameter θ for the Time Dependent Distances model (F.27, F.28) show that although the algorithm has captured the fact that θ increases between sets of simulations, it is poor at estimating the value of θ. This could be because of a wide range in the value of $t_{i,j}$ which is a factor in the full conditional of θ in this model. The plots for the estimation of the varied chain genetic parameter θ for the Chain Error model (F.17, F.18) and the Chain Poisson model (5.4, 5.5) show that θ is generally well estimated across the range 1 to 19 despite a couple of simulations where it is estimated poorly. We assume that is due to the particular transmission tree in those simulations as they had fewer than average infected farms and so gave less information for the estimate of θ. The plots (F.9, F.10, F.19, F.20,

F.29, F.30) for the estimation of the varied global genetic parameter θ_{gl} show that for all three models θ_{gl} is estimated well across the range 5 to 50.

5.7.2.2 Transmission tree estimation

We used the same set of simulations that were used for investigating parameter estimation in section 5.7.2.1 to investigate the strengths and weaknesses of the MCMC algorithm in recovering the network of transmission for simulations created with a range of values for each parameter.

In order to visualise how many of the transmission events were correctly estimated by the algorithm we took the output of the MCMC algorithm for each simulation and found the most likely source of exposure for each farm by finding which source was assigned to that particular farm for the largest number of iterations (after the burn in period had been excluded). We compared this estimated most likely source to the true source by producing separate plots for each parameter in each version of the model. These plots have a boxplot for each value of the parameter being investigated which plots the proportion of sources correctly identified for each of the 10 simulations with that value. Figure 5.6 shows an example of this type of plot for varied values of θ_{gl} for the Chain Error model. The plots for the other parameters and models can be found in appendix G but we discuss what they show here.

Effect of varying transmission parameter β_0
The plots (G.1, G.6, G.11) of the proportion of transmission sources recovered for varied β_0 show that for the Chain Poisson model and the Time Dependent Distances model the transmission is reasonably well estimated across the range of $\beta_0 \in (0, 1)$. However, for the Chain Error model we see that there is much more variability in the estimation of the transmission tree, especially when $\beta > 0.5$. This suggests that this model struggles when the transmission tree is larger and more complicated.

Effect of varying spatial parameter δ
The plots (G.2, G.7, G.12) of the proportion of transmission sources recovered for varied δ show that for all three models the transmission tree is well estimated across the range of $\delta \in (0.5, 1.5)$. All of the models cope well with varying values for the spatial parameter.

Effect of varying removal parameter γ

The plots (G.3, G.8, G.13) of the proportion of transmission sources recovered for varied γ show that for the Chain Poisson model and the Time Dependent Distances model the transmission is reasonably well estimated across the range of $\gamma \in (0,1)$. However, for the Chain Error model we see that the transmission tree recovery is generally poor when $\gamma \leq 0.2$, although for larger values the estimation is good.

Effect of varying genetics parameters θ and θ_{gl}

The plots (G.4, G.9, G.14) of the proportion of transmission sources recovered for varied θ show that for all three models the transmission tree is well estimated across a range of values, although the Chain Poisson model shows more variability in the success of transmission tree recovery as θ increases. This makes sense as it will be harder to estimate the transmission tree as θ and θ_{gl} become closer in value. This is seen also in the plots (G.5, 5.6, G.15) of the proportion of transmission sources recovered for varied θ_{gl} as the algorithm does markedly worse, for all three models, in estimating the tree when $\theta_{gl} \leq 10$ so it is close to the value of θ.

The MCMC algorithm appears to recover the transmission tree, and the parameters, well for a range of values of the parameters. The algorithm recovers the transmission tree best when $\beta_0 < 0.5$ and $\gamma > 0.3$ and θ_{gl} is obviously larger than θ.

5.8 Results for the Netherlands data

Having tested the algorithm on simulated data, we analysed the data available from the 2003 outbreak of avian influenza in the Netherlands under each of our three models in order to estimate the transmission tree, which describes which farms transmitted the disease to other farms, and the times of these transmission events. For each model we ran the MCMC algorithm for 400,000 iterations, with 10 augmented data steps taking place during each iteration. The parameters β_0, γ, θ and θ_{gl} were given $\Gamma(1, 10^{-6})$ prior distributions and the parameters δ and α (in the Chain Error model) were given improper uniform priors on $(0, \infty)$.

The infection times were initialised by setting them to be 24 hours prior to the farm's sampling time. If there was no sampling time then the infection time was set to be 48 hours before the farm's culling time. The exposure times were equal to these infection times minus the length of the latent period. The infection sources were initialised by

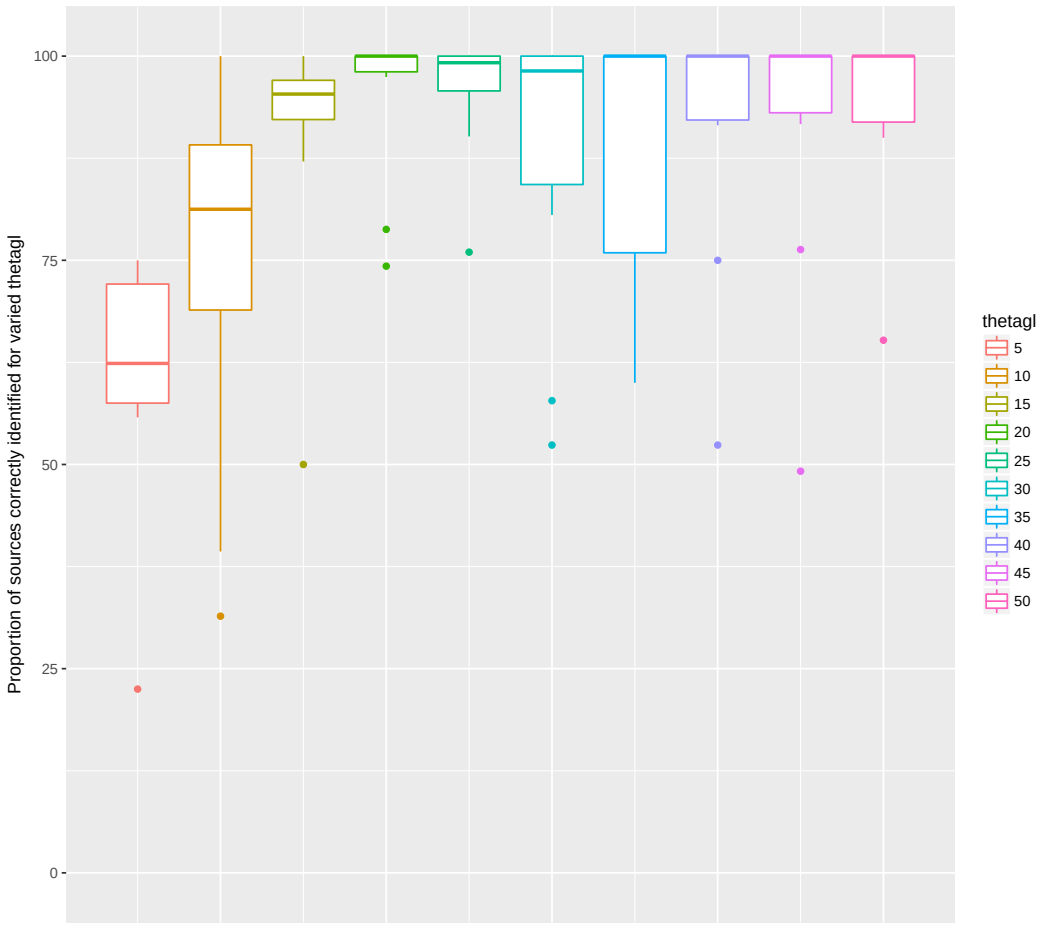


Figure 5.6: Boxplots to show the proportion of infection sources for farms recovered correctly for simulations with varied values for parameter θ_{gl} for the Chain Error model.

randomly picking one of the farms which were infectious at the time of the farm's exposure to be its source. If no source was available then we subtracted 0.5 from the exposure time until the farm had an exposure time at which there was an available source. The parameters were initialised using 'reasonable' values and we checked that these did not influence the results by running the MCMC algorithm multiple times from different initialisations. We checked for convergence using the traceplots of the posterior estimates of the parameters and of the likelihood.

5.8.1 Results from the Chain Error model

We performed analysis under the Chain Error model using a latent period of 1 day as in the model presented by Ypma et al. [19]. Table 5.3 on page 171 gives the posterior means with 95% equitailed credible intervals for the parameters of the model. This model estimated that sequences from farms which shared a transmission event had a mean genetic distance of 9 SNPs, and sequences from farms which were in distinct chains of transmission had a mean genetic distance of 11 SNPs. This reflects the data because, as we commented in section 5.3.1, the genetic distances are all fairly similar and fairly small. However, the 95% credible intervals for parameter θ and θ_{gl} do not overlap, so although the distributions for pairs of sequences from farms which share a transmission and those in separate chains are similar, they are distinctly different.

Figure 5.7 gives the posterior transmission tree for transmission events inferred with probability 0.4 or higher. Transmission events which have posterior probability smaller than 0.4 were not included to prevent the figure from becoming too noisy and unclear. This figure shows that the transmission events with the highest posterior probability are those between farms which are close together, which was expected. There are no transmission events between the two geographical clusters with high probability. In order to visualise the transmission tree for all the farms rather than just those with the highest posterior probabilities we created a median tree from the output of our MCMC algorithm, using the *treespace* package [84]. This package takes the transmission tree from each iteration of the MCMC algorithm output and finds a representative tree using a metric on the transmission trees which quantifies their differences. This median tree is shown in figure 5.8. This transmission tree proposed by our model seems to be intuitively sensible, as it only has one transmission event which occurs over the large geographic distance between the two clusters of farms.

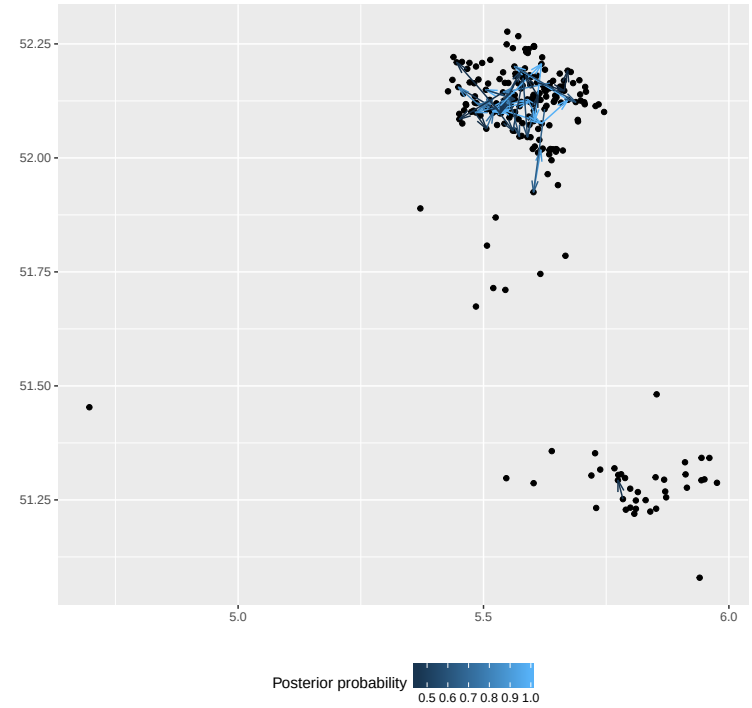


Figure 5.7: Transmission events with a posterior probability greater than 40% under the Chain Error model

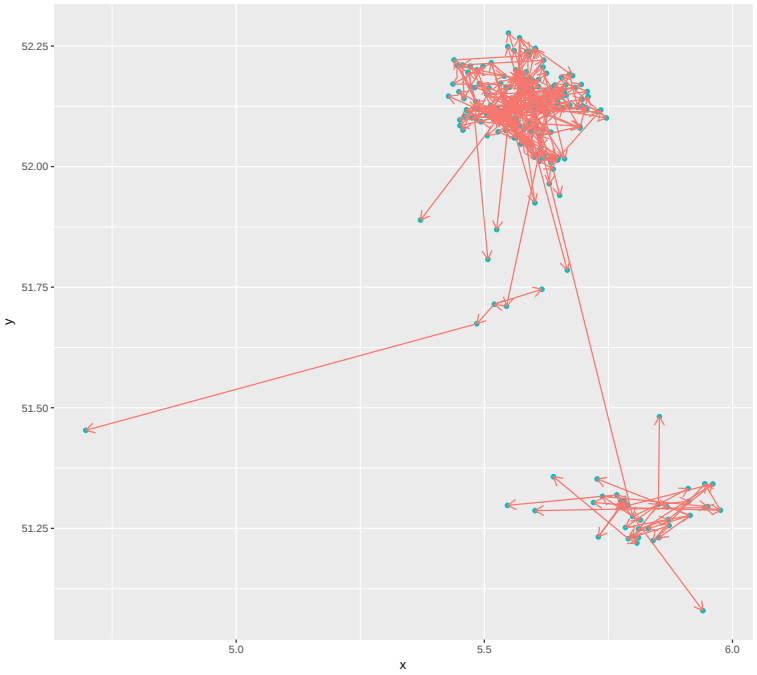


Figure 5.8: The mean transmission tree under the Chain Error model

5.8.2 Results from the Chain Poisson model

We performed analysis under the Chain Poisson model using a latent period of 1 day. Table 5.3 on page 171 gives the posterior means with 95% equitailed credible intervals for the parameters of the model. This model also estimated that sequences from farms which shared a transmission event had a mean genetic distance of 9 SNPs, and sequences from farms which were in distinct chains of transmission had a mean genetic distance of 11 SNPs. As for the Chain Error model this reflects the small, clustered genetic distances in the data. Again, the 95% credible intervals for parameter θ and θ_{gl} do not overlap, so although the distributions for sequences from pairs of farms which share a transmission and those in distinct chains are similar, they are distinctly different, suggesting that there is valuable information contained in the genetic data.

Figure 5.9 gives the posterior transmission tree for transmission events inferred with probability 0.4 or higher. Transmission events which have posterior probability smaller than 0.4 were not included as there were so many possible edges at smaller probability levels that the figure gives no information about the tree configuration. The figure shows that the transmission events with the highest posterior probability are those between farms which are close together, but there was also one transmission event estimated to have occurred between the two geographical clusters. This makes sense, as the disease has to move between the clusters, but it is unlikely that it would have been transmitted across such a big geographical distance multiple times. The average time that the algorithm estimates for this transmission (from farm 32 to farm 205) is day 37, and the first sample taken from the lower cluster in the data is from day 39.

We obtained a median tree from the output of our MCMC algorithm [84]. This median tree is shown in figure 5.10. This median tree again suggests that there was just one transmission of the disease between the two geographical clusters, and it also suggests a minimal number of transmission events across the other large geographical distances.

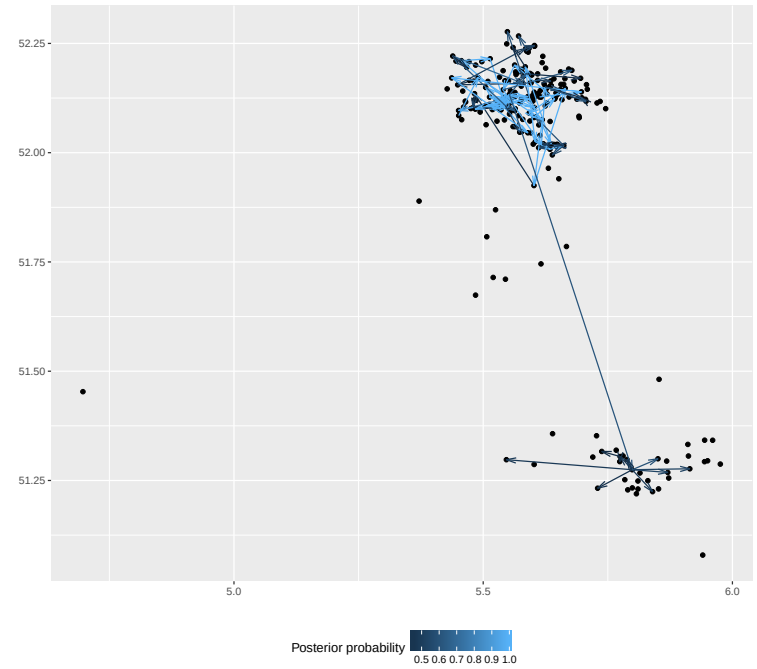


Figure 5.9: Transmission events with a posterior probability greater than 40% under the Chain Poisson model

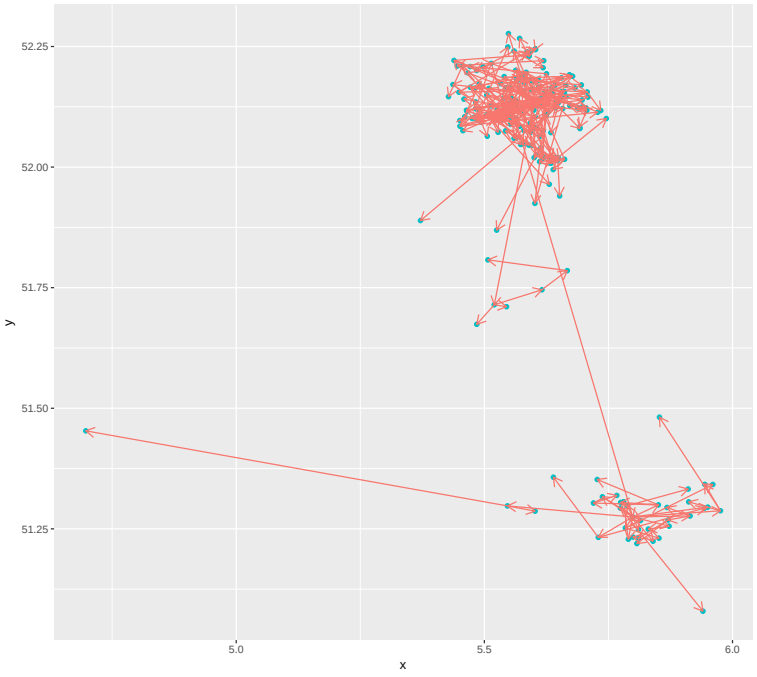


Figure 5.10: The median transmission tree under the Chain Poisson model

5.8.3 Results from the Time Dependent Distances model

We performed analysis under the Time Dependent Distances model using a latent period of 1 day. Table 5.3 on page 171 gives the posterior means with 95% equitailed

credible intervals for the parameters of the model. In this model the genetic diversity parameters have slightly different meanings because of the time dependence in the distributions for the genetic distances in the model, so the results can not be interpreted intuitively. It is notable that the posterior mean for θ_{gl} is considerably lower than for either of the other two models, suggesting that the transmission tree estimated under this model differs from the trees produced by the Chain Error and Chain Poisson models.

Figure 5.11 gives the posterior transmission tree for transmission events inferred with probability 0.4 or higher. Transmission events which have posterior probability smaller than 0.4 were not included. The figure shows that the transmission events with the highest posterior probability are those between farms which are close together, but there was also one transmission event estimated to have occurred between the two geographical clusters. The model estimates that on average this event happens on day 24 of the outbreak, from farm 17 in the upper cluster, to farm 207 in the lower cluster.

We obtained a median tree from the output of our MCMC algorithm [84]. This median tree is shown in figure 5.10. This median tree also suggests that there was only one transmission event between the two geographical clusters.

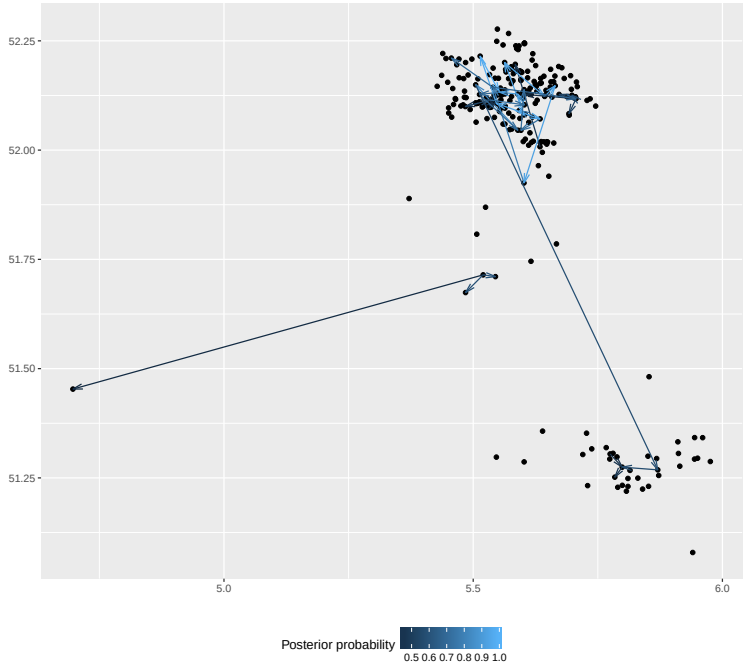


Figure 5.11: Transmission events with a posterior probability greater than 40% under the Time Dependent Distances model

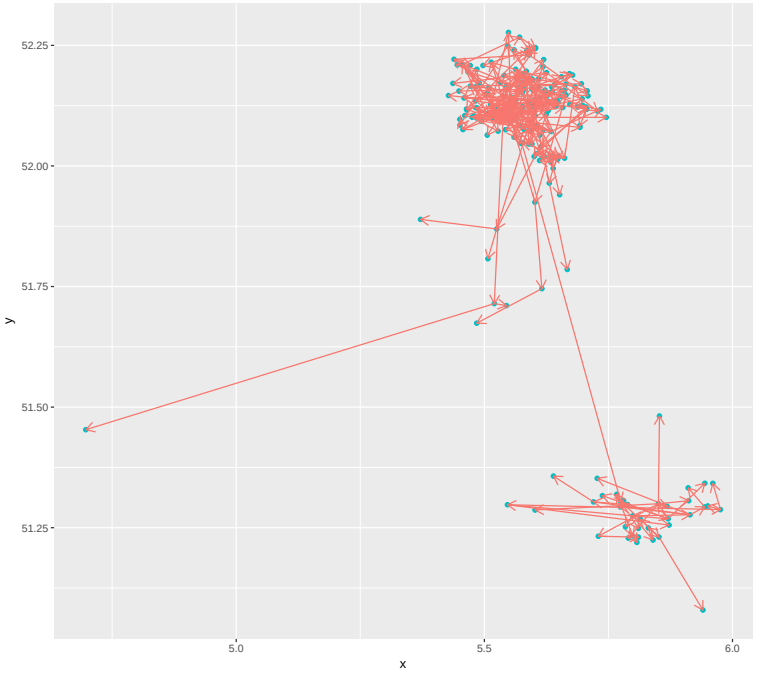


Figure 5.12: The median transmission tree under the Time Dependent Distances model

5.8.4 Comparison of results from the different models

The parameter estimates for all three models can be found in table 5.3, which shows that there is not a significant difference between the results from the three models, although the transmission rate β_0 and the removal rate γ are estimated to be slightly higher by the Chain Poisson model than by the other two models. The Chain Error model and the Chain Poisson model give very similar estimates for the genetic parameters, but we are unable to compare these with the θ parameter from the Time Dependent Distances model as it has a different meaning. We have already noted that although the θ_{gl} parameter should be comparable with the other two models, the Time Dependent Distances model gives a much lower estimate of 8 SNPs between sequences, rather than the 11 SNPs estimated by the other two models. This shows that when time is explicitly included in the model, some of the farms which have small genetic distances between their sequences are placed in separate transmission chains.

Comparing figures 5.7, 5.9 and 5.11 allows us to see the differences in the estimation of the transmission tree under each of the models. We can see that the Chain Error model infers the least number of transmission events with high probability so the tree is made up only of transmission events between farms that are near neighbours of each other, whereas both of the other two models not only infer transmission events over longer geographical distances, but also infer a greater number of these

Latent Period: 1 day						
Model	β_0	δ	γ	θ	θ_{gl}	α
Chain Error model	0.0000878 (0.000067,0.00012)	1.183 (0.95,1.46)	0.0286 (0.025,0.033)	9.256 (8.71,9.82)	11.163 (11.11,11.21)	1.485 (1.34,1.79)
Chain Poisson model	0.000152 (0.00012,0.00020)	1.393 (1.15,1.67)	0.0406 (0.035,0.047)	9.407 (8.86,9.90)	11.254 (11.21,11.31)	-
Time Dependent Distances model	0.000104 (0.000077,0.00015)	1.344 (1.06,1.70)	0.029 (0.025,0.034)	0.207 (0.19,0.23)	8.136 (8.08,8.22)	-

Table 5.3: Posterior mean estimates of the model parameters for each of the three models with a latent period of 1 day, with 95% equitailed credible intervals.

small distance transmission events within each cluster of farms. The transmission event which is inferred between the two clusters of farms in the Chain Poisson model and the Time Dependent Distances model is not exactly the same event (it occurs between different farms for each model), but we can see that the transmission originates from a broadly similar area, and is received by a farm in a similar location under each model. The Chain Poisson model estimates more transmission events with high probability over slightly longer distances in the bottom cluster of farms, and the Time Dependent Distances model is the only one to estimate some transmission events with high probability for the farms which are separate from either geographical cluster.

5.8.5 Results with a different value for the latent period

So far we have used a latent period of 1 day in our models. This value was chosen as it was used in the model presented by Ypma et al. [19] which we discussed in section 5.3.2. However, a further study of the available literature shows that the commonly used latent period is between 1 and 2 days, with 2 days often being used [77, 78, 80, 81]. In the study by van der Goot et al. [85] it is suggested that a latent period of 1-2 days fits the Netherland data better than a latent period of 1 day. Therefore we analysed the data once again using each of our three models with a latent period of 2 days to explore whether there would be a significant difference in the results. Again we ran the MCMC algorithm for 400,000 iterations for each model, with 10

Latent Period: 2 days						
Model	β_0	δ	γ	θ	θ_{gl}	α
Chain Error model	0.0000835 (0.000064,0.00011)	1.021 (0.84,1.24)	0.032 (0.027,0.039)	9.341 (8.78,9.93)	11.085 (11.03,11.13)	1.190 (1.07,1.62)
Chain Poisson model	0.000198 (0.00014,0.00029)	0.926 (0.63,1.32)	0.044 (0.037,0.055)	9.432 (8.87,9.98)	11.220 (11.15,11.31)	-
Time Dependent Distances model	0.000117 (0.000087,0.00016)	1.397 (1.11,1.75)	0.0306 (0.026,0.036)	0.208 (0.19,0.23)	8.128 (8.07,8.18)	-

Table 5.4: Posterior mean estimates of the model parameters for each of the three models with a latent period of 2 days, with 95% equitailed credible intervals.

augmented data steps at each iteration. The prior distributions for the parameters and the initialisation of the times, sources and parameters were the same as those described in section 5.8.

Table 5.4 gives the values of the parameters for the different models with the latent period fixed at 2 days. For all three of the models the parameters stay in the same regions when we extend the latent period, although the values for the transmission parameter β_0 each rise slightly, as do the values for the removal parameter γ. The only noticeable change in the distance parameter δ is for the Chain Poisson model, where it falls by 0.46 from 1.39 to 0.93. The genetic parameter estimates hardly change.

Figures 5.13, 5.14 and 5.15 show the estimated transmission trees under each of the three models with a latent period of 2 days. We have again plotted transmission events that were estimated with a posterior probability of 0.4 or greater. For all three models there appears to be greater resolution in the trees estimated under the longer latent period, as there are more transmission events estimated with a higher probability (lighter blue lines). The biggest difference between these trees and the trees with a shorter latent period is seen in the Chain Error model tree, where we now have an estimated transmission event linking the two geographical clusters rather than just transmission events between farms which are close together. With the latent period fixed at 2 days, the Chain Poisson model estimates that the source of the transmis-

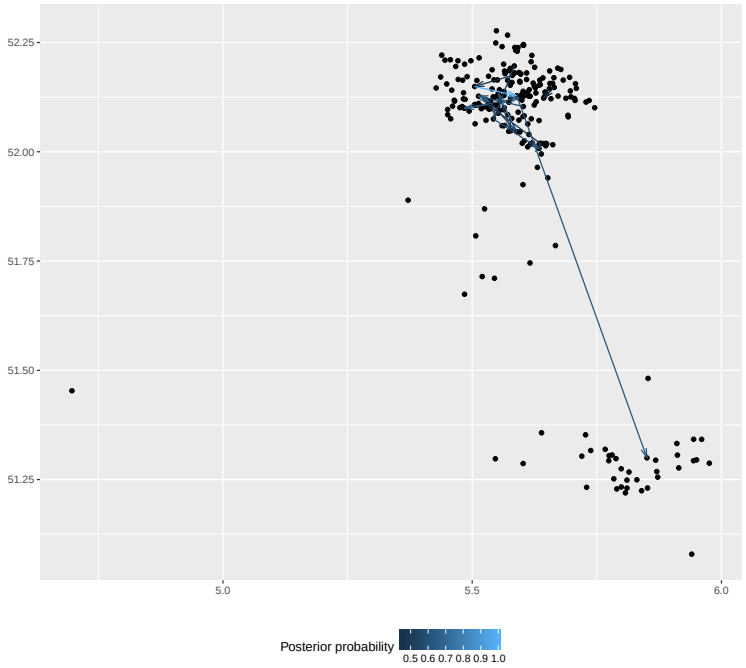


Figure 5.13: Transmission events with a posterior probability greater than 40% under the Chain Error model with a latent period of 2 days

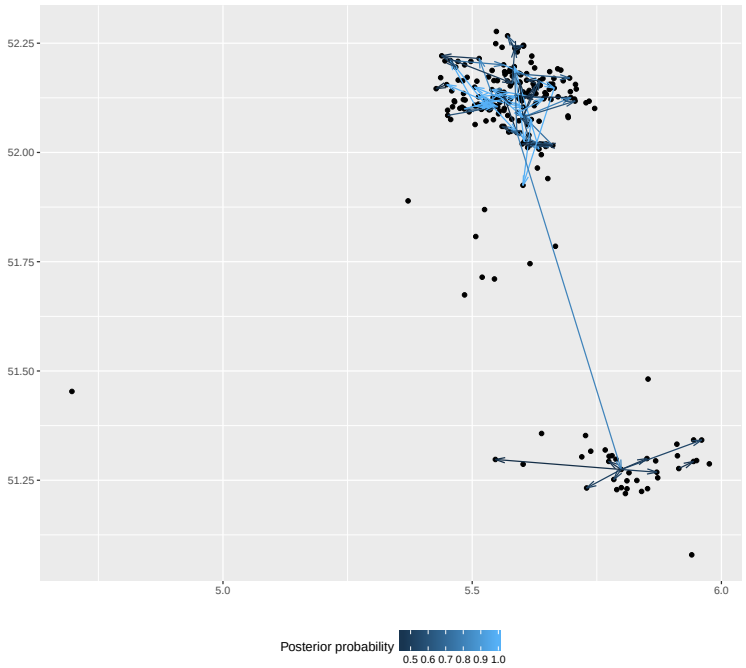


Figure 5.14: Transmission events with a posterior probability greater than 40% under the Chain Poisson model with a latent period of 2 days

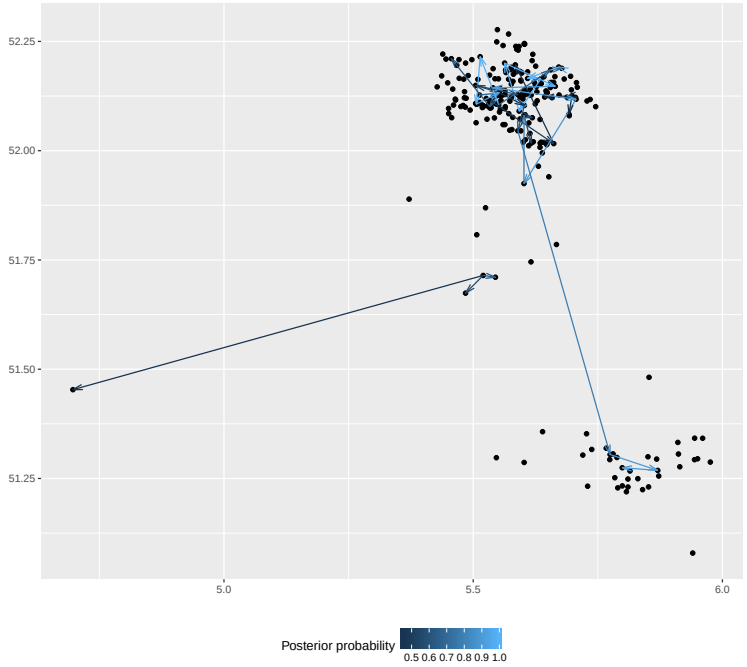


Figure 5.15: Transmission events with a posterior probability greater than 40% under the Time Dependent Distances model with a latent period of 2 days

sion event which links the clusters was likely to have been farm 167, which is a small backyard hobby farm. Although this is seems intuitively unlikely, Bataille et al. [72] estimated the same hobby farm to be the source of the transmission of the disease to the lower cluster. This is an important result to note, as it disagrees with the recognised idea that these small farms may not need to be culled as urgently as the bigger farms during epidemics.

5.9 Model assessment

In order to assess the goodness-of-fit of the models to the epidemiological data we used the final size of the epidemic (the number of farms to get infected and removed) as a summary statistic for a posterior predictive check. For each model we simulated 500 outbreaks, with locations the same as the farms in the Netherlands data, using values of the parameters drawn from their posterior densities. For each simulation we recorded the final size of the epidemic. These 500 values allowed us to estimate the posterior predictive distribution of the final size and to find where the observed final size fell in the distribution. Figure 5.16 shows the estimated posterior predictive distribution under the Time Dependent Distances model with a fixed latent period of 1 day. The observed final size is marked in red. It is clear that a large number of the simulations do not lead to large-scale outbreaks such as the one in the Netherlands as

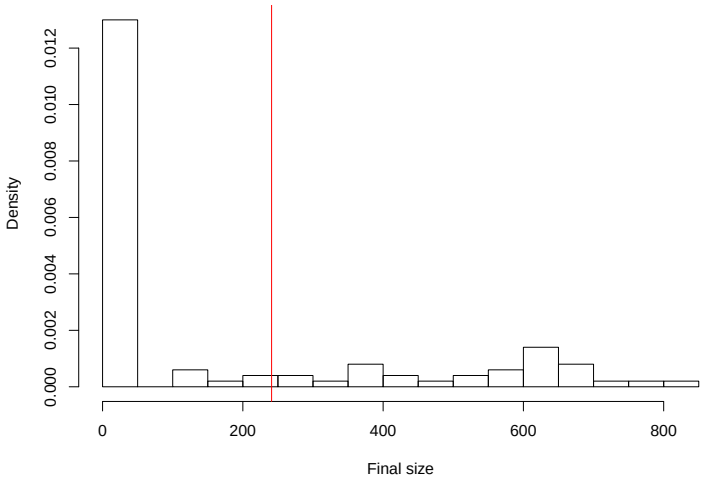
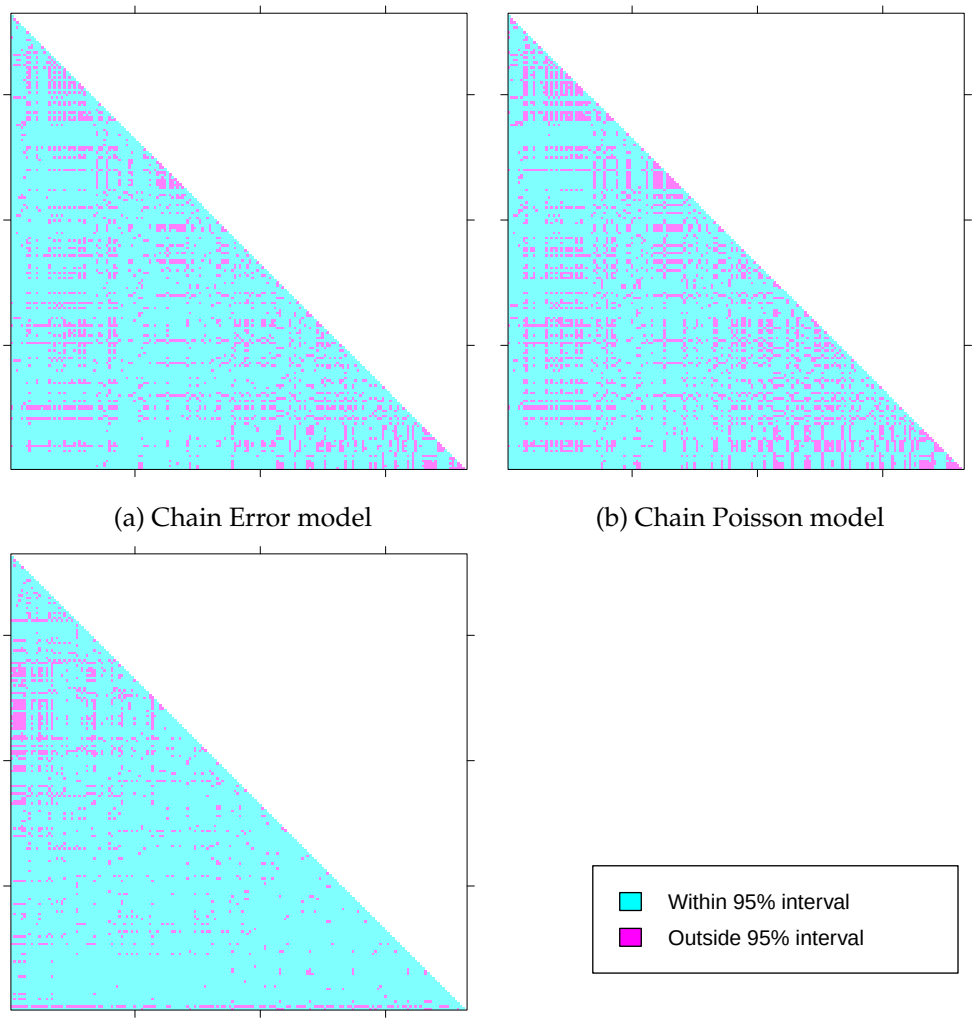


Figure 5.16: The estimated final size posterior predictive distribution for the Time Dependent Distances model with a fixed latent period of 1 day. The observed final size of the Netherland outbreak is marked in red.

half of the final sizes are under 20. However, the observed final size of 241 farms is still within the 95% highest density region of the posterior predictive distribution so there is no evidence against the model. We got similar results for each of the models with a fixed latent period of both 1 and 2 days. In each case we found that a large number of the outbreaks simulated had a small final size, but that the observed final size did not fall outside of the 95% highest density region of the posterior predictive distribution. Therefore there was no evidence against the fit of any of the models.

In order to assess the goodness-of-fit of the models to the genetic data a total of 1000 genetic distance matrices, $\tilde{\mathbf{\Psi}}$, were simulated for each model using values of the genetic parameters and values for the times and sources of infection events drawn from the posterior densities given by the output of the MCMC algorithm. If a particular observed distance is $\Psi_{i,j}$ then the posterior predictive p-value is defined as $\mathrm{P}(\Psi_{i,j} > \tilde{\Psi}_{i,j})$. Extreme p-values fall outside the 95% highest density region of the posterior predictive distribution and show that the model does not fit that distance well. We record the percentage of p-values that are not extreme over the whole genetic distance matrix in order to give a posterior predictive score for the model on that matrix. We can also provide a visual representation of how well the model fits the genetic distances matrix by plotting extreme p-values in one colour, and non-extreme p-values in another. Figures 5.17 and 5.18 give these matrix plots for each model, with fixed latent period of 1 day (figure 5.17) or 2 days (figure 5.18). The posterior predictive matrix

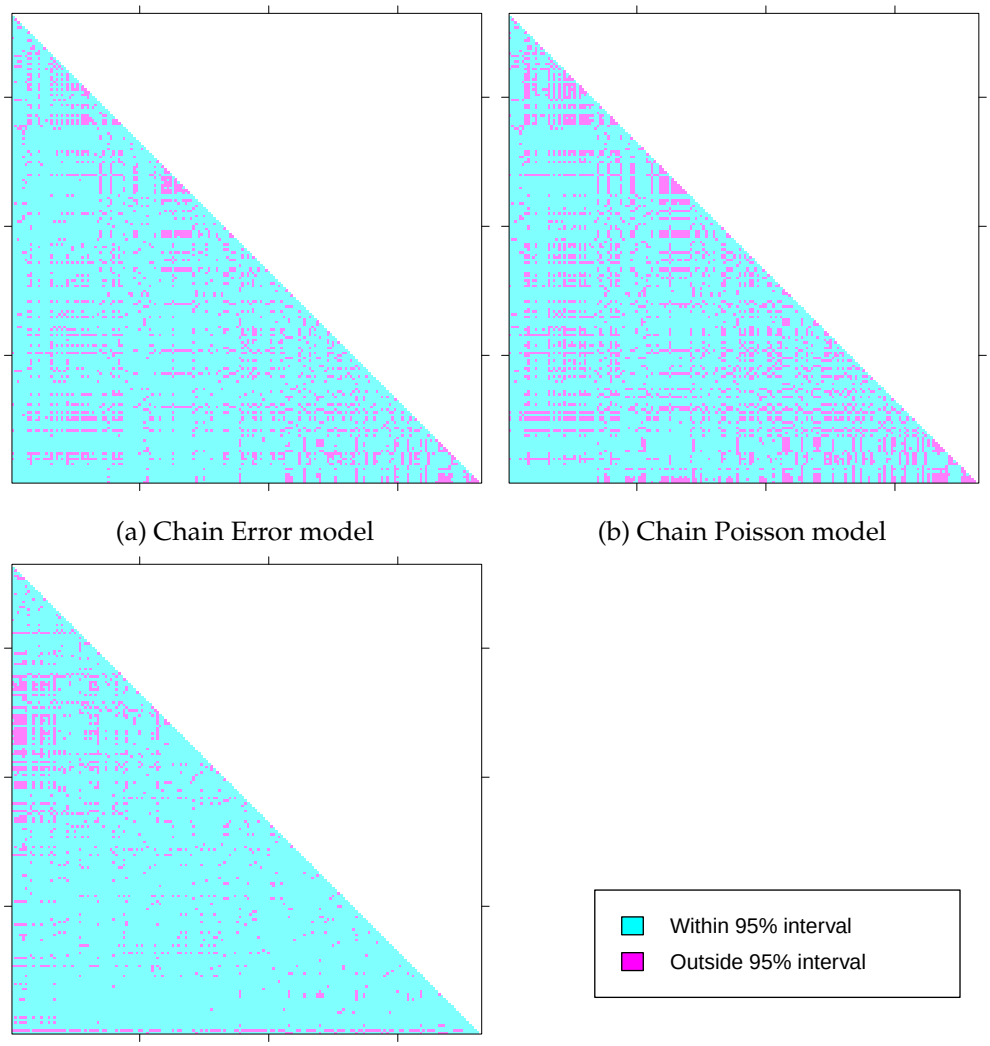


(a) Chain Error model

(b) Chain Poisson model

(c) Time Dependent Distances model

Figure 5.17: Posterior predictive checking to assess the fit of the three models with a fixed latent period of 1 day to the genetic distance matrix data. Blue cells indicate that the observed genetic distance fell within the 95% highest density region of the posterior predictive distribution and pink cells indicate that the observed genetic distance fell outside the interval. The layout of the matrices corresponds to the layout of the original genetic distance matrices, so the x-axis, left to right, runs from sequence 1 to sequence N, whereas the y-axis, from bottom to top, runs backwards from sequence N to sequence 1.



(a) Chain Error model

(b) Chain Poisson model

(c) Time Dependent Distances model

Figure 5.18: Posterior predictive checking to assess the fit of the three models with a fixed latent period of 2 days to the genetic distance matrix data. Blue cells indicate that the observed genetic distance fell within the 95% highest density region of the posterior predictive distribution and pink cells indicate that the observed genetic distance fell outside the interval. The layout of the matrices corresponds to the layout of the original genetic distance matrices, so the x-axis, left to right, runs from sequence 1 to sequence N, whereas the y-axis, from bottom to top, runs backwards from sequence N to sequence 1.

Latent period	Chain Error model	Chain Poisson model	Time Dependent Distances model
1 day	85.32%	80.05%	91.38%
2 days	86.50%	79.49%	91.35%

Table 5.5: The posterior predictive matrix scores of goodness-of-fit of the different models with different latent periods to the genetic data from the outbreak.

scores for each of the models and each of the latent periods are given in table 5.5. It is clear that all three of the models fit the data fairly well, but that the Time Dependent Distances model is the best fit for the genetic distance matrix.

5.10 Discussion

In this chapter we have adapted the novel models created in chapter 2 in order to analyse an epidemic of avian influenza that took place across farms in the Netherlands in 2003. We adapted the models to fit within a continuous time model for an epidemic with a spatial kernel and a single origin for the disease. Our models focused on utilising the information available in whole-genome sequence data in order to reconstruct the transmission tree and discover the dynamics of the disease spread.

These models were originally designed with regards to nosocomial pathogens, which tend to be modelled with multiple introductions of the disease to a much smaller sized population, but we felt that the simplicity of the model framework which simply uses the genetic distance between two pathogen samples instead of modelling each changing nucleotide in a long genetic sequence meant that it could be applied to a much broader range of diseases. This feature of the models also allowed us to simulate data from the model easily without having to simulate complicated genetic sequences, and therefore we were able to test the performance of the MCMC algorithm which we use to fit the models for a range of values of the parameters. We found that it performed well across the range of values which we could reasonably expect the parameters to take for these kinds of epidemics. The parameters, and the transmission trees, for the simulations were well recovered.

The analysis of the Netherlands avian influenza data under the three different models resulted in similar estimates for each of the parameters, even when we varied the latent period of the pathogen between 1 and 2 days. However, each model produced a different estimate for the transmission tree. The broad structure of the epidemic

transmission tree was estimated to be the same by each model: there were many transmission events inferred with high probability between farms which were close neighbours, and each model only inferred one transmission event which happened between the two distinct geographical clusters of farms. It has been shown previously by Bataille et al. [72] that it is highly likely that one long distance transmission event was the origin of the secondary outbreak of avian influenza in the lower cluster of farms. Although our estimates for the exact farms which were likely to have been the transmitter and receiver between the two clusters of farms tended to differ between models and depend on the length of the latent period, we did find that when the latent period was 2 days, the Chain Poisson model estimated the same farm as the source for this event as was estimated by Bataille et al. [72]. As we have discussed, the fact that we estimate a small hobby farm to be the source of a secondary outbreak of transmission has big consequences for epidemic control strategies.

In order to assess which model was the best fit for the data, we used the novel model assessment method introduced in chapter 3. This method showed that all three of the models were a reasonably good fit for the genetic data, but that the Time Dependent Distances model was the best fit.

5.10.1 Limitations and further work

There are, of course, some limitations to the models and methods that we have described here for analysing these data from the Netherlands outbreak. Some of these limitations stem from the nature of the data that were available to us. For example, we only had one genetic sequence from each farm from which to produce the genetic distances matrix which describes the genetic diversity between farms. However, each farm is likely to have hosted a variety of related variants of the pathogen, and the one consensus isolate which was sequenced from a sample of five birds was not guaranteed to be the dominant strain on that farm, and our model therefore has no mechanism for modelling within-host, or within-farm, genetic variation.

Although we consider the simplicity of just using the genetic distance between farms as our measure of genetic diversity as an advantage of our methods, there are ways in which we could broaden this without reintroducing the full genetic sequences. For example, we could use the numbers of transitions and transversions between sequences, as is done by Ypma et al. [19].

Another extension that could be made to the model framework would be to introduce a tailing-off of infectiousness after culling rather than a hard cut off point at the removal time. Ypma et al. [19] suggest that there may still be mechanisms for farms to infect other farms after the point at which they have been culled, through diseased particles remaining on the farm. There has been speculation that the method of transmission between farms may have been through the wind, or through contaminated vehicles moving between farms, which suggests that there would be potential for disease to be spread for a short time after a farm has been culled.

A model in which the transmissibility and susceptibility of farms is allowed to be heterogeneous may also be of interest. There are data available about the size and flock types of the farms which are currently not utilised by our model. This data could provide information about which types or size of farm are more likely to spread disease, and therefore which ones should be the priority targets for infection control strategies.

Notation used in Chapter 5

Model description	
N	Number of farms in the outbreak region
Ψ	Matrix of genetic distances
$\Psi_{i,j}$	Genetic distance between farms i and j
E_i	Exposure time of farm i
I_i	Infection time of farm i
S_i	Sampling time of farm i
R_i	Removal (culling) time of farm i
D_{ij}	Sum of consecutive distances in transmission chain between patients i and j $D_{ij} = \sum_{r=0}^{h-1} d_{p_r,p_{r+1}}$ where $p_0 = i$, $p_h = j$
$t_{i,j}$	Difference between the sampling times of farm i and farm j
$d_{i,j}$	Geographic distance between farm i and farm j
$\beta_{i,j}$	Infection pressure from farm i on farm j at E_j
P_i	Total infectious pressure on farm i at E_i
Model parameters	
ρ	Vector of parameters $\rho = \{\beta_0, \delta, \gamma, \Theta\}$
β_0	Transmission parameter
δ	Spatial parameter
γ	Removal parameter
Θ	Vector of genetic parameters
θ	Genetic diversity parameter for farms sharing a transmission event
θ_{gl}	Genetic diversity parameter for unrelated farms
α	Genetic chain error parameter for Chain Error model
Model inference and likelihood	
T	Unobserved transmission dynamics $T = \{\mathbf{I}, s, \Psi^a\}$
$\mathbf{I}$	Vector of unobserved infection times
s	Vector of unobserved sources of infection
Ψ^a	Matrix of unobserved genetic distances
n_{noseq}	Number of farms infected but no sequence sampled
n_{seq}	Number of sequences sampled from all infected farms
n_I	Total number of infected farms
n_R	Total number of removed farms
X	Vector of culling times
Z	Observed deterministic data: sampling times and number of susceptible farms
$\text{trans}(i,j)$	Number of transmission events between farms i and j

Notation used in Chapter 5

MCMC algorithm description

l	Initial infective
N_{par}	Number of (i,j) such that $\text{trans}(i,j) = 1$
N_{glo}	Number of (i,j) such that $\text{trans}(i,j) = \infty$
T^*	Candidate dataset proposed in augmented data update $T^* = \{\mathbf{I}^*, s^*, \Psi^a *\}$
q_{T,T^*}	Proposal ratio
	$q_{T,T^*} = P(T^* \to T)/(T \to T^*)$
O_i	Time at which i infects its first offspring

Simulation method

P	Length of fixed latent period
h	Number of days before culling that each farm is sampled
$\widetilde{\Psi}$	Set of simulated distance matrices

Chapter 6

Conclusions

This thesis aimed to develop new methods for incorporating whole-genome sequence data alongside epidemiological data into the analysis of epidemics. Such genetic data are becoming increasingly widely available as advances in technology drive down the cost and processing time involved in their collection. Analysing these data requires models for genetic diversity in sampled pathogen isolates. Many current methods either rely on complex micro-evolution models which model the process of genetic mutation and require high-dimensional data input, or on simplifying assumptions about the independence of the genetic distances within a transmission tree. The methods presented in chapter 2 use the number of single-nucleotide polymorphisms between pairs of sequences to measure genetic diversity, thus reducing the amount of data that needs to be stored and used to a single matrix of pairwise distances. We believe that this is a logical method, as these genetic distances are actually observed in the whole-genome sequence data and so we avoid having to make the many assumptions about the underlying processes that govern mutation that are necessary to model the micro-evolution process. Our models for the pairwise distances include dependence between sequences from individuals in the same chain of transmission, which other models overlook. This should lead to improved inference of transmission trees for outbreaks, which can inform decisions about prevention and control measures.

The stochastic epidemic model introduced in chapter 2, of which the genetic model is one part, has many advantages for modelling outbreaks. The model allows for each individual to have had multiple sequences taken, and includes a parameter for the within-host diversity observed between these sequences. Our model also allows for the pathogen to have been introduced into the population at more than one time, by more than one individual. The ability to infer a transmission forest rather than

a single transmission tree with one root is a valuable and unusual advantage of our method. The framework of our model is flexible and it could be of interest in the future to adapt it to include heterogeneity in the susceptibility or infectivity of different types of individuals, or to include a more complex model for the genetic distances which explicitly includes other forms of mutation such as recombination rather than just SNPs.

Recently there has been a focus on such methods which allow for the analysis of both epidemiological and genetic data, but methods for assessing the goodness-of-fit of the models used have not been developed. In chapter 3 we look at methods of model criticism for Bayesian epidemic models and attempt to expand these methods to allow for the assessment of how well a model fits the genetic data as well as the epidemiological data. It is important that there are tools available to assess the fit of the genetic part of a model as well as the epidemiological part in order to assess the goodness-of-fit of the model as a whole, which is necessary especially if the results are going to be used to inform public health procedures. Our method, which provides a 'posterior predictive matrix score' which is a percentage of the genetic distances which are 'well-fit', has been shown to allow for the most suitable model, out of a selection, to be found. It would be of interest to further develop this method of model fitting and potentially to establish a threshold which could determine ill-fitting models.

In chapters 4 and 5 we apply the models and methods developed in chapters 2 and 3 to data from two outbreaks of different pathogens in different populations. First, in chapter 4 we fit the new models to data from a nosocomial outbreak of MRSA. The discrete-time stochastic model with scope for multiple introductions and multiple sequences from the same individual is ideal for this hospital setting. A Bayesian inference scheme using a data-augmented MCMC algorithm allows us to infer unobserved colonisation times and sources and missing genetic data. Therefore the model can be used to take advantage of the information available in the WGS data to infer the pathways of transmission through each hospital ward in the study as well as the parameters of the model which tell us about transmission and importation probabilities and the sensitivity of the screening tests. Since there is currently considerable public health interest in antibiotic-resistance it is important that the dynamics of transmission of antibiotic-resistant pathogens such as MRSA are well understood so measures can be taken to prevent such transmission.

Using our novel model assessment techniques we showed that the genetic models introduced in chapter 2 would fit the data from this particular outbreak better if the Poisson distributions were replaced by either Geometric or Negative Binomial distributions due to their larger variances. This shows the importance of assessing the fit of a model to both the epidemiological and the genetic data. However, it would be of interest to develop a method for deciding, pre-analysis, which distribution would be the best fit for the particular data to be modelled.

Analysis of the data from the MRSA outbreak under our models led to the identification of a 'super-spreader' on each ward who was the source of a disproportionately large number of colonisations. It would be of interest to adapt the model to allow for heterogeneity of infectivity in order to investigate this further. One of these patients had three separate stays on the ward, and our model assumes that the patient's importation status for each admission was independent of their last stay. A natural development of our model would be to relax this assumption so that patients who left the ward colonised could return colonised with some probability.

In chapter 5 we used our genetic models to analyse WGS data from an outbreak of avian influenza in the Netherlands. Knowledge about the transmission dynamics of avian influenza is important for designing appropriate intervention strategies since it is difficult to prevent outbreaks due to the endemic carriage of the pathogen in wild flocks. The genetic models in this case formed part of a continuous-time stochastic epidemic model which included a spatial aspect. Fitting the models to data from this outbreak in a notably different sort of setting and population to the nosocomial pathogens discussed earlier allowed us to showcase the flexibility and wide-ranging applicability of our genetic models. Using our model assessment methods we showed that our models were a good fit for the data, and that the model which specifically includes time-dependence was the best fit. There is scope to develop our models further by allowing for heterogeneity of susceptibility and infectivity based on the type of birds kept on each farm, or on the size of the farm. These data are available for this outbreak. For future outbreak studies it would be of interest to collect more than one isolate per farm to allow for modelling of within-farm genetic diversity since our analysis had to assume that the one isolate taken was representative of the pathogen on the whole farm. As sequencing technology becomes faster, cheaper and more portable, it becomes more likely that it will be achievable to take multiple isolates for each infected farm in a future outbreak.

Overall, this thesis has developed methods and models for the analysis of both genetic and epidemiological data from an outbreak of a pathogen. These models are flexible and can be adapted to fit a wide range of settings and populations. We have also provided a novel method for assessing the goodness-of-fit of such models to the genetic data from an outbreak. We have demonstrated the capabilities of the models and the model assessment framework in two different contexts.

References

[1] A Kramer, A Manas, and M Kretzschmar. Principles of Infectious Disease Epidemiology. In A Kramer, M Kretzschmar, and K Krickeberg, editors, *Modern Infectious Disease Epidemiology*, chapter 5, pages 85–99. Springer, 2010. ISBN 978-1-4614-2507-6. doi: 10.1007/978-0-387-93835-6.

[2] N Bailey. *The Mathematical Theory of Infectious Diseases and its Applications. 2nd edition*. Charles Griffin & Company Limited, 1975. ISBN 0852642318. doi: 10.1111/j.1365-2586.2006.00647.x.

[3] N Becker. The Uses of Epidemic Models. *Biometrics*, 35:295–305, 1979. ISSN 0006341X. doi: 10.2307/2529951.

[4] M Kretzschmar and J Wallinga. Mathematical Models in Infectious Disease Epidemiology. In A Kramer, M Kretzschmar, and K Krickeberg, editors, *Modern Infectious Disease Epidemiology*, chapter 12, pages 209–221. Springer, 2010. doi: 10.1007/978-0-387-93835-6.

[5] H Andersson and T Britton. Stochastic epidemic models and their statistical analysis. *Springer Sceince and Business Media*, 2012. ISSN 0387950508. doi: 10.1007/978-1-4612-1158-7.

[6] C Robert. *The Bayesian Choice. From Decision-Theoretic Foundations to Computational Implementation*. Springer Texts in Statistics, 2006. ISBN 9780387715988. doi: 10.1007/0-387-71599-1.

[7] G Box and G Tiao. *Bayesian Inference in Statistical Analysis*. John Wiley & Sons, 1973. ISBN 0-471-57428-7. doi: 10.1002/9781118033197.

[8] W Gilks, S Richardson, and D Spiegelhalter. Introducing Markov Chain Monte Carlo. In W Gilks, S Richardson, and D Spiegelhalter, editors, *Markov chain Monte Carlo in practice*, pages 1–17. Chapman & Hall/CRC, 1998. ISBN 0-412-05551-1.

[9] W Hastings. Monte carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970. ISSN 00063444. doi: 10.1093/biomet/57.1.97.

[10] A Raftery and S Lewis. Implementing MCMC. In W Gilks, S Richardson, and D Spiegelhalter, editors, *Markov chain Monte Carlo in practice*. Chapman & Hall/CRC, 1996. ISBN 0-412-05551-1.

[11] A Gelfand and A Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 1990. ISSN 1537274X. doi: 10.1080/01621459.1990.10476213.

[12] A Gelfand, S Hills, A Racine-Poon, and A Smith. Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, 85:398–409, 1990. ISSN 1537274X. doi: 10.1080/01621459.1990.10474968.

[13] P O'Neill and G Roberts. Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 162 (1):121–129, 1999. ISSN 09641998. doi: 10.1111/1467-985X.00125.

[14] C Worby, P O'Neill, T Kypraios, J Robotham, D De Angelis, E Cartwright, S Peacock, and B Cooper. Reconstructing transmission trees for communicable diseases using densely sampled genetic data. *Annals of Applied Statistics*, 10(1):395–417, 2016. ISSN 19417330. doi: 10.1214/15-AOAS898.

[15] T Kypraios, P O'Neill, S Huang, S Rifas-Shiman, and B Cooper. Assessing the role of undetected colonization and isolation precautions in reducing Methicillin-Resistant Staphylococcus aureus transmission in intensive care units. *BMC Infectious Diseases*, 10:29, 2010. ISSN 14712334. doi: 10.1186/1471-2334-10-29.

[16] S Cauchemez, F Carrat, C Viboud, A Valleron, and P Boëlle. A Bayesian MCMC approach to study transmission of influenza: Application to household longitudinal data. *Statistics in Medicine*, 23:3469–3487, 2004. ISSN 02776715. doi: 10.1002/sim.1912.

[17] M Morelli, G Thébaud, J Chadœuf, D King, D Haydon, and S Soubeyrand. A Bayesian Inference Framework to Reconstruct Transmission Trees Using Epidemiological and Genetic Data. *PLoS Computational Biology*, 8(11):e1002768, 2012. ISSN 1553734X. doi: 10.1371/journal.pcbi.1002768.

[18] T Jombart, A Cori, X Didelot, S Cauchemez, C Fraser, and N Ferguson. Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. *PLoS Computational Biology*, 10(1):e1003457, 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003457.

[19] R Ypma, A Bataille, A Stegeman, G Koch, J Wallinga, and W van Ballegooijen. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proceedings of the Royal Society B: Biological Sciences*, 279 (1728):444–450, 2012. ISSN 0962-8452. doi: 10.1098/rspb.2011.0913.

[20] M Tanner and W Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–540, 1987. ISSN 1537274X. doi: 10.1080/01621459.1987.10478458.

[21] A Gelman, X Meng, and H Stern. Posterior Predictive Assessment of Model Fitness via Realized Discrepancies. *Statistica Sinica*, 6:733–807, 1996. ISSN 10170405. doi: 10.1.1.142.9951.

[22] X Meng. Posterior Predictive *p*-Values. *The Annals of Statistics*, 22:1142–1160, 1994. ISSN 0090-5364. doi: 10.1214/aos/1176325622.

[23] N Croucher, S Harris, Y Grad, and W Hanage. Bacterial genomes in epidemiology–present and future. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 368(1614):20120202, 2013. ISSN 1471-2970. doi: 10.1098/rstb.2012.0202.

[24] P Tang, M Croxen, M Hasan, W Hsiao, and L Hoang. Infection control in the new age of genomic epidemiology. *American Journal of Infection Control*, 45:170–197, 2017. ISSN 15273296. doi: 10.1016/j.ajic.2016.05.015.

[25] E van Dijk, Y Jaszczyszyn, D Naquin, and C Thermes. The Third Revolution in Sequencing Technology. *Trends in Genetics*, 34(9):666–681, 2018. ISSN 13624555. doi: 10.1016/j.tig.2018.05.008.

[26] S Dzidic and V Bedeković. Horizontal gene transfer-emerging multidrug resistance in hospital bacteria. *Acta pharmacologica Sinica*, 24:519–526, 2003. ISSN 1671-4083.

[27] A Lesk. *Introduction to Genomics*. Oxford University Press, second edition, 2012. ISBN 978-0-19-956435-4.

[28] Illumina Inc. Explore Illumina sequencing technology. Massively parallel sequencing with optimized SBS chemistry, 2019. URL

https://emea.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html.

[29] J Shendure and H Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 26: 1135–1145, 2008. ISSN 10870156. doi: 10.1038/nbt1486.

[30] J Quick, N Loman, S Duraffour, J Simpson, E Severi, L Cowley, J Bore, R Koundouno, G Dudas, A Mikhail, N Ouédraogo, B Afrough, A Bah, J Baum, B Becker-Ziaja, J Boettcher, M Cabeza-Cabrerizo, Á Camino-Sánchez, L Carter, J Doerrbecker, T Enkirch, I Garciá-Dorival, N Hetzelt, J Hinzmann, T Holm, L Kafetzopoulou, M Koropogui, A Kosgey, E Kuisma, C Logue, A Mazzarelli, S Meisel, M Mertens, J Michel, D Ngabo, K Nitzsche, E Pallasch, L Patrono, J Portmann, J Repits, N Rickett, A Sachse, K Singethan, I Vitoriano, R Yemanaberhan, E Zekeng, T Racine, A Bello, A Sall, O Faye, O Faye, N Magassouba, C. Williams, V Amburgey, L Winona, E Davis, J Gerlach, F Washington, V Monteil, M Jourdain, M Bererd, A Camara, H Somlare, A Camara, M Gerard, G Bado, B Baillet, D Delaune, K Nebie, A Diarra, Y Savane, R Pallawo, G Gutierrez, N Milhano, I Roger, C Williams, F Yattara, K Lewandowski, J Taylor, P Rachwal, D Turner, G Pollakis, J Hiscox, D Matthews, M O'Shea, A Johnston, D Wilson, E Hutley, E Smit, A Di Caro, R Wolfel, K Stoecker, E Fleischmann, M Gabriel, S Weller, L Koivogui, B Diallo, S Keita, A Rambaut, P Formenty, S Gunther, and M Carroll. Real-time, portable genome sequencing for Ebola surveillance. *Nature*, 530:228–232, 2016. ISSN 14764687. doi: 10.1038/nature16996.

[31] N Faria, E Sabino, M Nunes, L Alcantara, N Loman, and O Pybus. Mobile real-time surveillance of Zika virus in Brazil. *Genome Medicine*, 8:97, 2016. ISSN 1756994X. doi: 10.1186/s13073-016-0356-2.

[32] D Loughran and J Harrison. Antibiotic resistance: A long term, serious problem...getting worse. Thoughts on the future of surgery in a post-antibiotic era. *European Surgery - Acta Chirurgica Austriaca*, 46(2):55–56, 2014. ISSN 16824016. doi: 10.1007/s10353-014-0253-0.

[33] Centers for Disease Control and Prevention. Antibiotic/Antimicrobial Resistance, 2018. URL https://www.cdc.gov/drugresistance/index.html.

[34] R Norrby, M Powell, B Aronsson, D Monnet, I Lutsat, I Bocsan, O Cars, H Giamarellou, and I Gyssens. The Bacterial Challenge : Time to React. Technical report, European Centre for Disease Prevention and Control and European Medicines Agency, 2009. URL

https://ecdc.europa.eu/en/publications-data/ecdcemea-joint-technical-report-bacterial-challenge-time-react.

[35] R Laxminarayan, A Duse, C Wattal, A Zaidi, H Wertheim, N Sumpradit, E Vlieghe, G Hara, I Gould, H Goossens, C Greko, A So, M Bigdeli, G Tomson, W Woodhouse, E Ombaka, A Peralta, F Qamar, F Mir, S Kariuki, Z Bhutta, A Coates, R Bergstrom, G Wright, E Brown, and O Cars. Antibiotic resistance-the need for global solutions. *The Lancet Infectious Diseases*, 13:1057–1098, 2013. ISSN 14733099. doi: 10.1016/S1473-3099(13)70318-9.

[36] P Pumart, T Phodha, V Thamlikitkul, A Riewpaiboon, P Prakongsai, and S Limwattananon. Health and economic impacts of antimicrobial resistance in Thailand. *Journal of Health Services Research & Policy*, 6:352–360, 2012.

[37] M Honigsbaum. Superbugs and us. *The Lancet*, 391:420, 2018. doi: 10.1016/S0140-6736(18)30110-7.

[38] M.J.M Bonten and M.C.J Bootsma. Nosocomial Transmission: Methicillin-Resistant Staphylococcus aureus (MRSA). In A Kramer, M Kretzschmar, and K Krickeberg, editors, *Modern Infectious Disease Epidemiology*, pages 395–407. Springer, 2010. ISBN 978-1-4614-2507-6. doi: 10.1007/978-0-387-93835-6.

[39] J Lindsay and M Holden. Staphylococcus aureus: Superbug, super genome? *Trends in Microbiology*, 12(8):378–385, 2004. ISSN 0966842X. doi: 10.1016/j.tim.2004.06.004.

[40] J Kluytmans, A Van Belkum, and H Verbrugh. Nasal carriage of Staphylococcus aureus: Epidemiology, underlying mechanisms, and associated risks. *Clinical Microbiology Reviews*, 10:85–99, 1997. ISSN 08938512. doi: 10.1007/s15010-005-4012-9.

[41] R Williams. Healthy carriage of Staphylococcus aureus: its prevalence and importance. *Bacteriological reviews*, 27(96):56–71, 1963. ISSN 0005-3678.

[42] L Mermel, J Cartony, P Covington, G Maxey, and D Morse. Methicillin-resistant Staphylococcus aureus colonization at different body sites: A prospective, quantitative analysis. *Journal of Clinical Microbiology*, 49:1119–1121, 2011. ISSN 00951137. doi: 10.1128/JCM.02601-10.

[43] S Blot, K Vandewoude, E Hoste, and F Colardyn. Outcome and attributable mortality in critically ill patients with bacteremia involving methicillin-susceptible

and methicillin-resistant Staphylococcus aureus. *Archives of Internal Medicine*, 162:2229–2235, 2002. ISSN 00039926. doi: 10.1001/archinte.162.19.2229.

[44] S Cosgrove, G Sakoulas, E Perencevich, M Schwaber, A Karchmer, and Y Carmeli. Comparison of mortality related to methicillin-resistant and methicillin-susceptible Staphylococcus aureus bacteremia: a metaanalysis. *Clinical Infectious Diseases*, 36:53–59, 2003. ISSN 1537-6591. doi: doi:10.1086/345476.

[45] S Cosgrove, Y Qi, K Kaye, S Harbarth, A Karchmer, and Y Carmeli. The Impact of Methicillin Resistance in Staphylococcus aureus Bacteremia on Patient Outcomes: Mortality, Length of Stay, and Hospital Charges. *Infection Control & Hospital Epidemiology*, 26:166–174, 2005. ISSN 0899-823X. doi: 10.1086/502522.

[46] D Venkatesh, M Poen, T Bestebroer, R Scheuer, O Vuong, M Chkhaidze, A Machablishvili, J Mamuchadze, L Ninua, N Fedorova, R Halpin, X Lin, A Ransier, T Stockwell, D Wentworth, D Kriti, J Dutta, H van Bakel, A Puranik, M Slomka, S Essen, I Brown, R Fouchier, and N Lewis. Avian Influenza Viruses in Wild Birds: Virus Evolution in a Multihost Ecosystem. *Journal of Virology*, 92(15):1–20, 2018. ISSN 0022-538X. doi: 10.1128/JVI.00433-18.

[47] T Ulrichs. Airborne Transmission: Influenza and Tuberculosis. In A Kramer, M Kretzschmar, and K Krickeberg, editors, *Modern Infectious Disease Epidemiology*, pages 279–290. Springer, 2010. ISBN 978-1-4614-2507-6. doi: 10.1007/978-0-387-93835-6.

[48] D Alexander. A review of avian influenza in different bird species. In *Veterinary Microbiology*, 2000. ISBN 4419323574. doi: 10.1016/S0378-1135(00)00160-7.

[49] D Alexander and I Brown. History of highly pathogenic avian influenza. *Revue Scientifique et Technique de l'OIE*, 28(1):19–38, 2009. ISSN 0253-1933. doi: 10.20506/rst.28.1.1856.

[50] D Xiang, Z Pu, T Luo, F Guo, X Li, X Shen, D Irwin, R Murphy, M Liao, and Y Shen. Evolutionary dynamics of avian influenza A H7N9 virus across five waves in mainland China, 2013–2017. *Journal of Infection*, 77(3):205–211, 2018. ISSN 15322742. doi: 10.1016/j.jinf.2018.05.006.

[51] S Su, M Gu, D Liu, J Cui, G Gao, J Zhou, and X Liu. Epidemiology, Evolution, and Pathogenesis of H7N9 Influenza Viruses in Five Epidemic Waves since 2013 in China. *Trends in Microbiology*, 25(9):713–728, 2017. ISSN 18784380. doi: 10.1016/j.tim.2017.06.008.

[52] C Jewell, T Kypraios, R Christley, and G Roberts. A novel approach to real-time risk prediction for emerging infectious diseases: A case study in Avian Influenza H5N1. *Preventive Veterinary Medicine*, 91:19–28, 2009. ISSN 01675877. doi: 10.1016/j.prevetmed.2009.05.019.

[53] E Cottam, G Thebaud, J Wadsworth, J Gloster, L Mansley, D Paton, D King, and D Haydon. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proceedings of the Royal Society B: Biological Sciences*, 275(1637):887–895, 2008. ISSN 0962-8452. doi: 10.1098/rspb.2007.1442.

[54] E Numminen, C Chewapreecha, J Siren, C Turner, P Turner, S Bentley, and J Corander. Two-phase importance sampling for inference about transmission trees. *Proceedings of the Royal Society B: Biological Sciences*, 281(1794):20141324, 2014. ISSN 1471-2954 (Electronic). doi: 10.1098/rspb.2014.1324.

[55] M Hall, M Woolhouse, and A Rambaut. Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set. *PLoS Computational Biology*, 11(12):e1004613, 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004613.

[56] N Mollentze, L Nel, S Townsend, K le Roux, K Hampson, D Haydon, and S Soubeyrand. A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. *Proceedings. Biological sciences / The Royal Society*, 281(1782):20133251, 2014. ISSN 1471-2954. doi: 10.1098/rspb.2013.3251.

[57] A Drummond and A Rambaut. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7(1):214, 2007. ISSN 1471-2148. doi: 10.1186/1471-2148-7-214.

[58] T Jukes and C Cantor. Evolution of Protein Molecules. *Mammalian Protein Metabolism*, pages 21–132, 1969. ISSN 0022-2143. doi: 10.1016/B978-1-4832-3211-9.50009-7.

[59] T Jombart, R Eggo, P Dodd, and F Balloux. Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity*, 106(2):383–390, 2011. ISSN 0018-067X. doi: 10.1038/hdy.2010.78.

[60] C Cespedes, B Said-Salim, M Miller, S Lo, B Kreiswirth, R Gordon, P Vavagiakis, R Klein, and F Lowy. The clonality of Staphylococcus aureus nasal carriage. *Journal of Infectious Diseases*, 191:444–452, 2005. ISSN 0022-1899. doi: 10.1086/427240.

[61] K Mongkolrattanothai, B Gray, P Mankin, A Stanfill, R Pearl, L Wallace, and R Vegunta. Simultaneous carriage of multiple genotypes of Staphylococcus aureus in children. *Journal of Medical Microbiology*, 60:317–322, 2011. ISSN 00222615. doi: 10.1099/jmm.0.025841-0.

[62] O Pybus, C Fraser, and A Rambaut. Evolutionary epidemiology: preparing for an age of genomic plenty. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, 368(1614):20120193–20120193, 2013.

[63] S Stefani, D Ryeon, J Lindsay, A Friedrich, A Kearns, H Westh, and F Mackenzie. Meticillin-resistant Staphylococcus aureus (MRSA): global epidemiology and harmonisation of typing methods. *International Journal of Antimicrobial Agents*, 39(4):273–282, 2012. ISSN 0924-8579. doi: 10.1016/j.ijantimicag.2011.09.030.

[64] A Stegeman, A Bouma, A Elbers, M de Jong, G Nodelijk, F de Klerk, G Koch, and M van Boven. Avian Influenza A Virus (H7N7) Epidemic in The Netherlands in 2003: Course of the Epidemic and Effectiveness of Control Measures. *Journal of Infectious Diseases*, 190:2088–2095, 2004. ISSN 0022-1899. doi: 10.1086/425583.

[65] M Koopmans, B Wilbrink, M Conyn, G Natrop, H van der Nat, H Vennema, A Meijer, J van Steenbergen, R Fouchier, A Osterhaus, and A Bosman. Transmission of H7N7 avian influenza A virus to human beings during a large outbreak in commercial poultry farms in the Netherlands. *The Lancet*, 363(9409):587–593, 2004. ISSN 1474-547X. doi: 10.1016/S0140-6736(04)15589-X.

[66] G Smith, D Vijaykrishna, J Bahl, S Lycett, M Worobey, O Pybus, S Ma, C Cheung, J Raghwani, S Bhatt, J Peiris, Y Guan, and A Rambaut. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature*, 459 (7250):1122–1125, 2009. ISSN 0028-0836. doi: 10.1038/nature08182.

[67] N Naffakh and S van der Werf. April 2009: an outbreak of swine-origin influenza A(H1N1) virus with evidence for human-to-human transmission. *Microbes and Infection*, 11(8-9):725–728, 2009. ISSN 12864579. doi: 10.1016/j.micinf.2009.05.002.

[68] M Lau, G Marion, G Streftaris, and G Gibson. A Systematic Bayesian Integration of Epidemiological and Genetic Data. *PLoS Computational Biology*, 11:1–27, 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004633.

[69] C Worby. *Statistical inference and modelling for nosocomial infections and the incorporation of whole genome sequence data*. PhD thesis, University of Nottingham, 2013.

[70] D Eyre, T Golubchik, N Gordon, R Bowden, P Piazza, E Batty, C Ip, D Wilson, X Didelot, L O'Connor, R Lay, D Buck, A Kearns, A Shaw, J Paul, M Wilcox, P Donnelly, T Peto, A Walker, and D Crook. A pilot study of rapid benchtop sequencing of Staphylococcus aureus and Clostridium difficile for outbreak detection and surveillance. *BMJ Open*, 2(3):e001124–e001124, 2012. ISSN 2044-6055. doi: 10.1136/bmjopen-2012-001124.

[71] S Harris, E Cartwright, M Török, M Holden, N Brown, A Ogilvy-Stuart, M Ellington, M Quail, S Bentley, J Parkhill, and S Peacock. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study. *The Lancet Infectious Diseases*, 13(2):130–136, 2013. ISSN 14733099. doi: 10.1016/S1473-3099(12)70268-2.

[72] A Bataille, F van der Meer, A Stegeman, and G Koch. Evolutionary analysis of inter-farm transmission dynamics in a highly pathogenic avian influenza epidemic. *PLoS pathogens*, 7(6):e1002094, 2011. ISSN 1553-7374. doi: 10.1371/journal.ppat.1002094.

[73] G Gibson, G Streftaris, and D Thong. Comparison and Assessment of Epidemic Models. *Statistical Science*, 33(1):19–33, 2018. ISSN 0883-4237. doi: 10.1214/17-STS615.

[74] M Alharthi. *Bayesian Model Assessment for Stochastic Epidemic Models*. PhD thesis, University of Nottingham, 2016.

[75] S Tong, M Holden, E Nickerson, B Cooper, A Cori, T Jombart, S Cauchemez, C Fraser, V Wuthiekanun, J Thaipadungpanit, M Hongsuwan, N Day, D Limmathurotsakul, J Parkhill, and S Peacock. Genome sequencing defines phylogeny and spread of methicillin-resistant Staphylococcus aureus in a high transmission setting. *Genome Research*, 25:111–118, 2015. doi: 10.1101/gr.174730.114.Freely.

[76] S Harris, E Feil, M Holden, M Quail, E Nickerson, N Chantratita, S Gardete, A Tavares, N Day, J Lindsay, J Edgeworth, H de Lencastre, J Parkhill, S Peacock, and S Bentley. Evolution of MRSA During Hospital Transmission and Intercontinental Spread. *Science*, 327(5964):469–474, 2010. ISSN 0036-8075. doi: 10.1126/science.1182395.

[77] G Boender, T Hagenaars, A Bouma, G Nodelijk, A Elbers, M de Jong, and M van Boven. Risk Maps for the Spread of Highly Pathogenic Avian Influenza

in Poultry. *PLoS Computational Biology*, 3(4):e71, 2007. ISSN 1553-734X. doi: 10.1371/journal.pcbi.0030071.

[78] V Bavinck, A Bouma, M van Boven, M Bos, E Stassen, and J Stegeman. The role of backyard poultry flocks in the epidemic of highly pathogenic avian influenza virus (H7N7) in the Netherlands in 2003. *Preventive Veterinary Medicine*, 88(4): 247–254, 2009. ISSN 01675877. doi: 10.1016/j.prevetmed.2008.10.007.

[79] X Zhang, M Friedl, and C Schaaf. Intra- and interspecies transmission of H7N7 highly pathogenic avian influenza virus during the avian influenza epidemic in The Netherlands in 2003. *Journal of Geophysical Research*, 111(1):333–340, 2006. ISSN 0253-1933.

[80] M Bos, M Van Bovena, M Nielena, A Boumaa, A Elbersc, G Nodelijkb, G Koch, A Stegemana, and M De Jong. Estimating the day of highly pathogenic avian influenza (H7N7) virus introduction into a poultry flock based on mortality data. *Veterinary Research*, 38:493–504, 2007. ISSN 09284249. doi: 10.1051/vetres:2007008.

[81] A Le Menach, E Vergu, R Grais, D Smith, and A Flahault. Key strategies for reducing spread of avian influenza among commercial poultry holdings: lessons for transmission to humans. *Proceedings of the Royal Society B: Biological Sciences*, 273:2467–2475, 2006. ISSN 0962-8452. doi: 10.1098/rspb.2006.3609.

[82] G Boender, T Hagenaars, A Bouma, G Nodelijk, A Elbers, M de Jong, and M van Boven. Risk Maps for the Spread of Highly Pathogenic Avian Influenza in Poultry. *PLoS Computational Biology*, 3(4):e71, 2007. ISSN 1553-734X. doi: 10.1371/journal.pcbi.0030071.

[83] N Ferguson, C Donnelly, and R Anderson. The Foot-and-Mouth Epidemic in Great Britain: Pattern of Spread and Impact of Interventions. *Science*, 292(April), 2001. doi: 10.1126/science.1061020.

[84] M Kendall, D Ayabina, and C Colijn. Estimating transmission from genetic and epidemiological data: a metric to compare transmission trees. *Statistical Science*, 33(1):70–85, 2016. ISSN 0883-4237. doi: 10.1214/17-STS637.

[85] J van der Goot, M de Jong, G Koch, and M Van Boven. Comparison of the transmission characteristics of low and high pathogenicity avian influenza A virus (H5N2). *Epidemiology & Infection*, 131(2):1003–1013, 2003.

Appendix A

Proposal ratios for augmented data step in the MCMC algorithm for MRSA models

A.1 Add colonisation

A.1.1 Add importation

In this move the proposal ratio is for adding an importation is

$$q_{T,T^*} = \frac{n_{sus} - n_{add}}{w(1 + n_{add_0})Y_{add}}.$$

This is because the probability of proposing this move is equal to:

- the probability of choosing the particular susceptible patient from the total number available to be picked (all susceptible patients, n_{sus}, minus those already with an added colonisation time, n_{add}), which is $1/(n_{sus} - n_{add})$,
- multiplied by the probability of assigning the patient as an importation, w,
- multiplied by the probability of picking the set of genetic distances, Y_{add}, where

$$Y_{add} = \prod_{j=1}^{n_{seqs}+n_{noseqs}+n_{add}} \mathrm{P}\left[\Psi_{i*,j} = \Psi^{a*}_{i*,j}|\Theta\right]$$

and the probability of proposing the reverse move is simply the probability of picking this patient from the set of patients with an added colonisation but no offspring which is $1/(1 + nadd_0)$ where the 1 is to account for the patient which we are proposing to add in this move.

A.1.2 Add acquisition

The proposal ratio for adding an acquisition is

$$q_{T,T^*} = \frac{C(t_i^{c*})(n_{sus} - n_{add})(t_i^d - t_i^a + 1)}{(1-w)(n_{add_0} + 1)Y_{add}}.$$

Here the probability of proposing this move is:

- the probability of choosing the particular susceptible patient from the total number available to be picked (all susceptible patients, n_{sus}, minus those already with an added colonisation time, n_{add}), which is $1/(n_{sus} - n_{add})$,
- multiplied by the probability of assigning the patient as an acquisition, $1 - w$,
- multiplied by the probability of choosing the particular day of colonisation, t_i^{c*}, from the set of days that the patient is on the ward, which is $1/(t_i^d - t_i^a + 1)$,
- multiplied by the probability of choosing the source for this patient's colonisation from the set of patients colonised on the chosen day, t_i^{c*}, which is $1/C(t_i^{c*})$,
- multiplied by the probability of picking the set of genetic distances, Y_{add}, where

$$Y_{add} = \prod_{j=1}^{n_{seqs}+n_{noseqs}+n_{add}} \mathrm{P}\left[\Psi_{i*,j} = \Psi_{i*,j}^{a*} | \Theta\right].$$

The probability of proposing the reverse move is simply the probability of choosing this particular patient to have their colonisation removed from the set of patients with an added colonisation but no offspring which is $1/(1 + nadd_0)$ where the 1 is to account for the patient which we are proposing to add in this move.

A.2 Remove colonisation

A.2.1 Removing an importation

In this move the proposal ration for removing a patient who was assumed to be an importation of the pathogen is

$$q_{T,T^*} = \frac{n_{add_0} \cdot w \cdot Y_{rm}}{n_{sus} - n_{add} + 1}.$$

Here the probability of proposing this move is the probability of picking this patient from the set of patients with an added colonisation but no offspring which is $1/(nadd_0)$. The probability of proposing the reverse move, which is the probability of proposing to add this patient as an importation, is

- the probability of choosing the particular susceptible patient from the total number available to be picked (all susceptible patients, n_{sus}, minus those already with an added colonisation time, n_{add}), which is $1/(n_{sus} - n_{add} + 1)$ where the 1 accounts for the patient which we are proposing to remove in this step,
- multiplied by the probability of assigning the patient as an importation, w,
- multiplied by the probability of picking the set of genetic distances, Y_{rm}, where

$$Y_{rm} = \prod_{j=1}^{n_{seqs}+n_{noseqs}+n_{add}} \mathrm{P}\left[\Psi_{i,j} = \Psi_{i,j}^{a*}|\Theta\right].$$

A.2.2 Removing an acquisition

The proposal ratio for removing the colonisation time of a patient who was assumed to be an acquisition is

$$q_{T,T^*} = \frac{Y_{rm} \cdot n_{add_0} \cdot (1-w)}{(t_i^d - t_i^a + 1)(n_{sus} - n_{add} + 1)(C(t_i^c) - 1)}.$$

Here the probability of proposing this move is the probability of picking this patient from the set of patients with an added colonisation but no offspring which is $1/(nadd_0)$. The probability of proposing the reverse move, which is the probability of proposing to add this patient as an acquisition, is

- the probability of choosing the particular susceptible patient from the total number available to be picked (all susceptible patients, n_{sus}, minus those already with an added colonisation time, n_{add}), which is $1/(n_{sus} - n_{add} + 1)$ where the 1 accounts for the patient which we are proposing to remove in this step,
- multiplied by the probability of assigning the patient as an acquisition, $1 - w$,
- multiplied by the probability of choosing the particular day of colonisation, t_i^c, from the set of days that the patient is on the ward, which is $1/(t_i^d - t_i^a + 1)$,
- multiplied by the probability of choosing the source for this patient's colonisation from the set of patients colonised on the chosen day, t_i^c, which is $1/C(t_i^c - 1)$ where the -1 accounts for the patient who's colonisation time we are proposing to remove,
- multiplied by the probability of picking the set of genetic distances, Y_{rm}, where

$$Y_{rm} = \prod_{j=1}^{n_{seqs}+n_{noseqs}+n_{add}} \mathrm{P}\left[\Psi_{i,j} = \Psi_{i,j}^{a*}|\Theta\right].$$

A.3 Moving a colonisation time

A.3.1 Moving an acquisition that remains an acquisition

If we propose to change the colonisation time of a patient that was previously colonised whilst on the ward to another time at which they may by colonised on the ward, then the proposal ratio is

$$q_{T,T^*} = \frac{C(t_i^{c*})}{C(t_i^c)}.$$

The probability of proposing this move is

- the probability of choosing the particular colonised patient from the set of all colonised patients, which is $1/(n_{seqs} + n_{noseq} + n_{add})$,
- multiplied by the probability of assigning the patient as an acquisition, $(1 - w)$,
- multiplied by the probability of choosing a colonisation time, t_i^{C*}, from the days that the patient was on the ward, which is $1/(t_i^d - t_i^a + 1)$,
- multiplied by the probability of choosing the source of colonisation from those available on the new colonisation day, t_i^{c*}, which is $1/C(t_i^{c*})$.

The probability of proposing the reverse move is

- the probability of choosing the particular colonised patient from the set of all colonised patients, which is $1/(n_{seqs} + n_{noseq} + n_{add})$,
- multiplied by the probability of assigning the patient as an acquisition, $(1 - w)$,
- multiplied by the probability of choosing a colonisation time, t_i^c, from the days that the patient was on the ward, which is $1/(t_i^d - t_i^a + 1)$,
- multiplied by the probability of choosing the source of colonisation from those available on the colonisation day, t_i^c, which is $1/C(t_i^c)$.

The probability of choosing the patient, the acquisition probability and the probability of choosing the day cancel to leave $q_{T,T^*} = \frac{C(t_i^{c*})}{C(t_i^c)}$.

A.3.2 Reassigning an acquisition as an importation

If we propose to reassign a patient who was previously colonised whilst on the ward as an importation of the pathogen to the ward, then the proposal ratio is

$$q_{T,T^*} = \frac{1 - w}{w(f_i - t_i^a + 1)C(t_i^c)}.$$

The probability of proposing this move is

- the probability of choosing the particular colonised patient from the set of all colonised patients, which is $1/(n_{seqs} + n_{noseq} + n_{add})$,
- multiplied by the probability of assigning the patient as an importation, w.

The probability of proposing the reverse move of changing an importation patient to an acquisition is

- the probability of choosing the particular colonised patient from the set of all colonised patients, which is $1/(n_{seqs} + n_{noseq} + n_{add})$,
- multiplied by the probability of assigning the patient as an acquisition, $1 - w$,
- multiplied by the probability of choosing the day of the patient's colonisation, t_i^c, from the set of days between their admission, t_i^a, and their last possible susceptible day, f_i, which is $1/(f_i - t_i^a + 1)$,
- multiplied by the probability of choosing the source of the patient's colonisation from the set of colonised patients present on day t_i^c which is $1/C(t_i^c)$.

The probabilities of choosing the patient cancel in the proposal ratio.

A.3.3 Reassigning an importation as an acquisition

If we proposed to reassign a patient who was previously an importation as an acquisition the proposal ratio is

$$q_{T,T^*} = \frac{w \cdot (f_i - t_i^a + 1) \cdot C(t_i^{c*})}{1 - w}.$$

The probability of proposing this move is

- the probability of choosing the particular colonised patient from the set of all colonised patients, which is $1/(n_{seqs} + n_{noseq} + n_{add})$,
- multiplied by the probability of assigning the patient as an acquisition, $1 - w$,
- multiplied by the probability of choosing the day of the patient's colonisation, t_i^{c*}, from the set of days between their admission, t_i^a, and their last possible susceptible day, f_i, which is $1/(f_i - t_i^a + 1)$,
- multiplied by the probability of choosing the source of the patient's colonisation from the set of colonised patients present on day t_i^{c*} which is $1/C(t_i^{c*})$.

The probability of proposing the reverse move is

- the probability of choosing the particular colonised patient from the set of all colonised patients, which is $1/(n_{seqs} + n_{noseq} + n_{add})$,
- multiplied by the probability of assigning the patient as an importation, w.

The probabilities of choosing the patient cancel in the proposal ratio.

A.4 Changing a patient's genetic distances

The proposal ratio for changing a patients genetic distances is

$$q_{T,T^*} = \frac{\prod\limits_{j:\, i \neq j} \mathrm{P}\left[\Psi_{i,j} = \Psi^{a}_{i,j} | \Theta\right]}{\prod\limits_{j:\, i \neq j} \mathrm{P}\left[\Psi_{i,j} = \Psi^{a*}_{i,j} | \Theta\right]}$$

which is simply the probability of choosing the current genetic distances divided by the probability of choosing the proposed genetic distances. The probability of choosing the particular patient is the same for both this move and the reverse move and so cancels.

Appendix B

Traceplots from the MCMC algorithm output for each of the three Poisson-based models on Ward 1 of the Thai data

B.1 Chain Error model

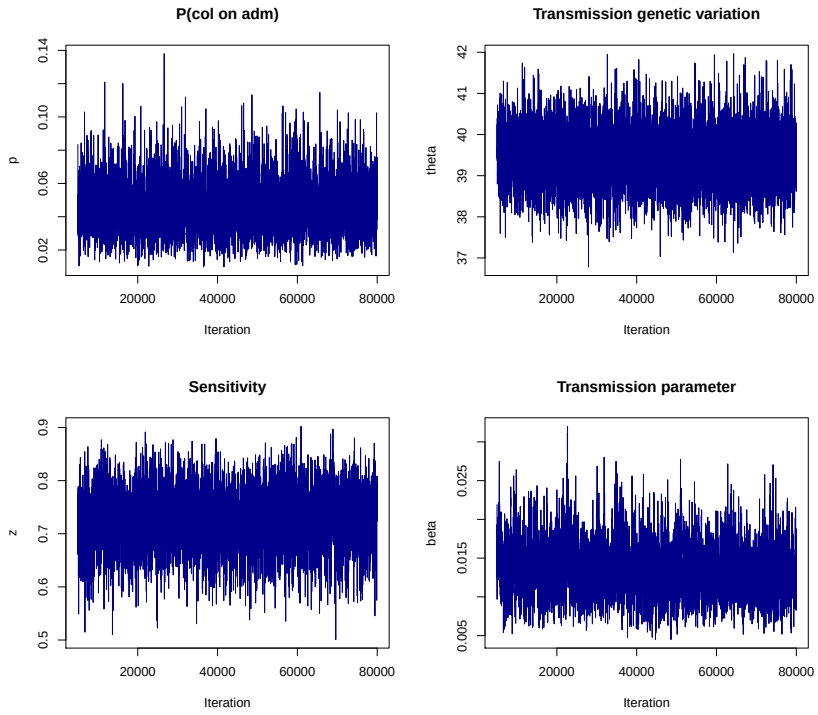


Figure B.1: Traceplots of estimates of parameters p, z, θ and β under the Chain Error model.

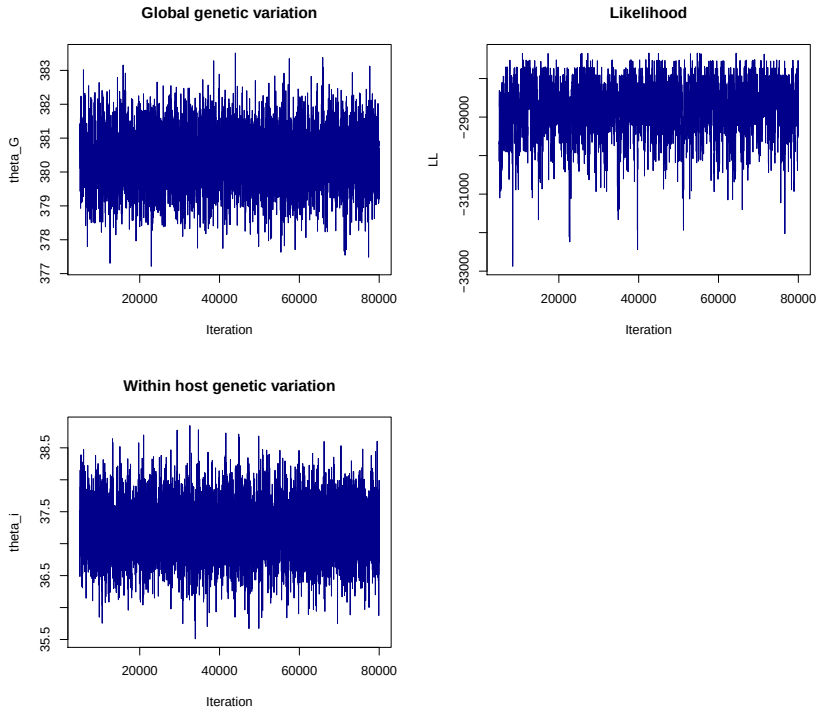


Figure B.2: Traceplots of estimates of parameters θ_{gl}, θ_i and the log likelihood under the Chain Error model.

B.2 Chain Poisson model

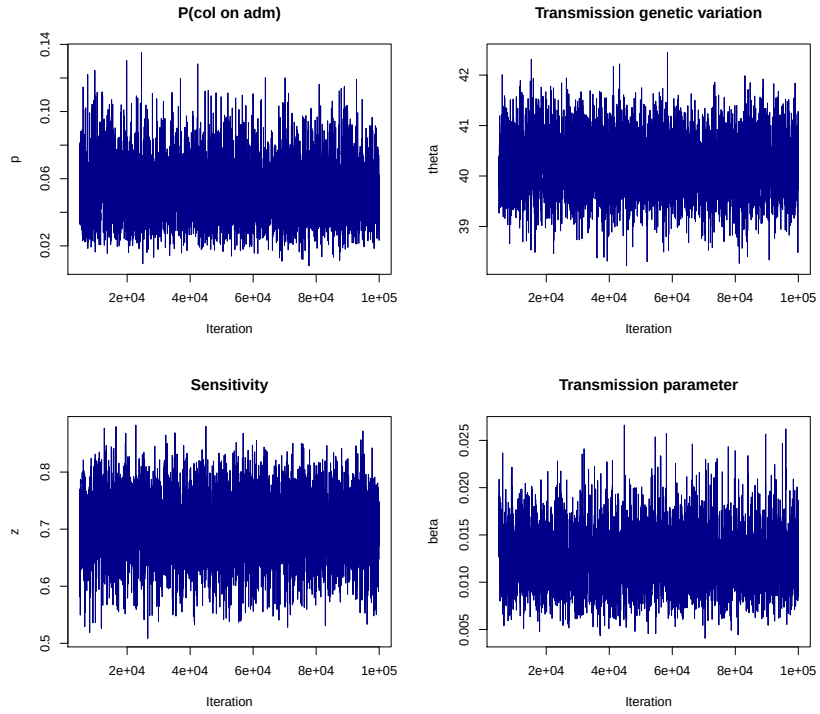


Figure B.3: Traceplots of estimates of parameters p, z, θ and β under the Chain Poisson model.

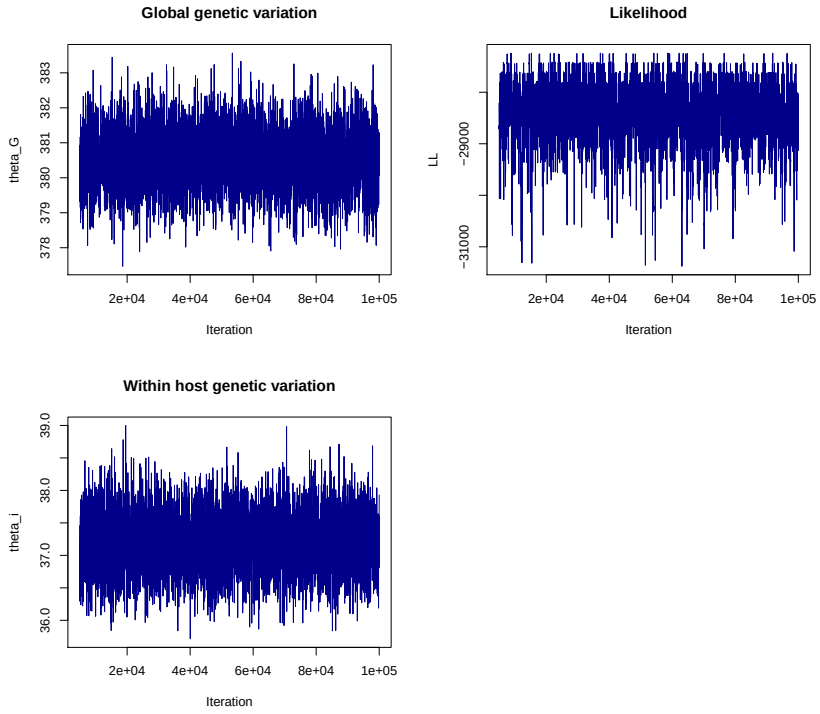


Figure B.4: Traceplots of estimates of parameters θ_{gl}, θ_i and the log likelihood under the Chain Poisson model.

B.3 Time Dependent Distances model

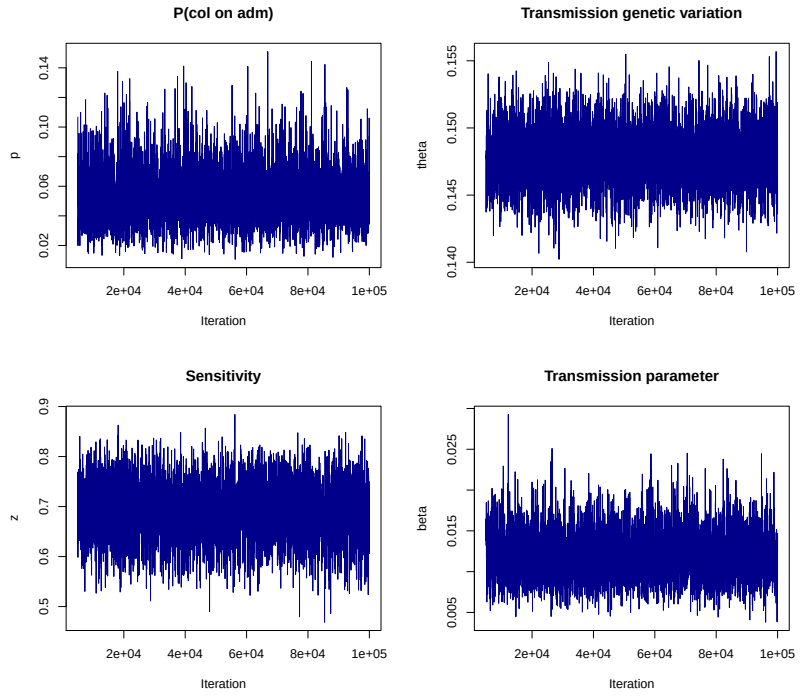


Figure B.5: Traceplots of estimates of parameters p, z, θ and β under the Time Dependent Distances model.

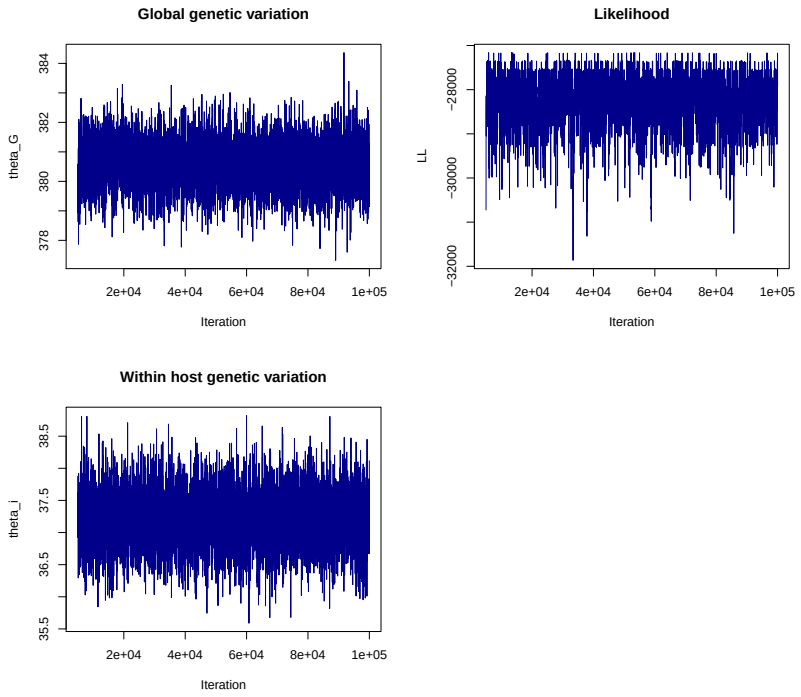


Figure B.6: Traceplots of estimates of parameters θ_{gl}, θ_i and the log likelihood under the Time Dependent Distances model.

Appendix C

Graphs for estimation of simulation parameters for MRSA

C.1 Chain Error model

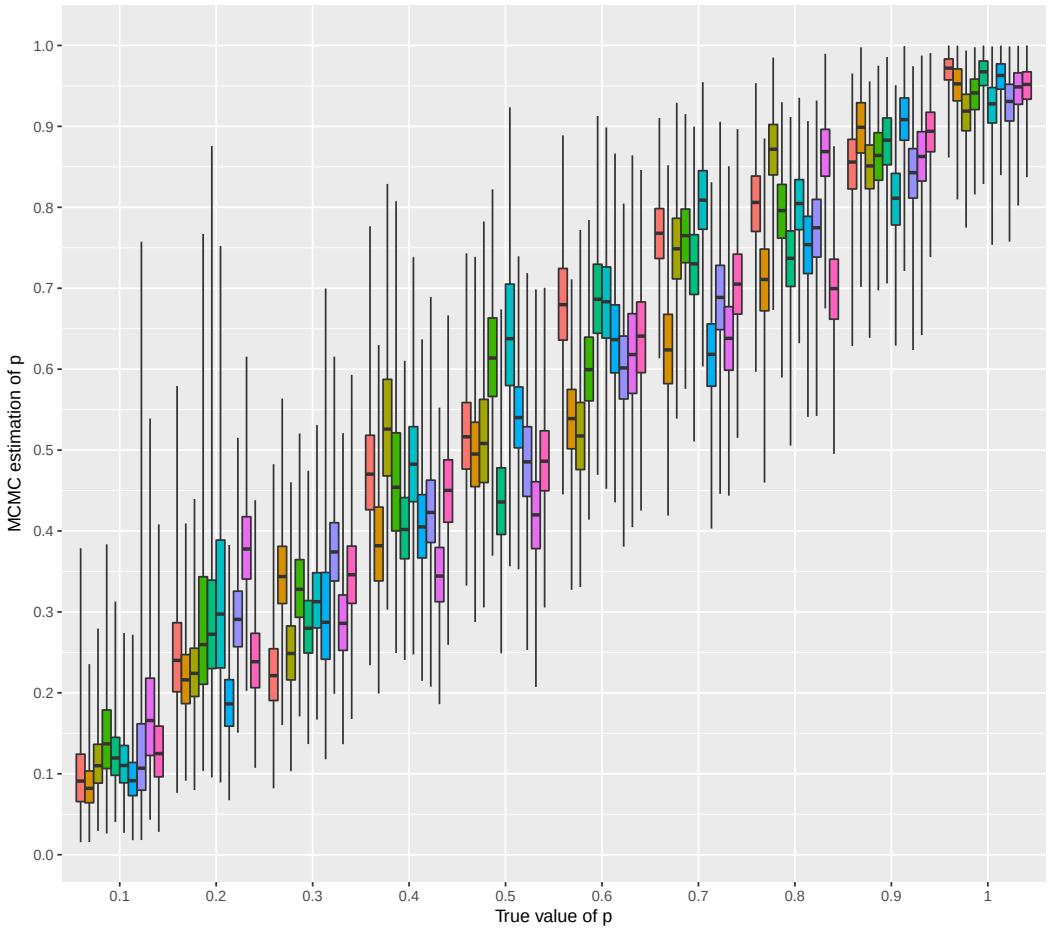


Figure C.1: The posterior estimates of parameter p from fitting the Chain Error model to 100 simulated datasets from the same model with varied input value for p, with 10 simulations for each value.

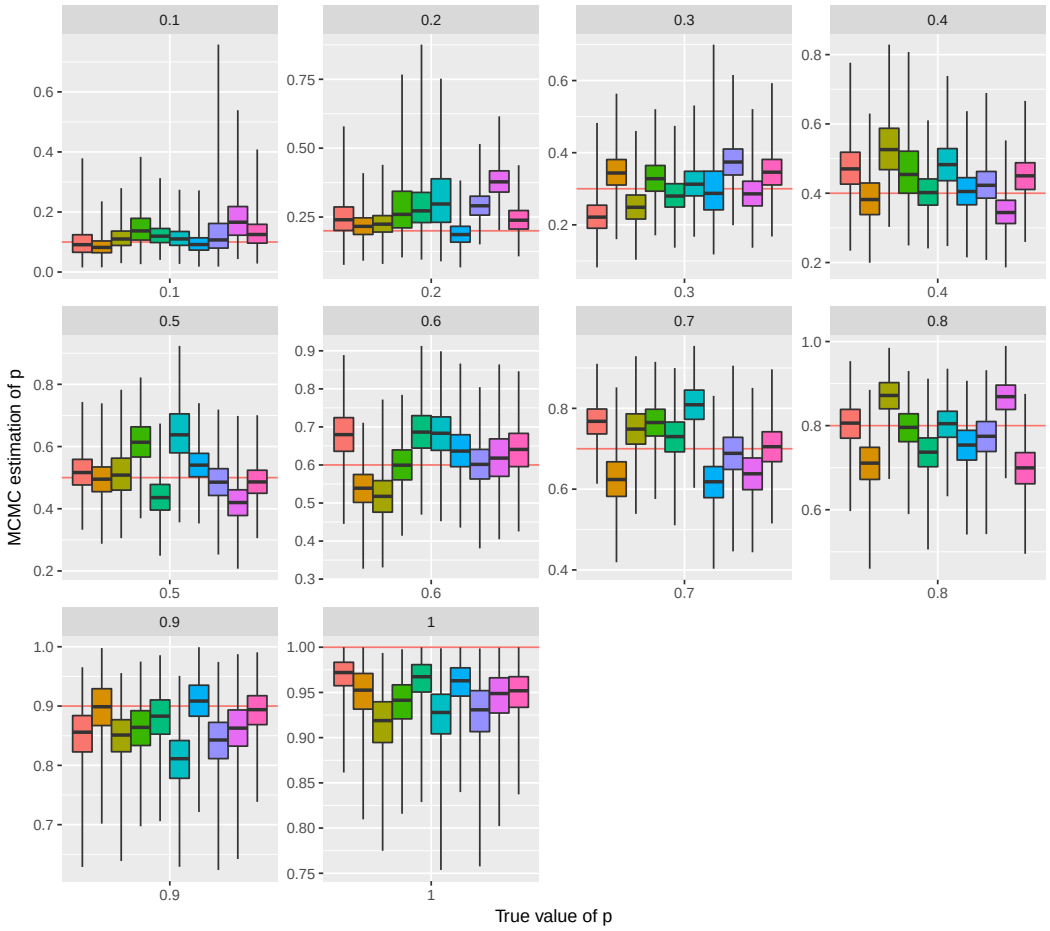


Figure C.2: The posterior estimates of parameter p from fitting the Chain Error model to 100 simulated datasets from the same model with varied input value for p, with 10 simulations for each value.

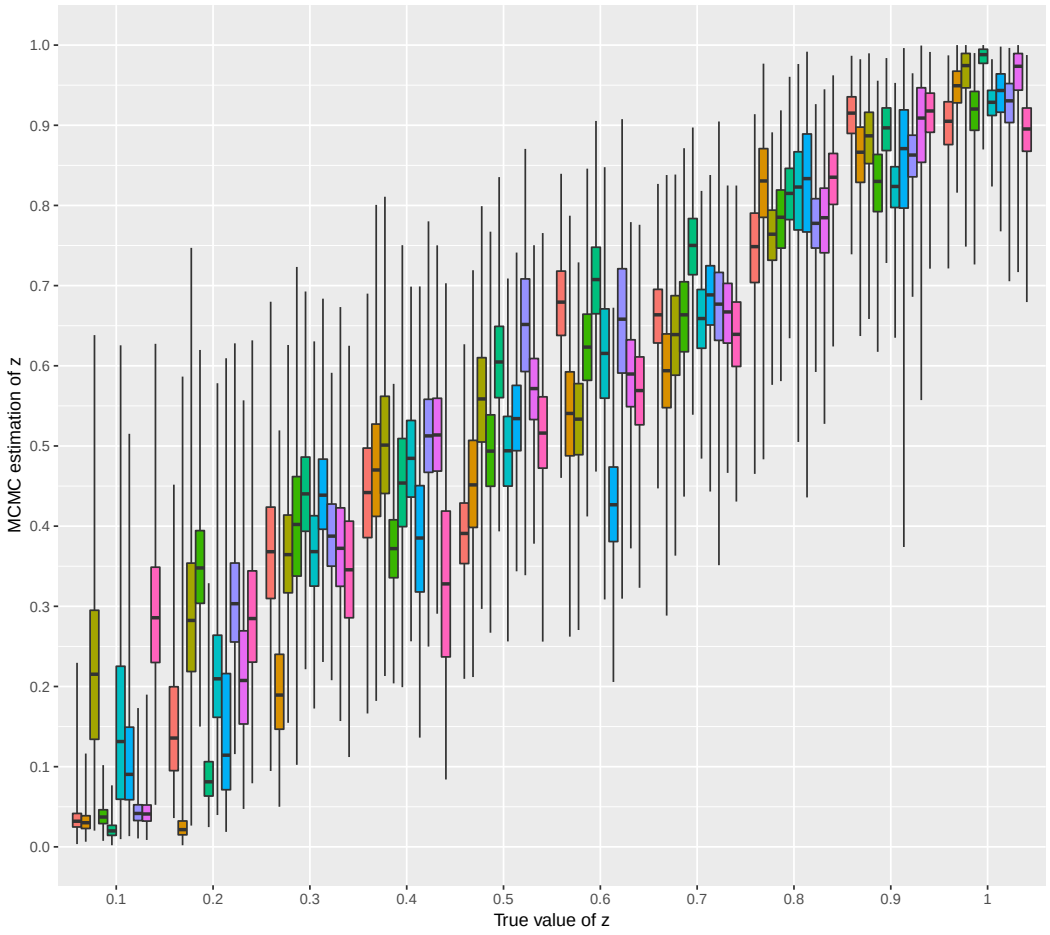


Figure C.3: The posterior estimates of parameter z from fitting the Chain Error model to 100 simulated datasets from the same model with varied input value for z, with 10 simulations for each value.

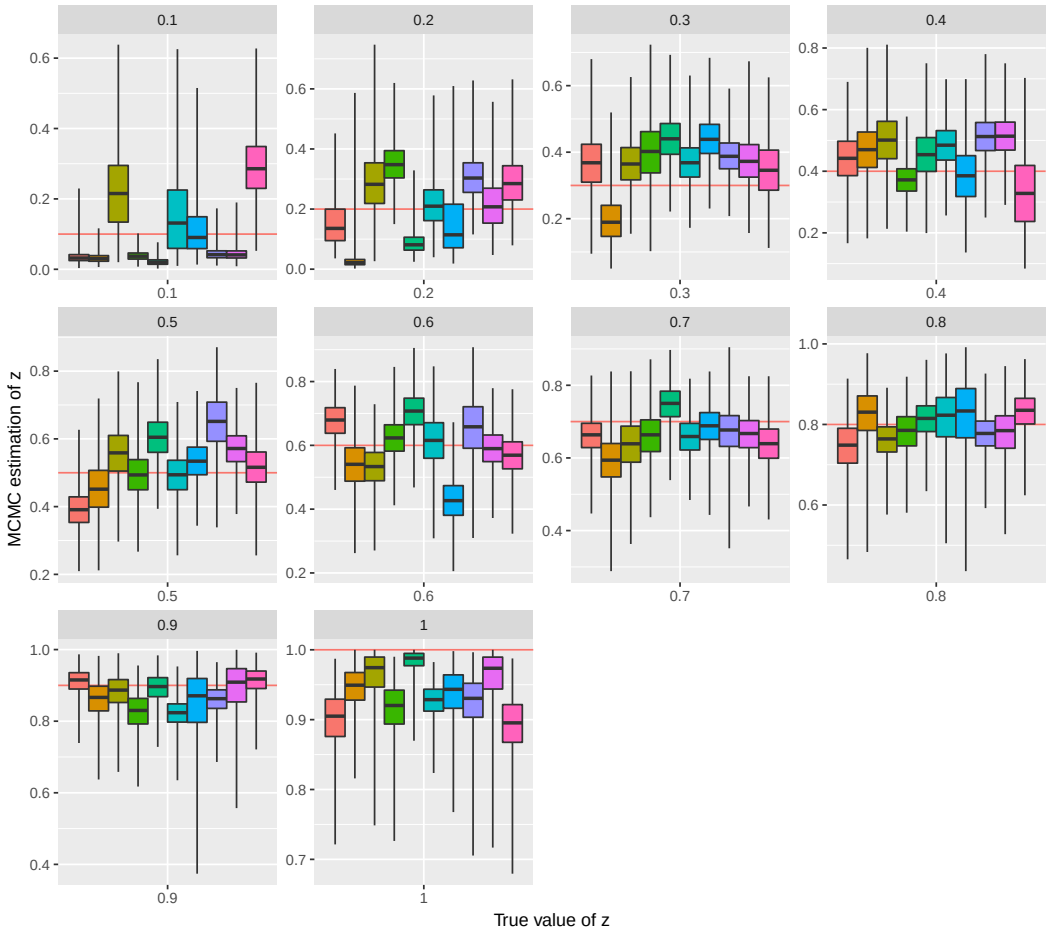


Figure C.4: The posterior estimates of parameter z from fitting the Chain Error model to 100 simulated datasets from the same model with varied input value for z, with 10 simulations for each value.

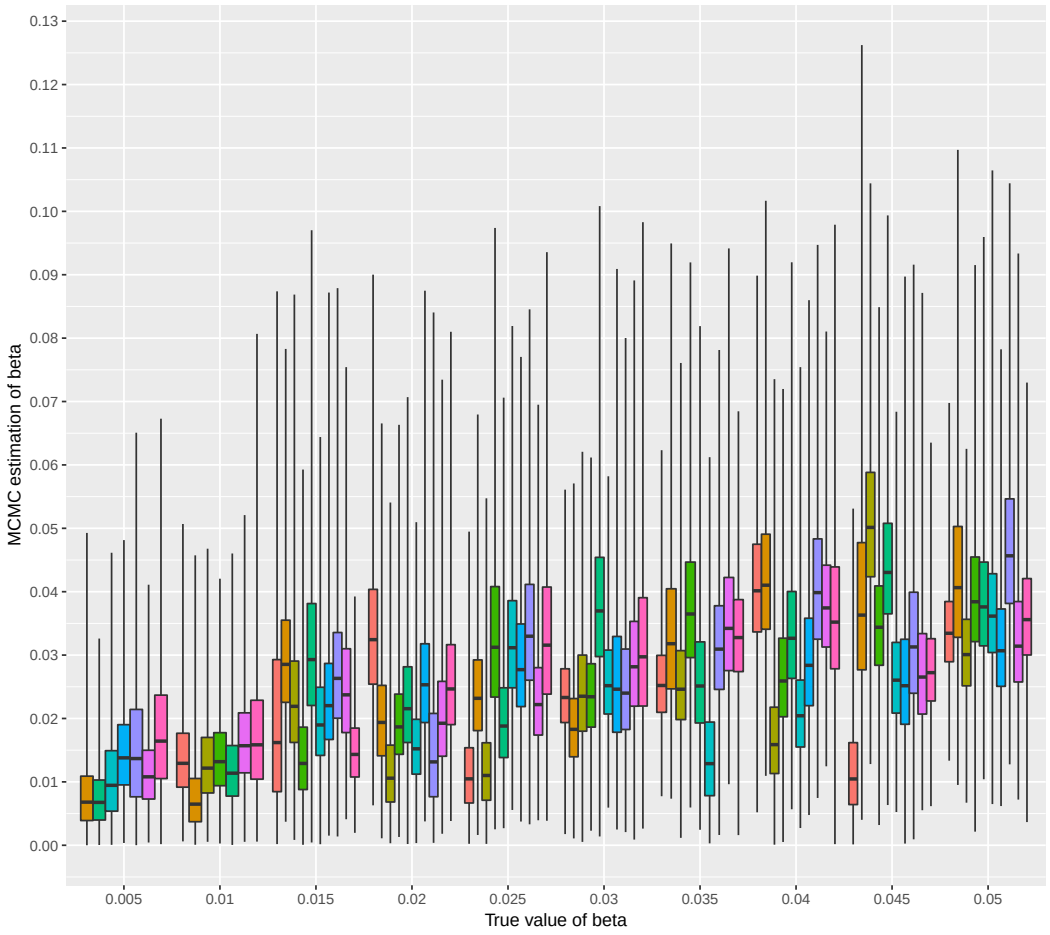


Figure C.5: The posterior estimates of parameter β from fitting the Chain Error model to 100 simulated datasets from the same model with varied input value for β, with 10 simulations for each value.

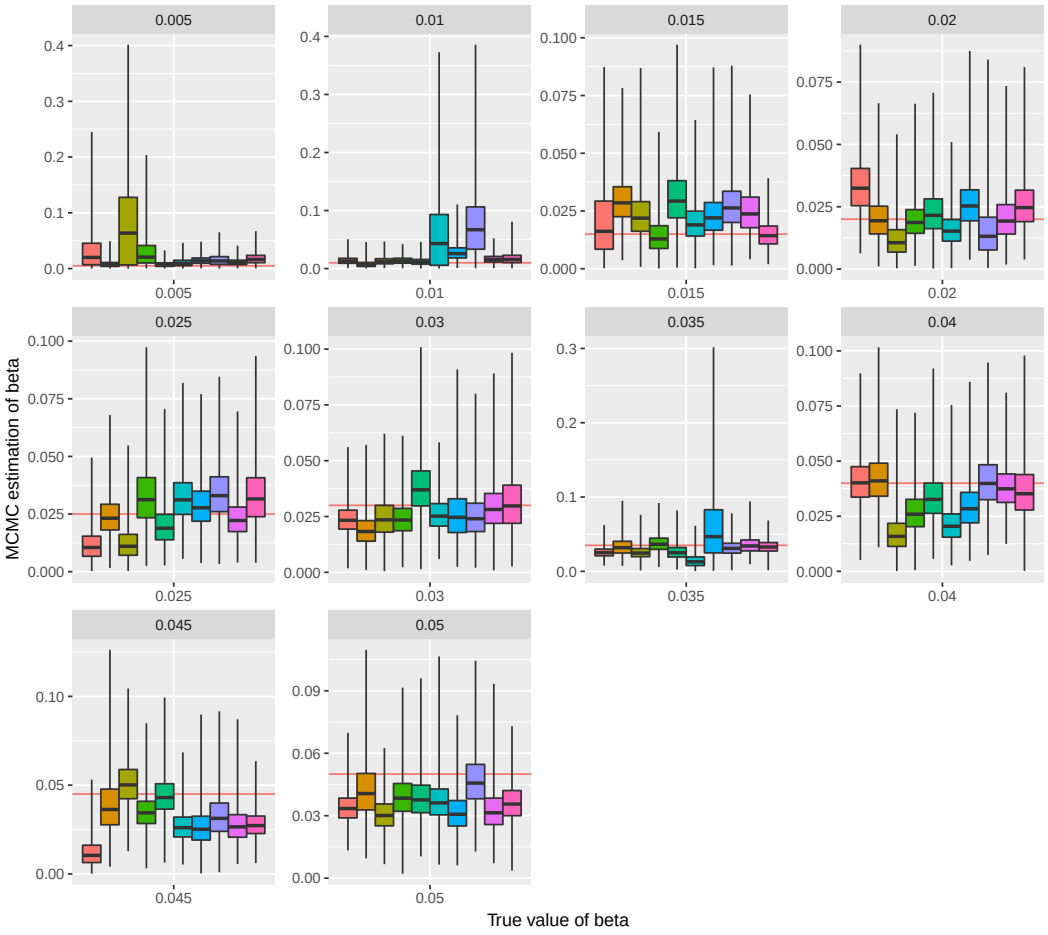


Figure C.6: The posterior estimates of parameter β from fitting the Chain Error model to 100 simulated datasets from the same model with varied input value for β, with 10 simulations for each value.

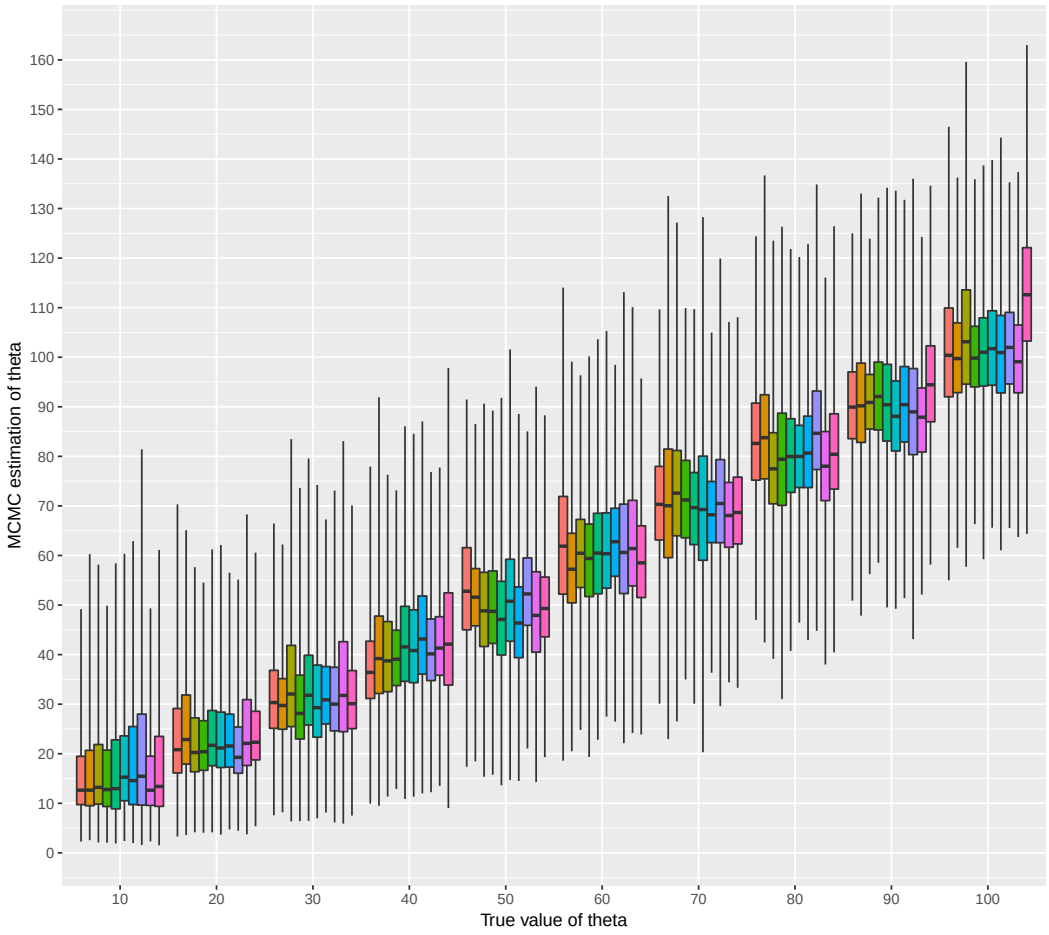


Figure C.7: The posterior estimates of parameter θ from fitting the Chain Error model to 100 simulated datasets from the same model with varied input value for θ, with 10 simulations for each value.

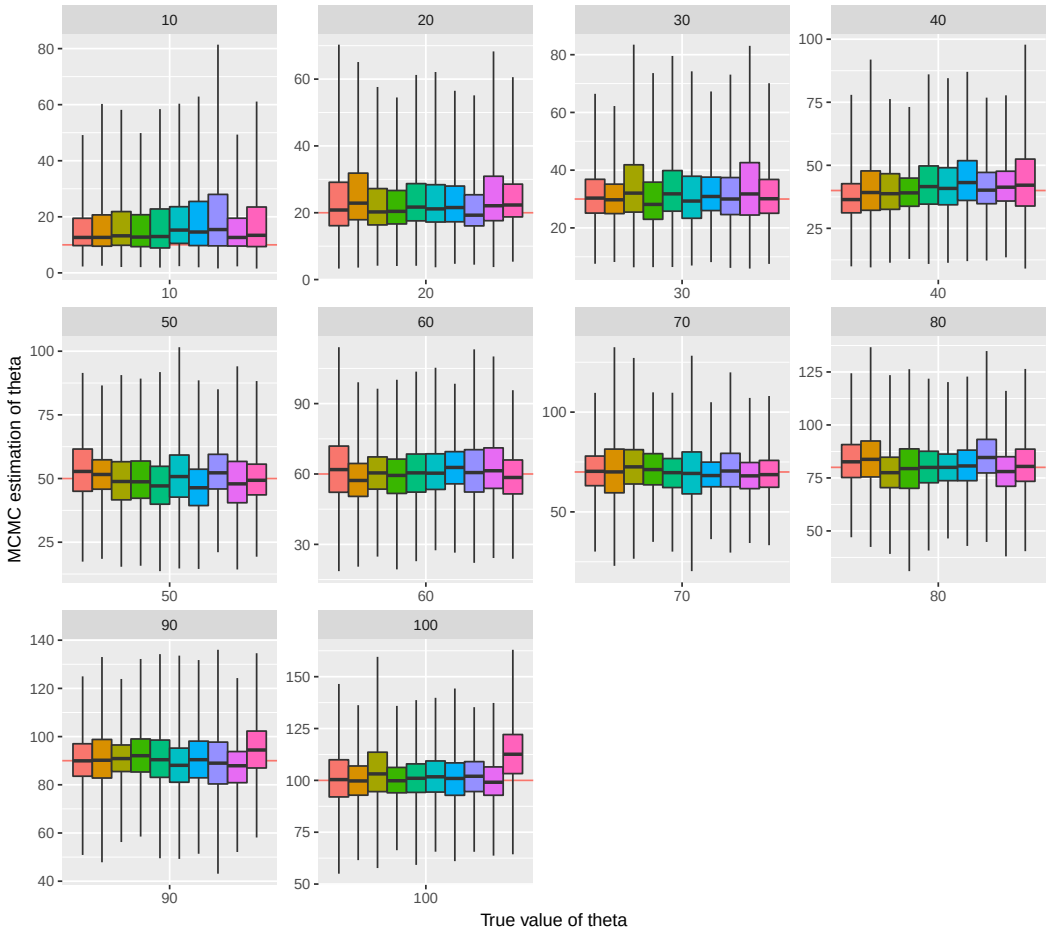


Figure C.8: The posterior estimates of parameter θ from fitting the Chain Error model to 100 simulated datasets from the same model with varied input value for θ, with 10 simulations for each value.

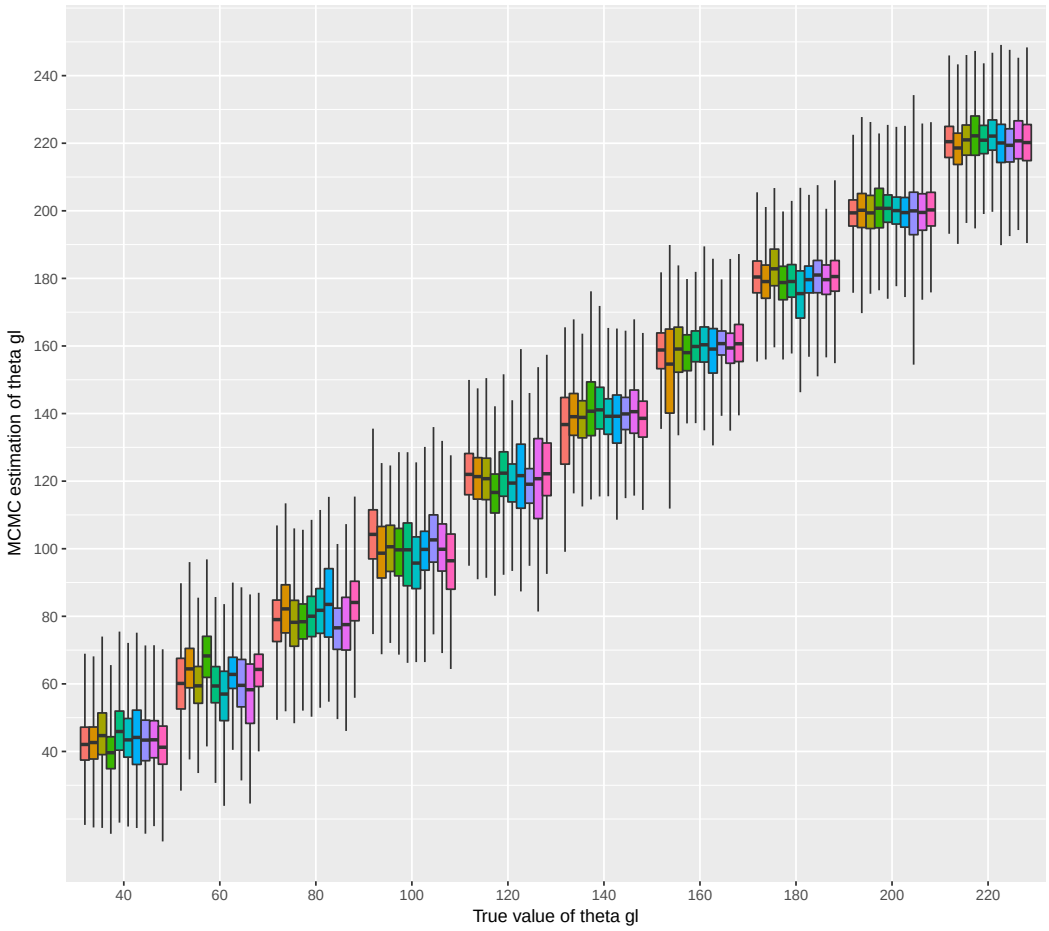


Figure C.9: The posterior estimates of parameter θ_{gl} from fitting the Chain Error model to 100 simulated datasets from the same model with varied input value for θ_{gl}, with 10 simulations for each value.

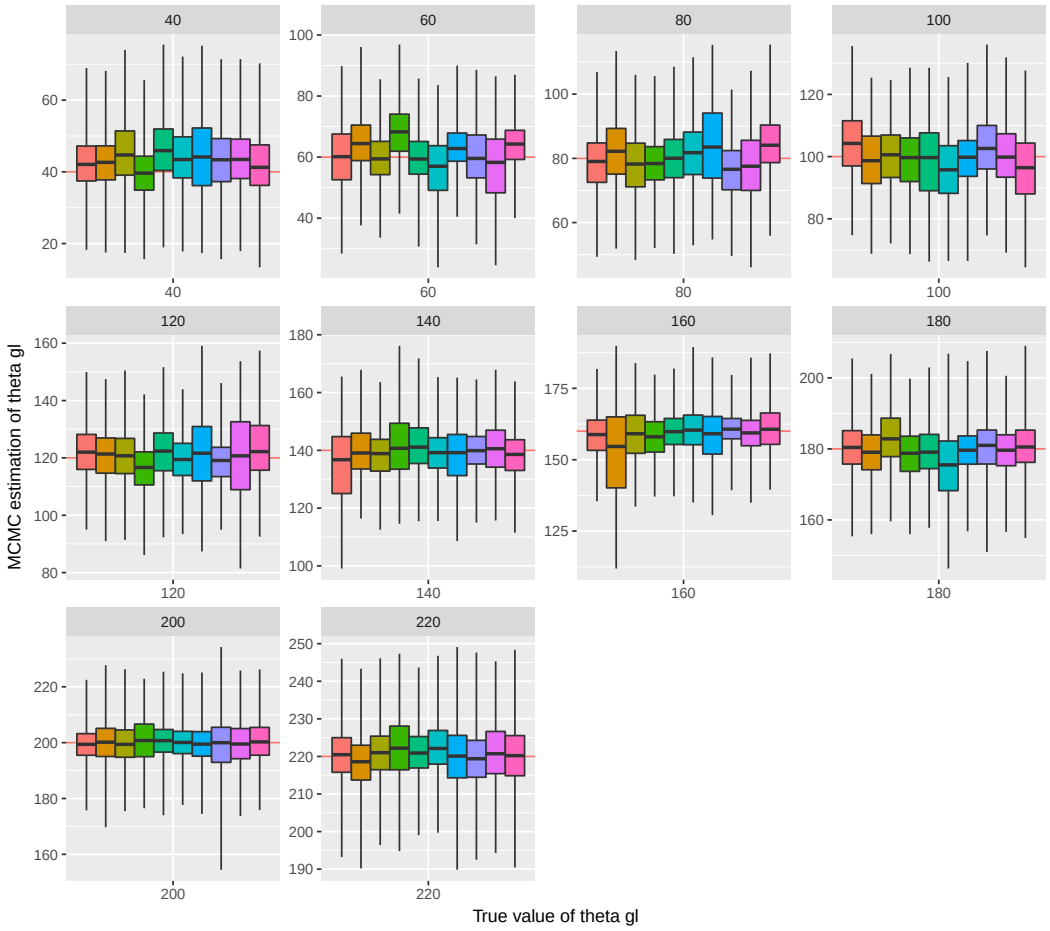


Figure C.10: The posterior estimates of parameter θ_{gl} from fitting the Chain Error model to 100 simulated datasets from the same model with varied input value for θ_{gl}, with 10 simulations for each value.

C.2 Chain Poisson model

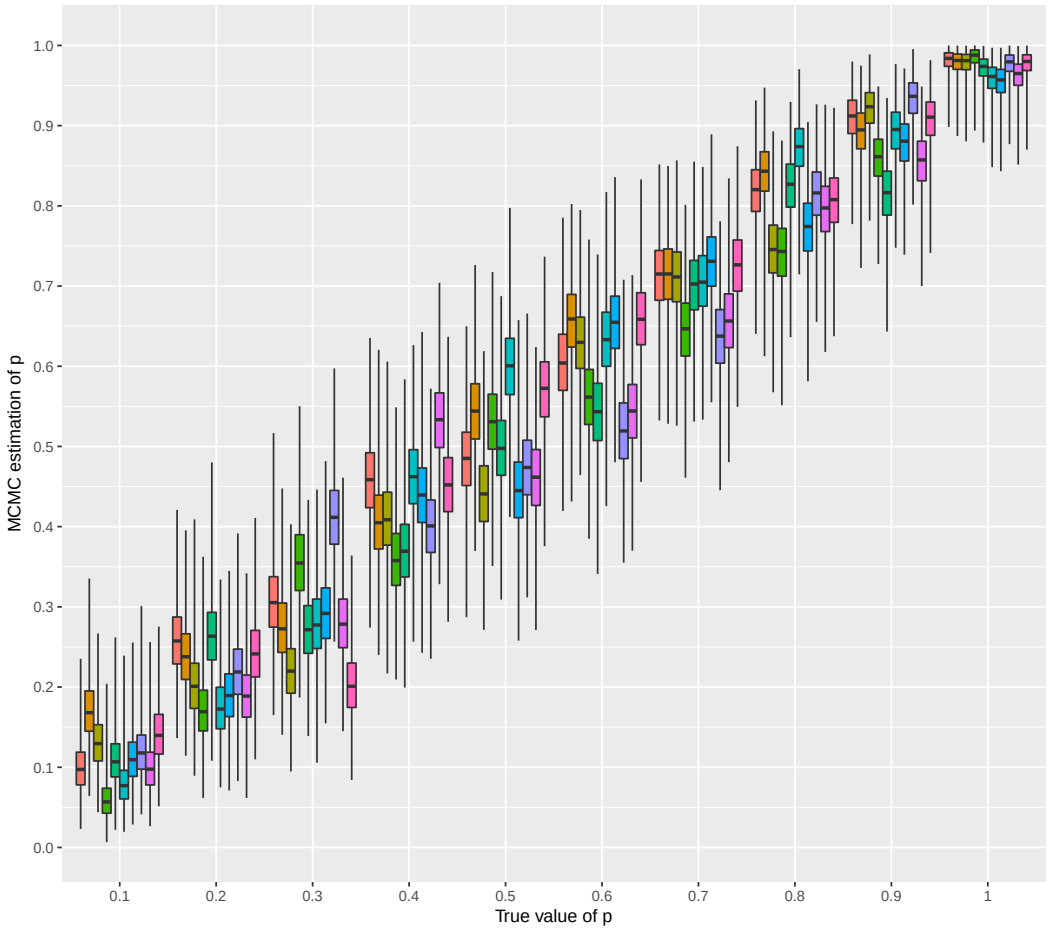


Figure C.11: The posterior estimates of parameter p from fitting the Chain Poisson model to 100 simulated datasets from the same model with varied input value for p, with 10 simulations for each value.

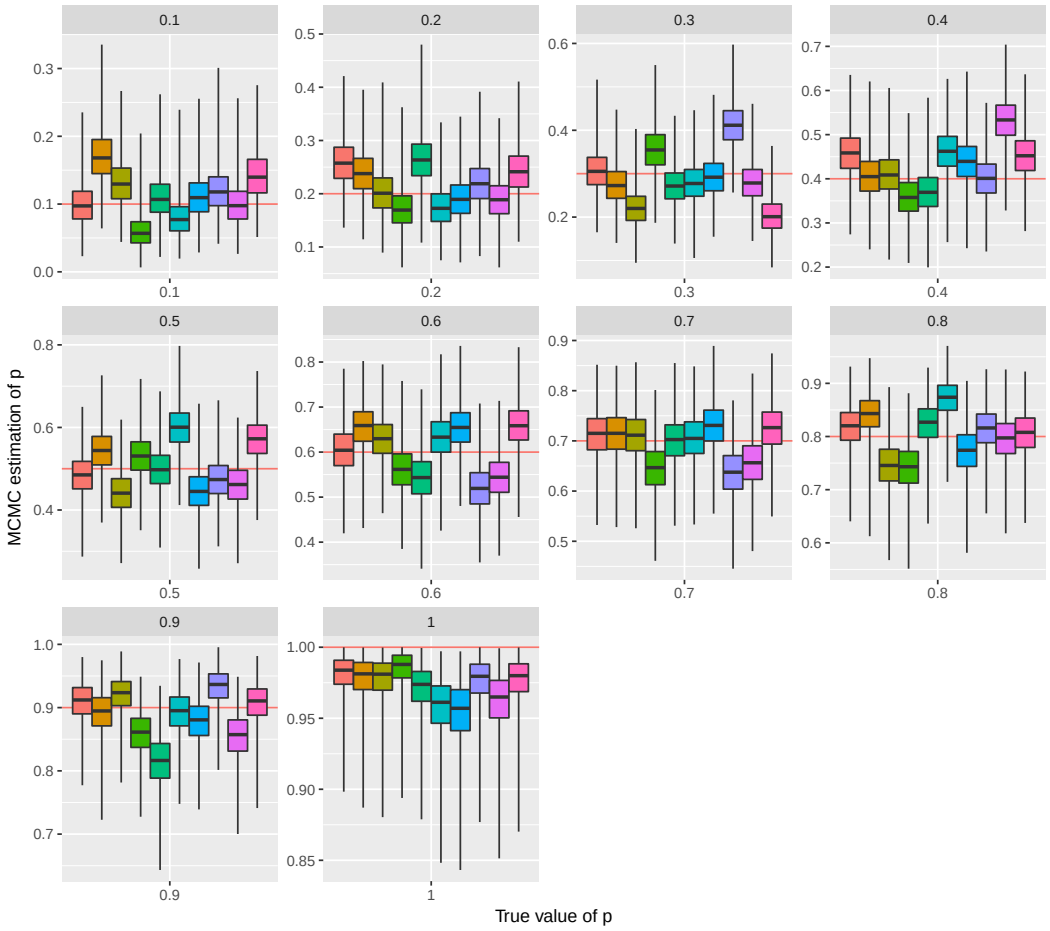


Figure C.12: The posterior estimates of parameter p from fitting the Chain Poisson model to 100 simulated datasets from the same model with varied input value for p, with 10 simulations for each value.

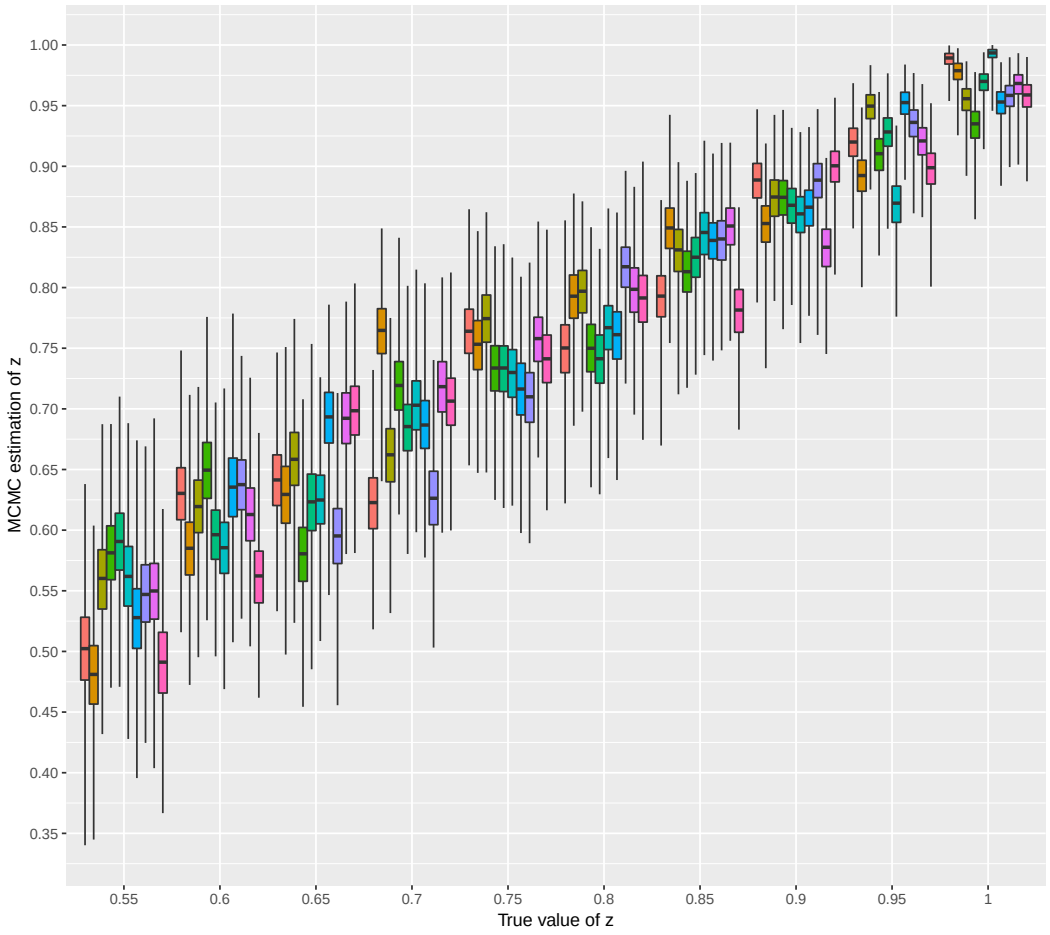


Figure C.13: The posterior estimates of parameter z from fitting the Chain Poisson model to 100 simulated datasets from the same model with varied input value for z, with 10 simulations for each value.

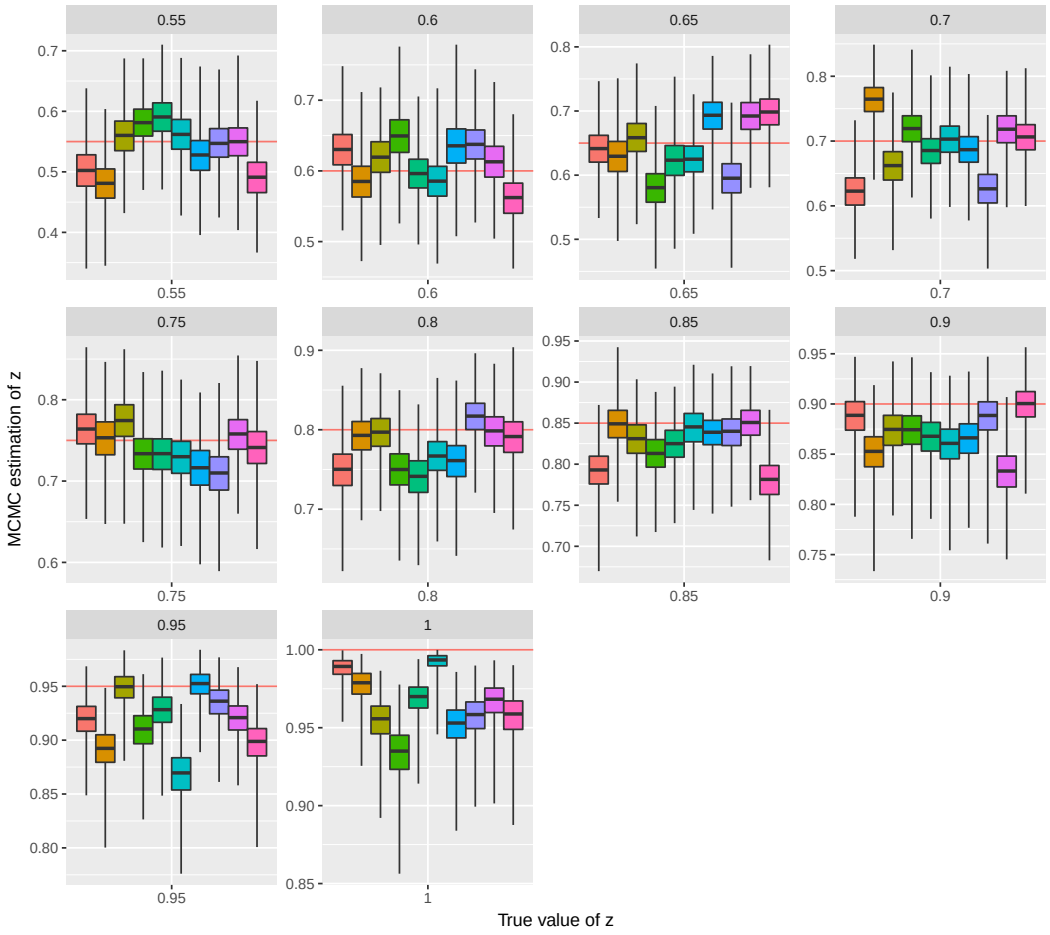


Figure C.14: The posterior estimates of parameter z from fitting the Chain Poisson model to 100 simulated datasets from the same model with varied input value for z, with 10 simulations for each value.

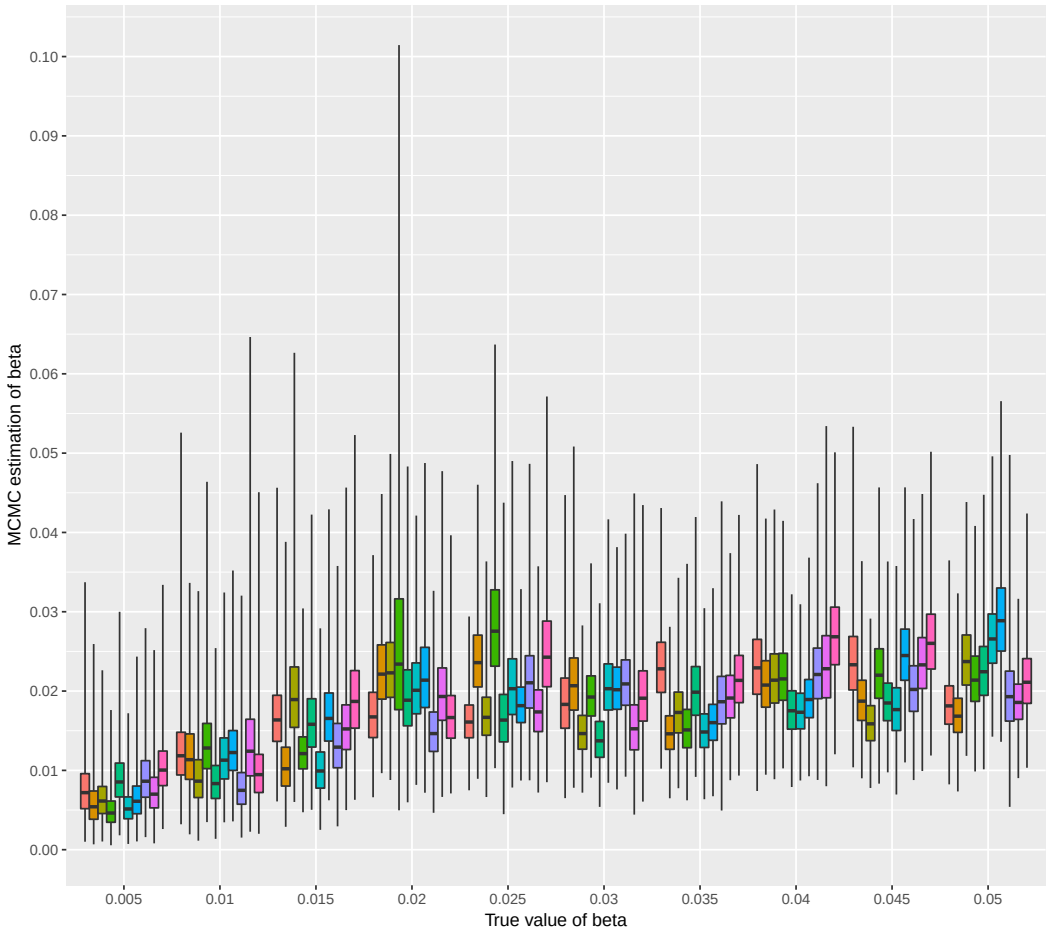


Figure C.15: The posterior estimates of parameter β from fitting the Chain Poisson model to 100 simulated datasets from the same model with varied input value for β, with 10 simulations for each value.

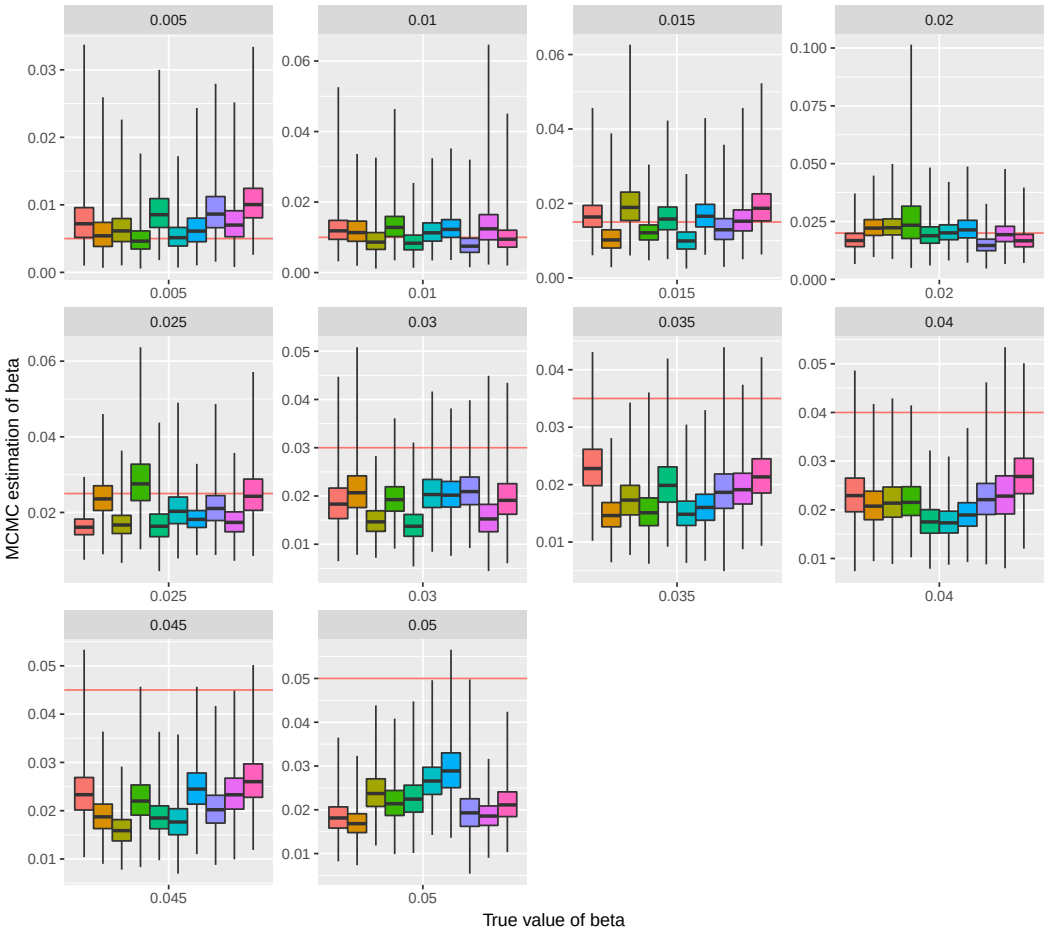


Figure C.16: The posterior estimates of parameter β from fitting the Chain Poisson model to 100 simulated datasets from the same model with varied input value for β, with 10 simulations for each value.

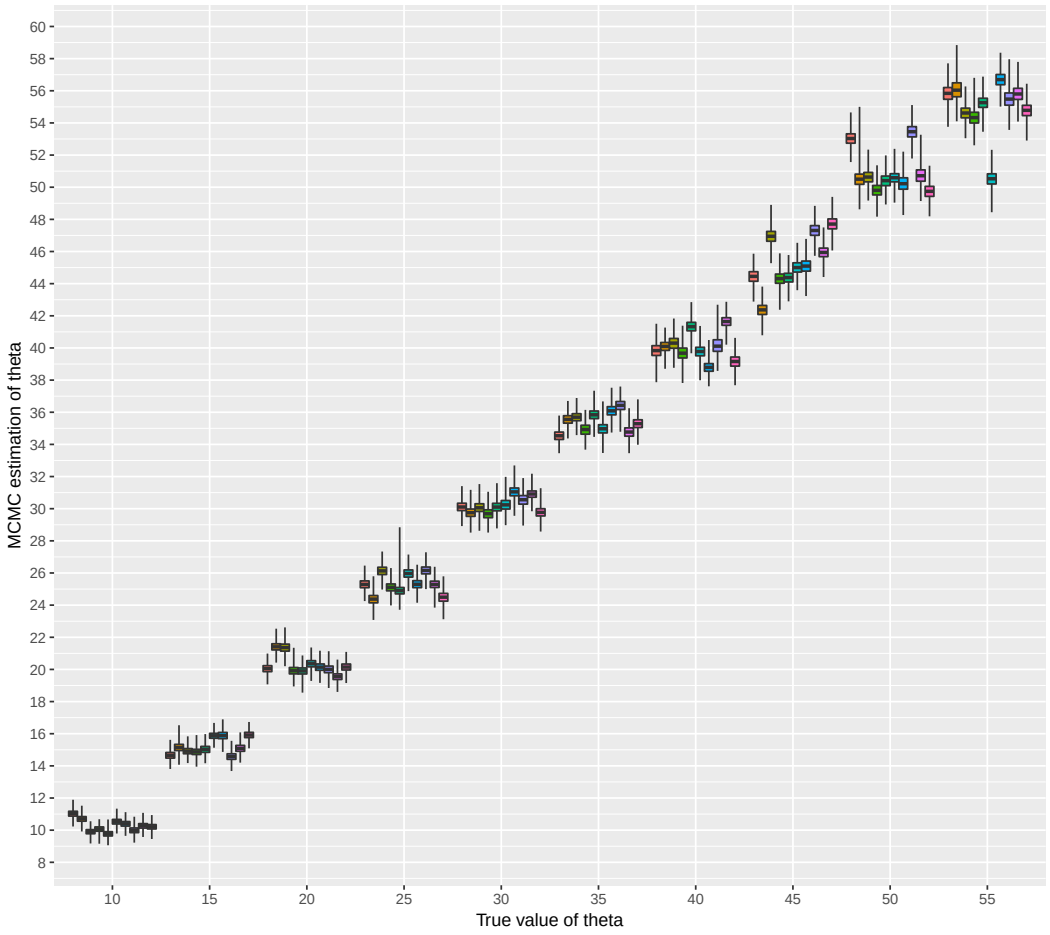


Figure C.17: The posterior estimates of parameter θ from fitting the Chain Poisson model to 100 simulated datasets from the same model with varied input value for θ, with 10 simulations for each value.

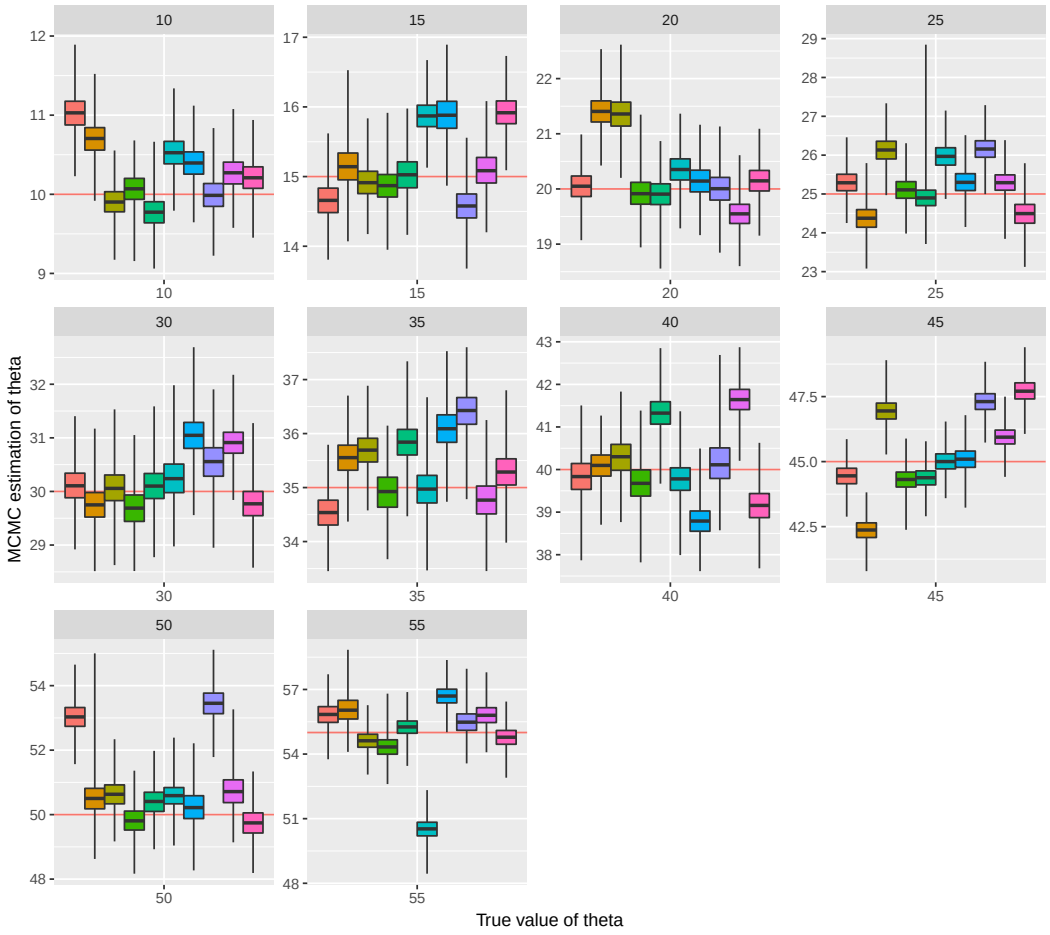


Figure C.18: The posterior estimates of parameter θ from fitting the Chain Poisson model to 100 simulated datasets from the same model with varied input value for θ, with 10 simulations for each value.

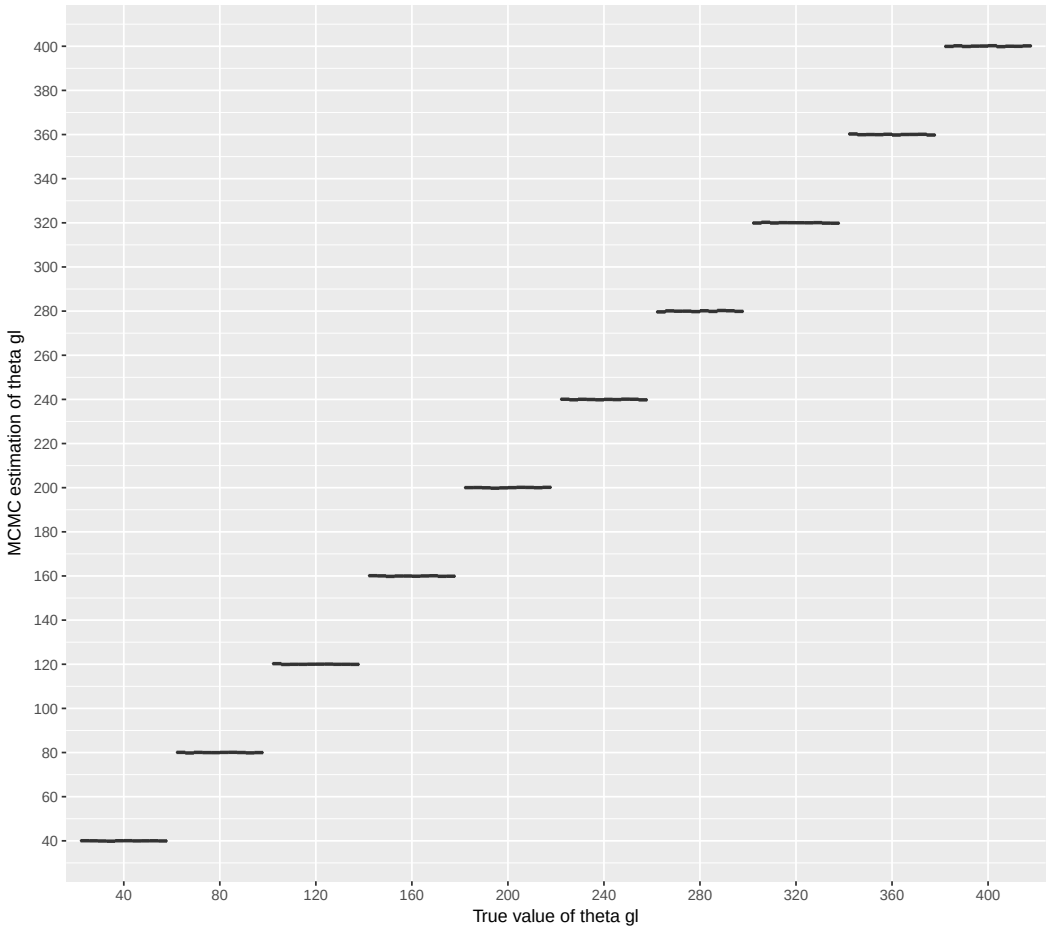


Figure C.19: The posterior estimates of parameter θ_{gl} from fitting the Chain Poisson model to 100 simulated datasets from the same model with varied input value for θ_{gl}, with 10 simulations for each value.

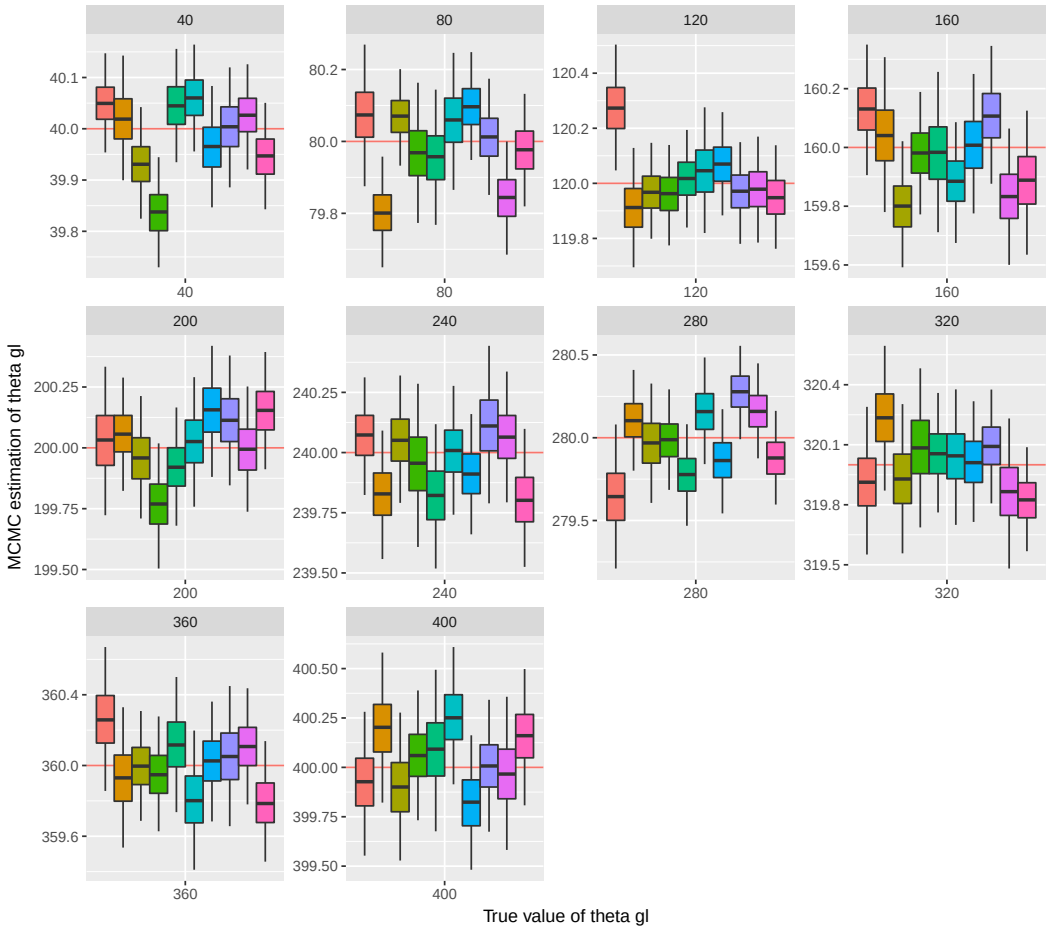


Figure C.20: The posterior estimates of parameter θ_{gl} from fitting the Chain Poisson model to 100 simulated datasets from the same model with varied input value for θ_{gl}, with 10 simulations for each value.

C.3 Time Dependent Distances model

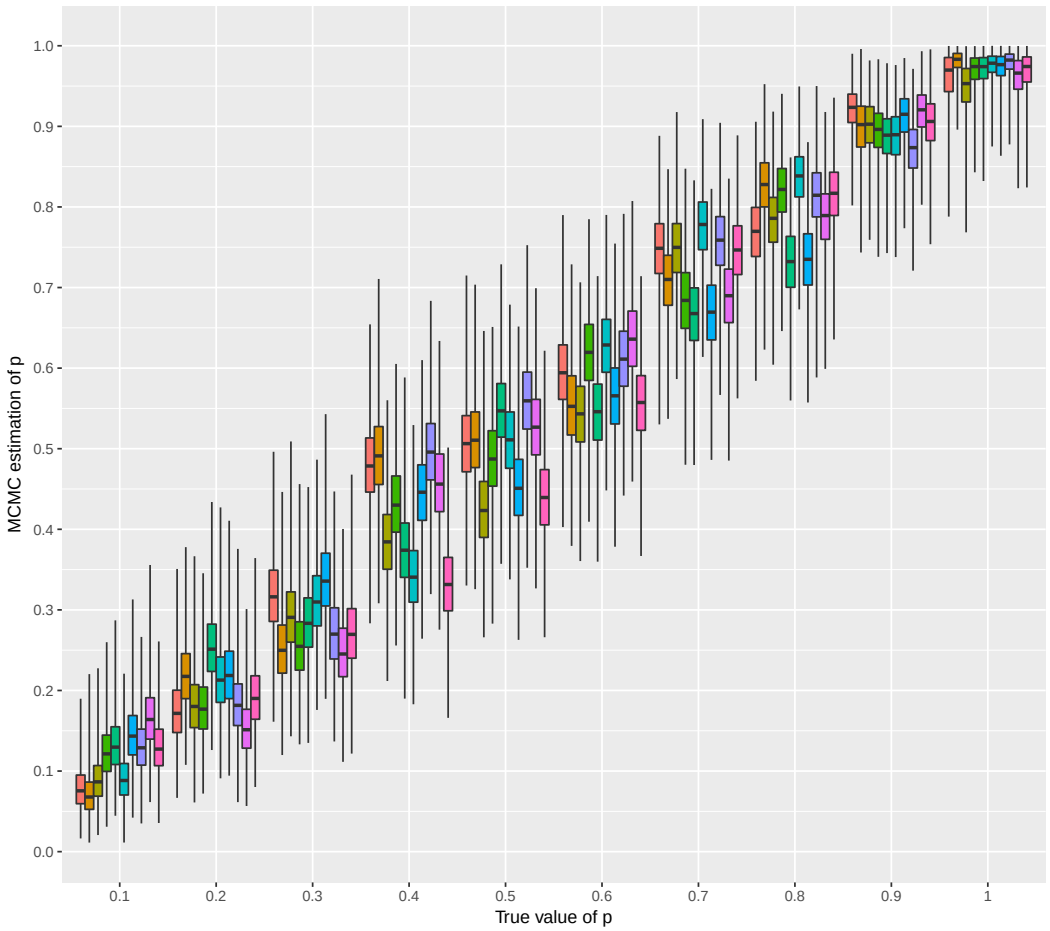


Figure C.21: The posterior estimates of parameter p from fitting the Time Dependent Distances model to 100 simulated datasets from the same model with varied input value for p, with 10 simulations for each value.

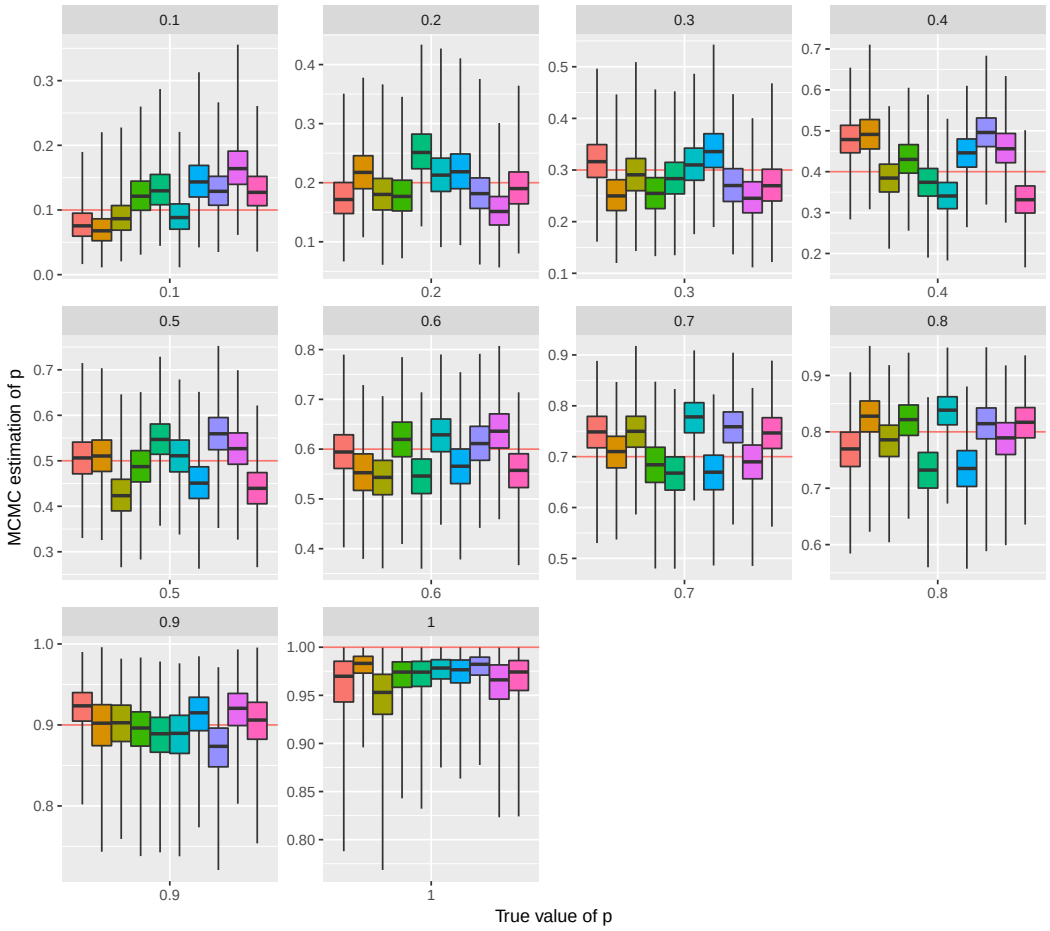


Figure C.22: The posterior estimates of parameter p from fitting the Time Dependent Distances model to 100 simulated datasets from the same model with varied input value for p, with 10 simulations for each value.

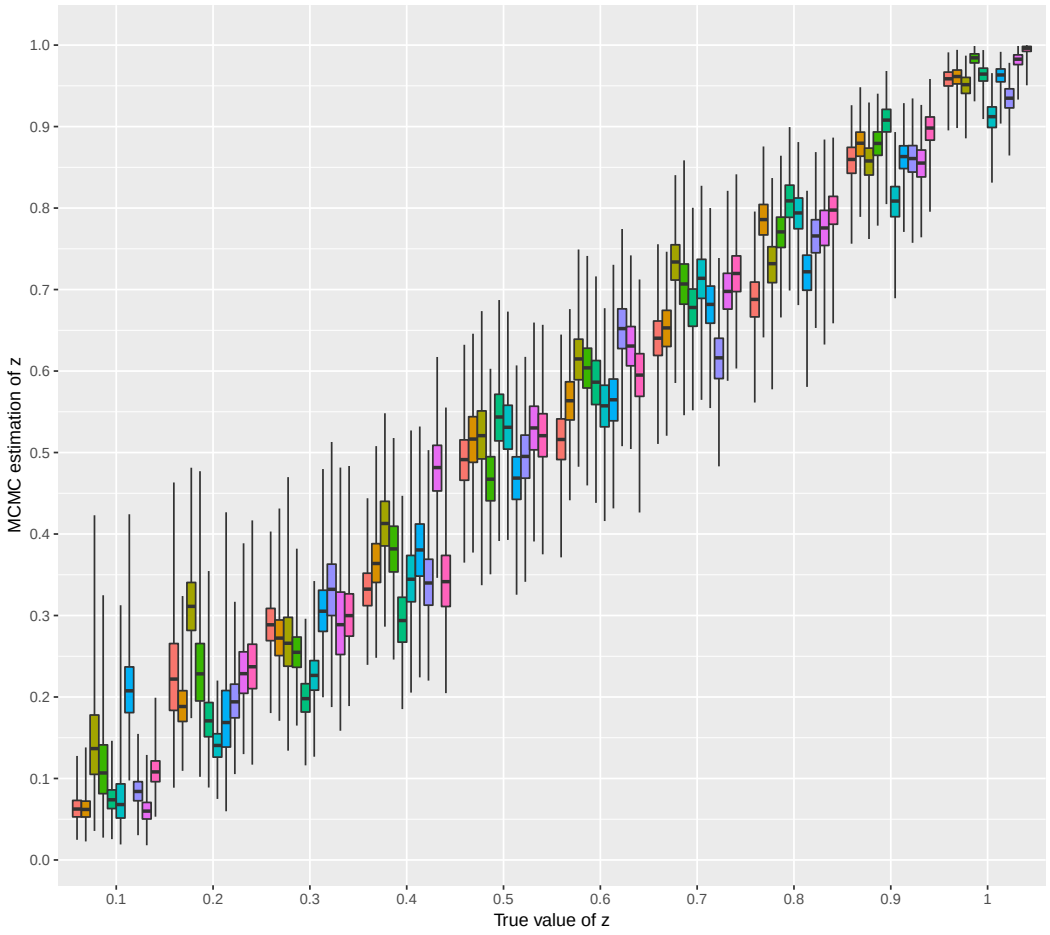


Figure C.23: The posterior estimates of parameter z from fitting the Time Dependent Distances model to 100 simulated datasets from the same model with varied input value for z, with 10 simulations for each value.

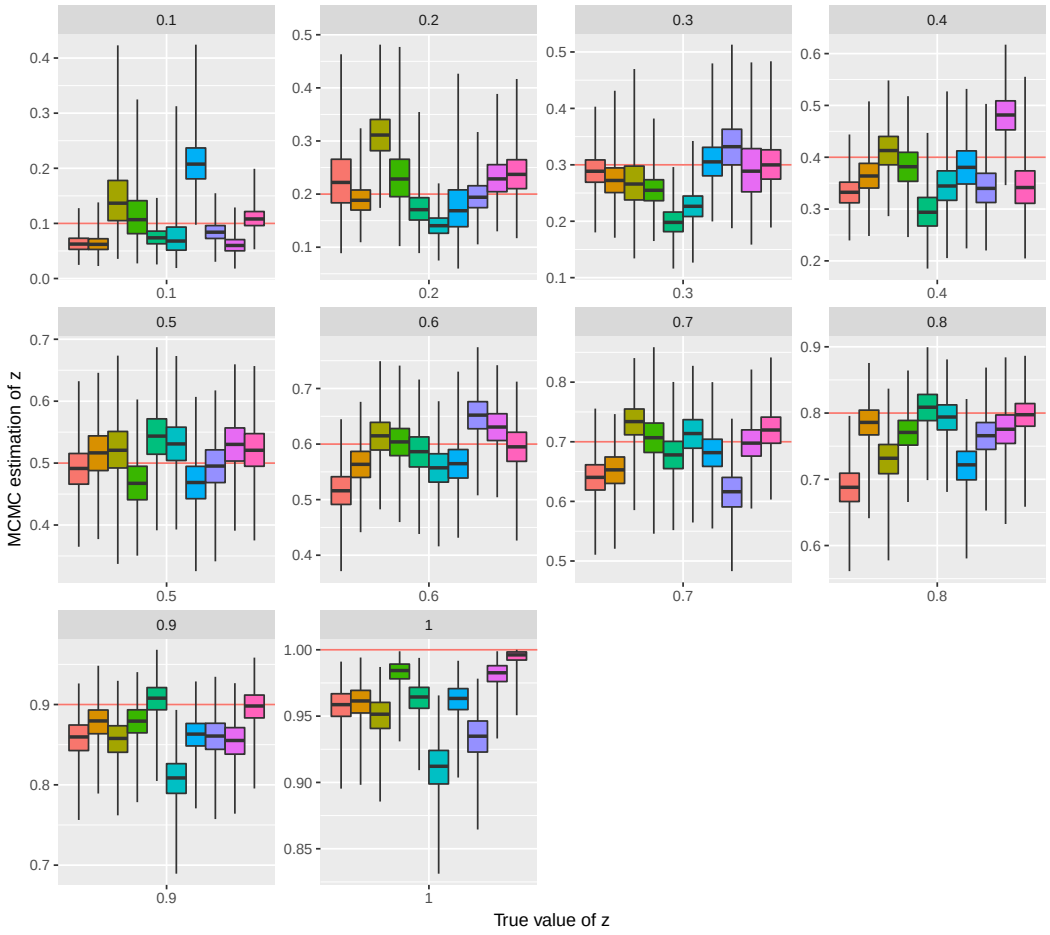


Figure C.24: The posterior estimates of parameter z from fitting the Time Dependent Distances model to 100 simulated datasets from the same model with varied input value for z, with 10 simulations for each value.

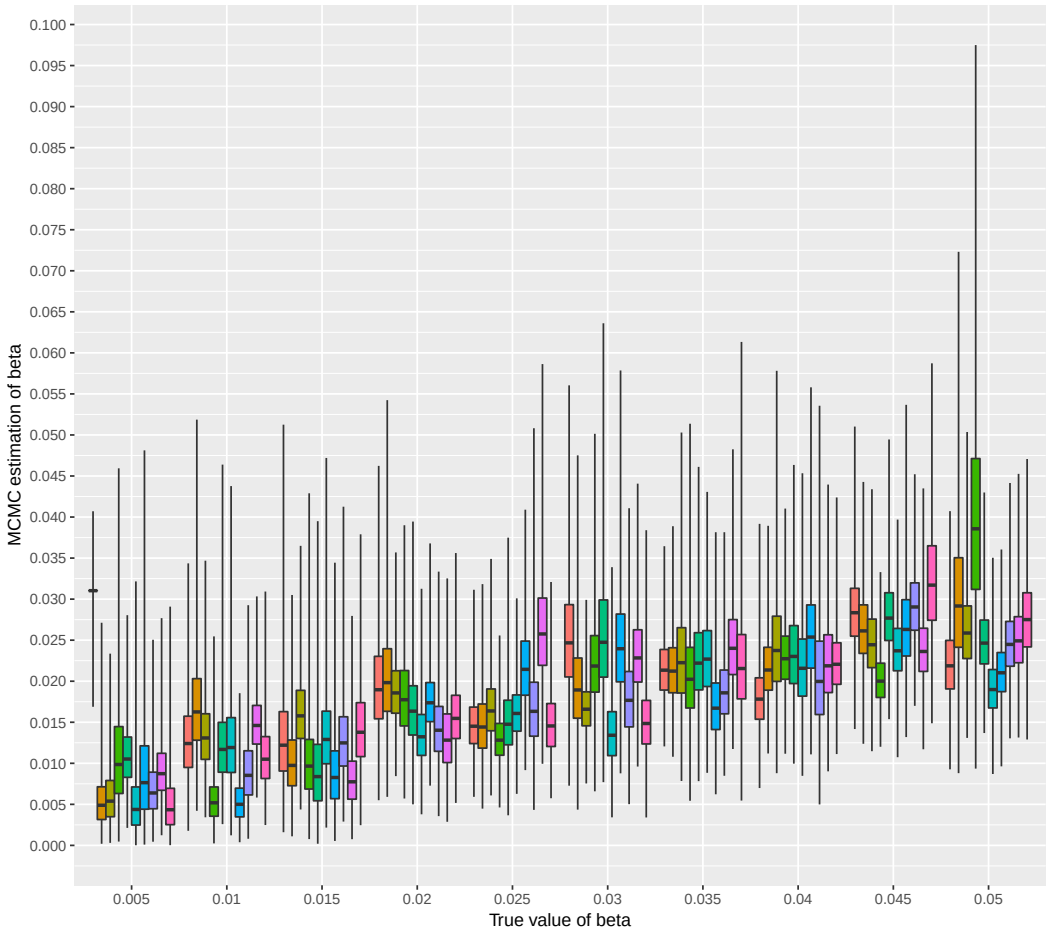


Figure C.25: The posterior estimates of parameter β from fitting the Time Dependent Distances model to 100 simulated datasets from the same model with varied input value for β, with 10 simulations for each value.

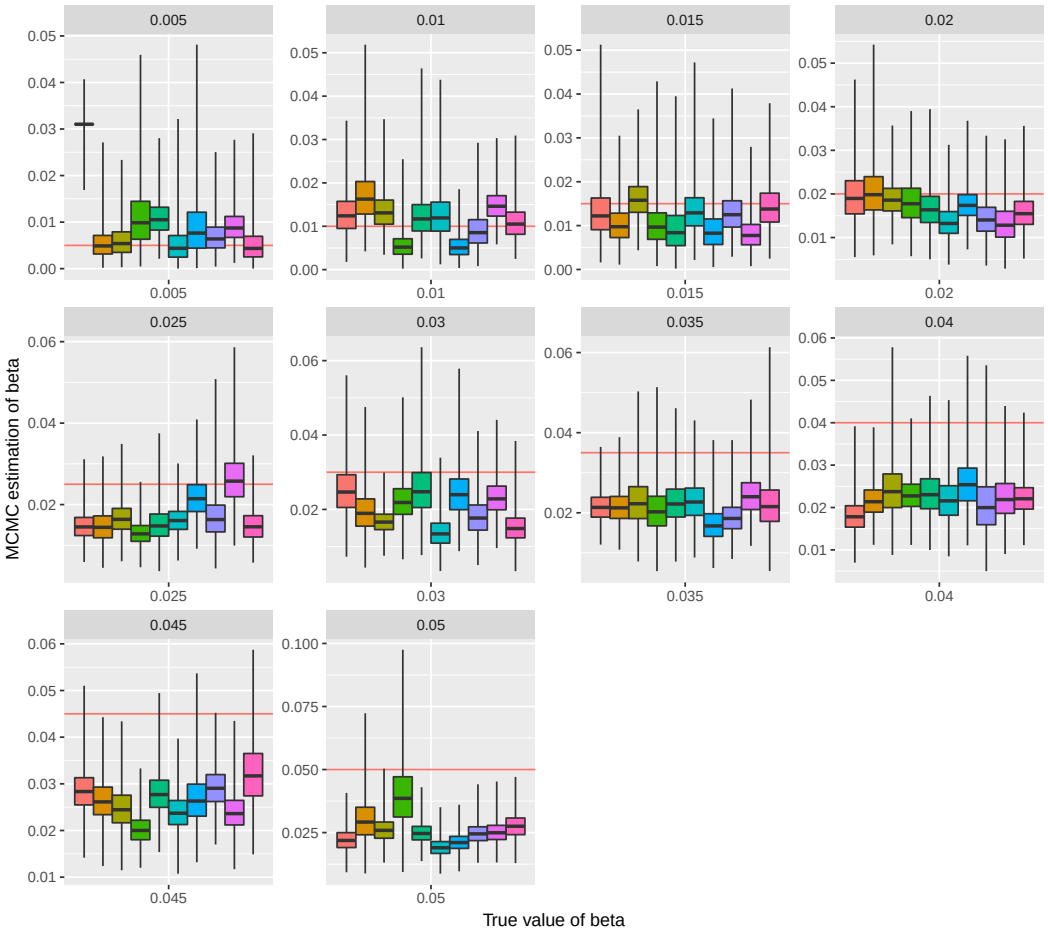


Figure C.26: The posterior estimates of parameter β from fitting the Time Dependent Distances model to 100 simulated datasets from the same model with varied input value for β, with 10 simulations for each value.

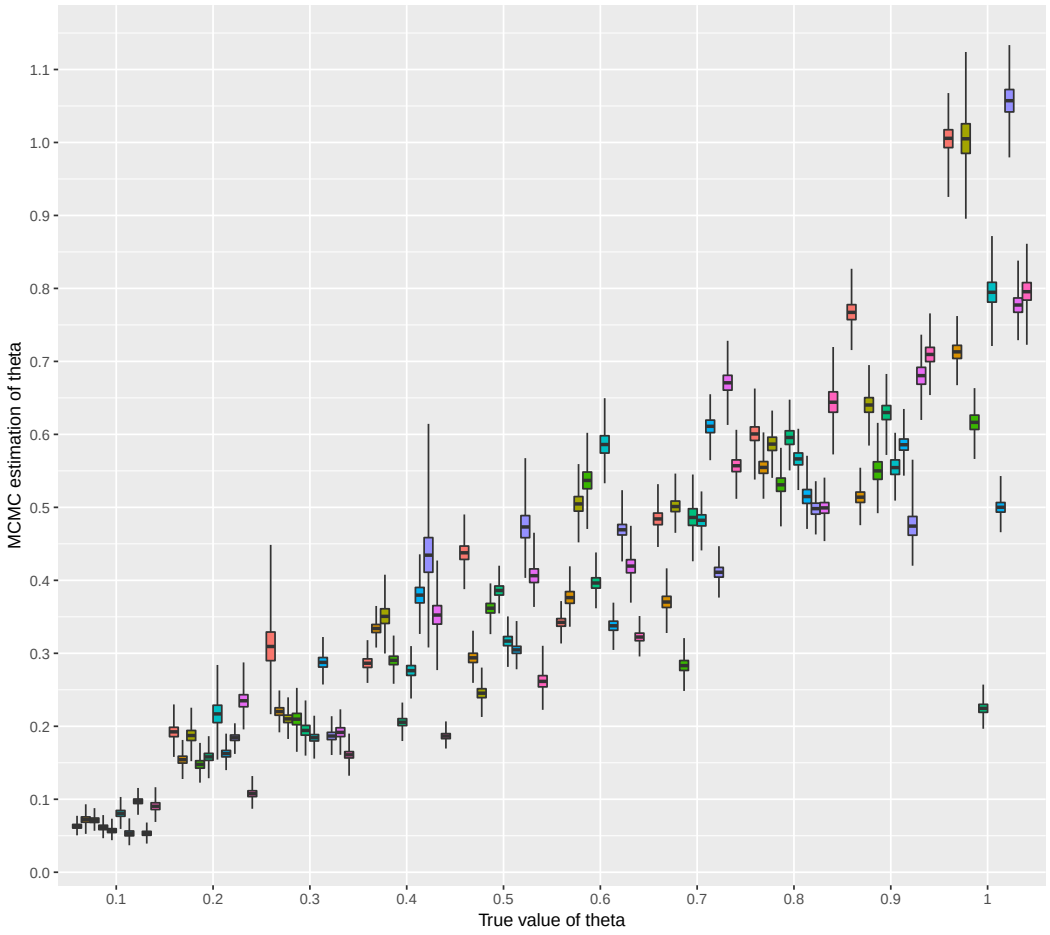


Figure C.27: The posterior estimates of parameter θ from fitting the Time Dependent Distances model to 100 simulated datasets from the same model with varied input value for θ, with 10 simulations for each value.

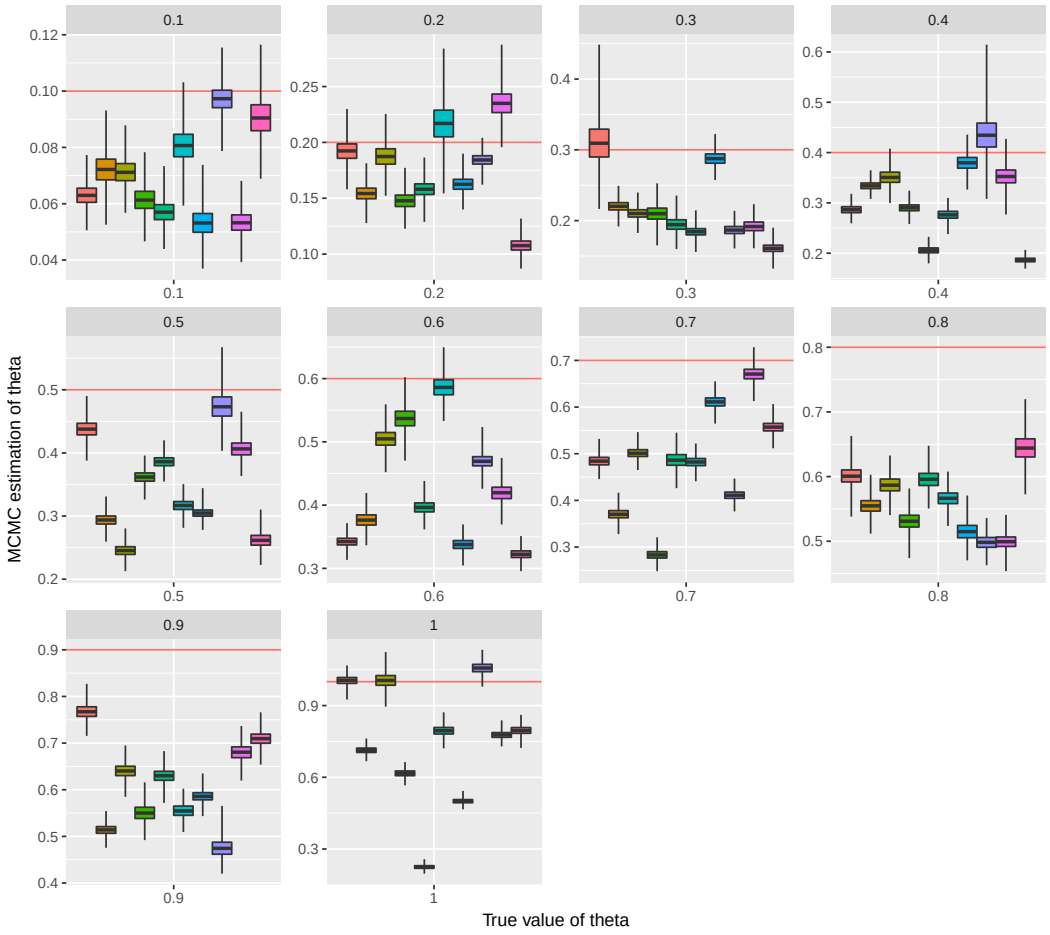


Figure C.28: The posterior estimates of parameter θ from fitting the Time Dependent Distances model to 100 simulated datasets from the same model with varied input value for θ, with 10 simulations for each value.

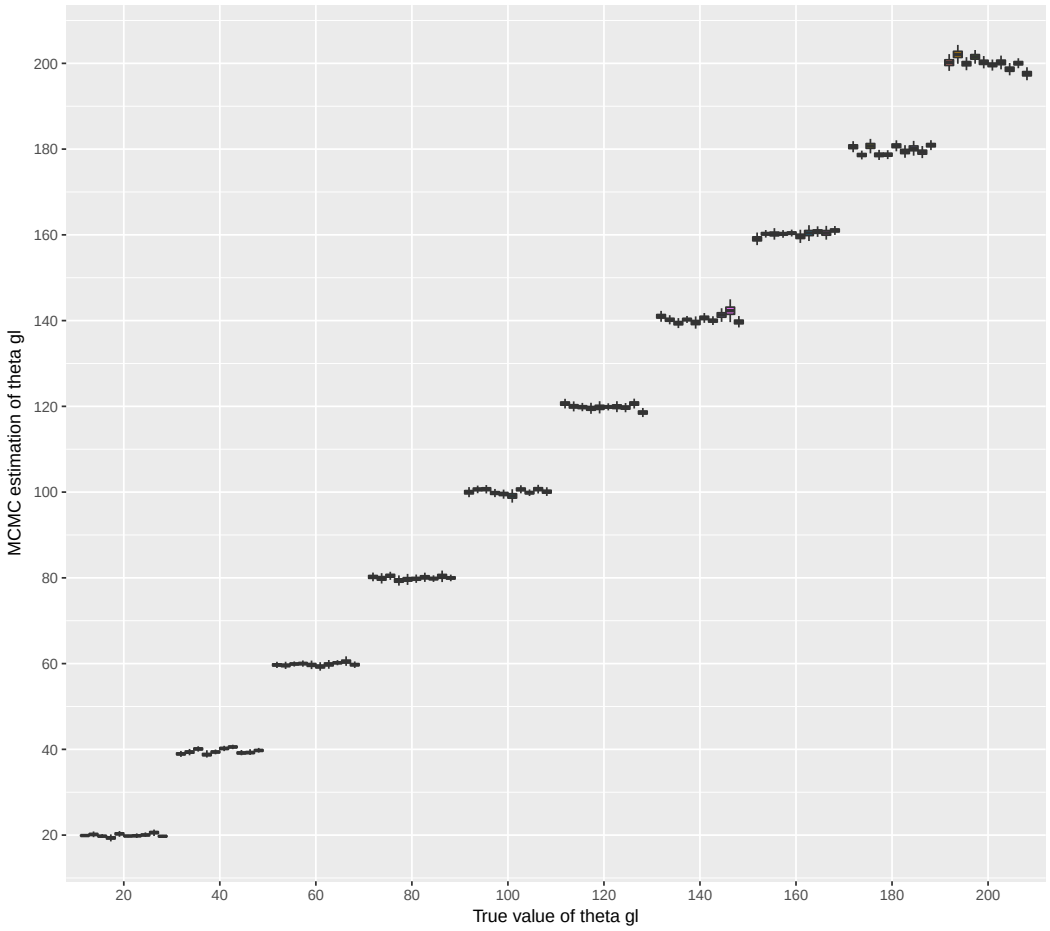


Figure C.29: The posterior estimates of parameter θ_{gl} from fitting the Time Dependent Distances model to 100 simulated datasets from the same model with varied input value for θ_{gl}, with 10 simulations for each value.

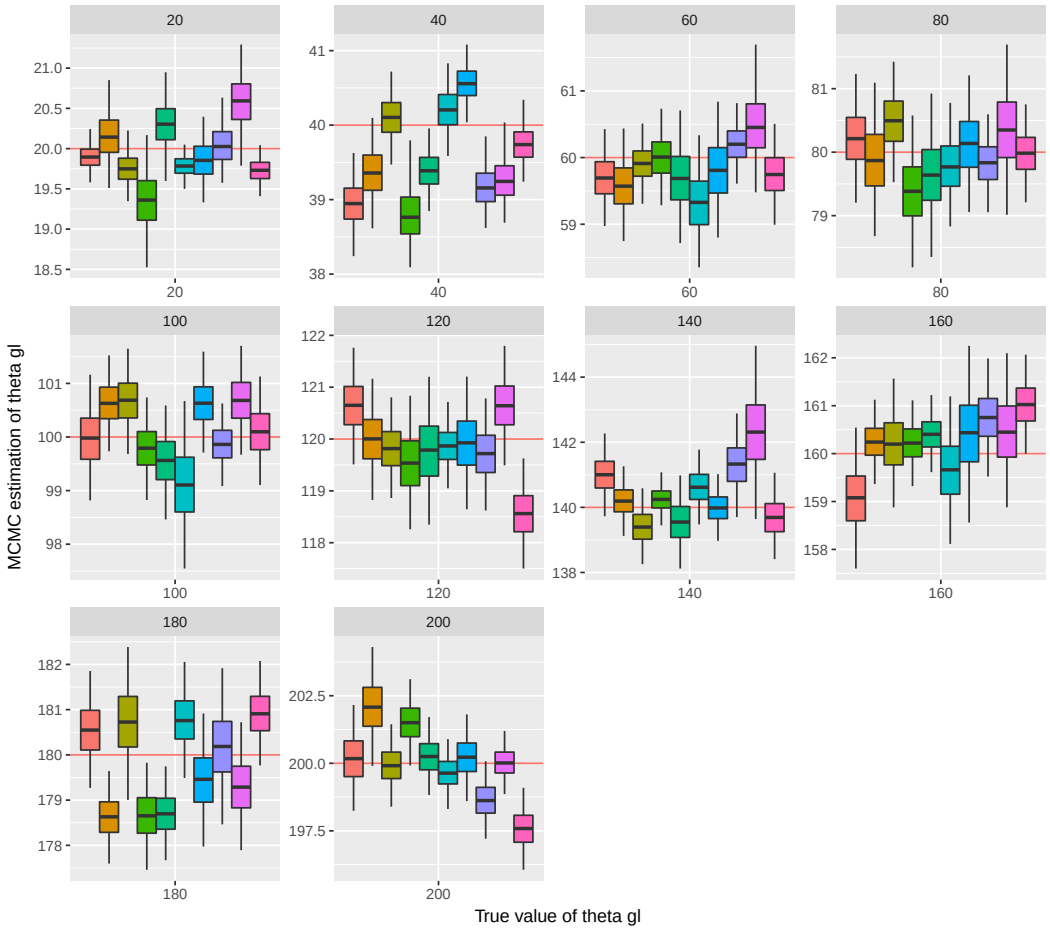


Figure C.30: The posterior estimates of parameter θ_{gl} from fitting the Time Dependent Distances model to 100 simulated datasets from the same model with varied input value for θ_{gl}, with 10 simulations for each value.

Appendix D

Graphs for network reconstruction for simulations for MRSA

D.0.1 Chain Error model

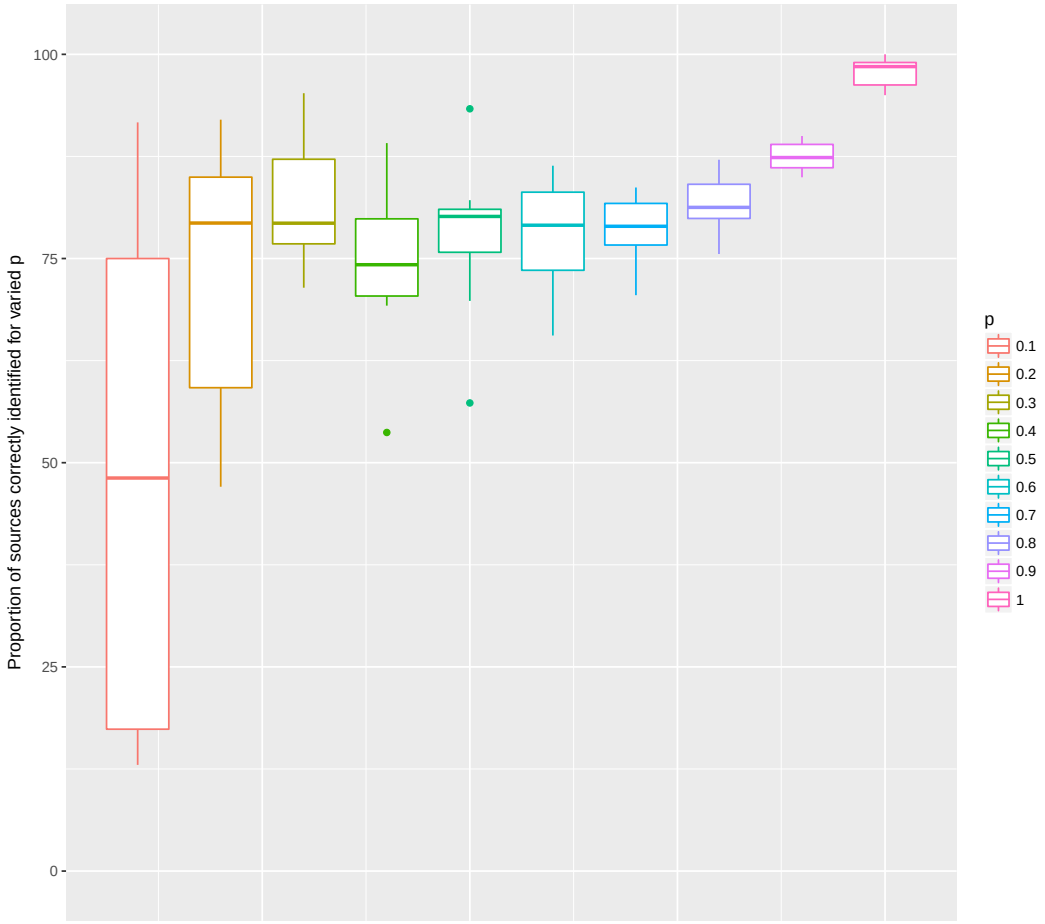


Figure D.1: Boxplots to show the proportion of infection sources for patients recovered correctly for simulations with varied values for parameter p for the Chain Error model.

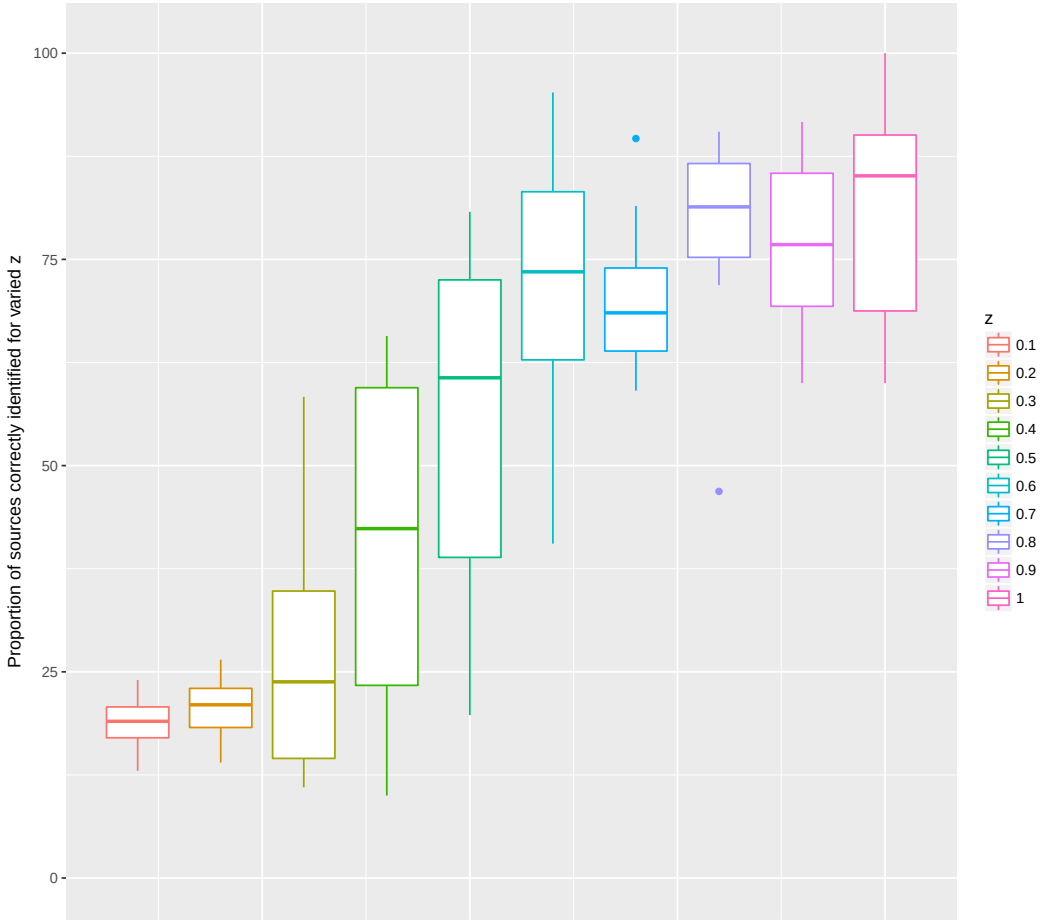


Figure D.2: Boxplots to show the proportion of infection sources for patients recovered correctly for simulations with varied values for parameter z for the Chain Error model.

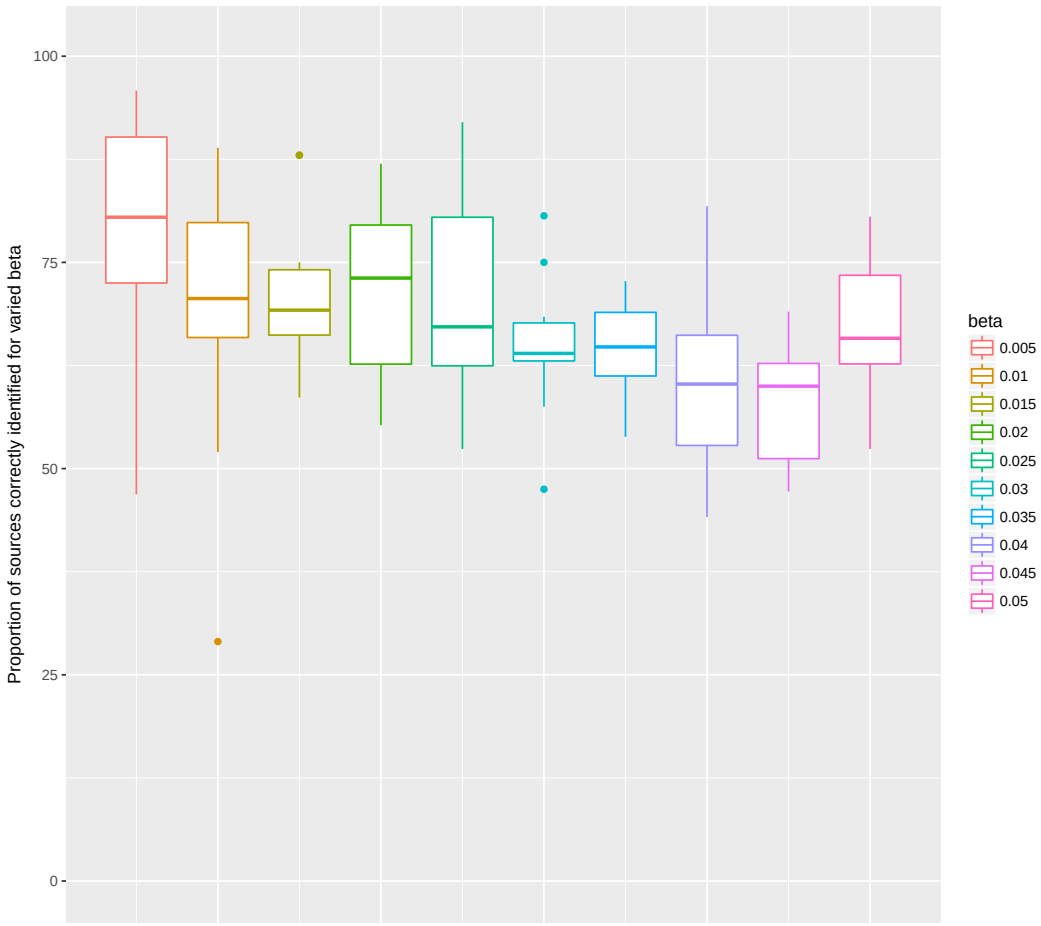


Figure D.3: Boxplots to show the proportion of infection sources for patients recovered correctly for simulations with varied values for parameter β for the Chain Error model.

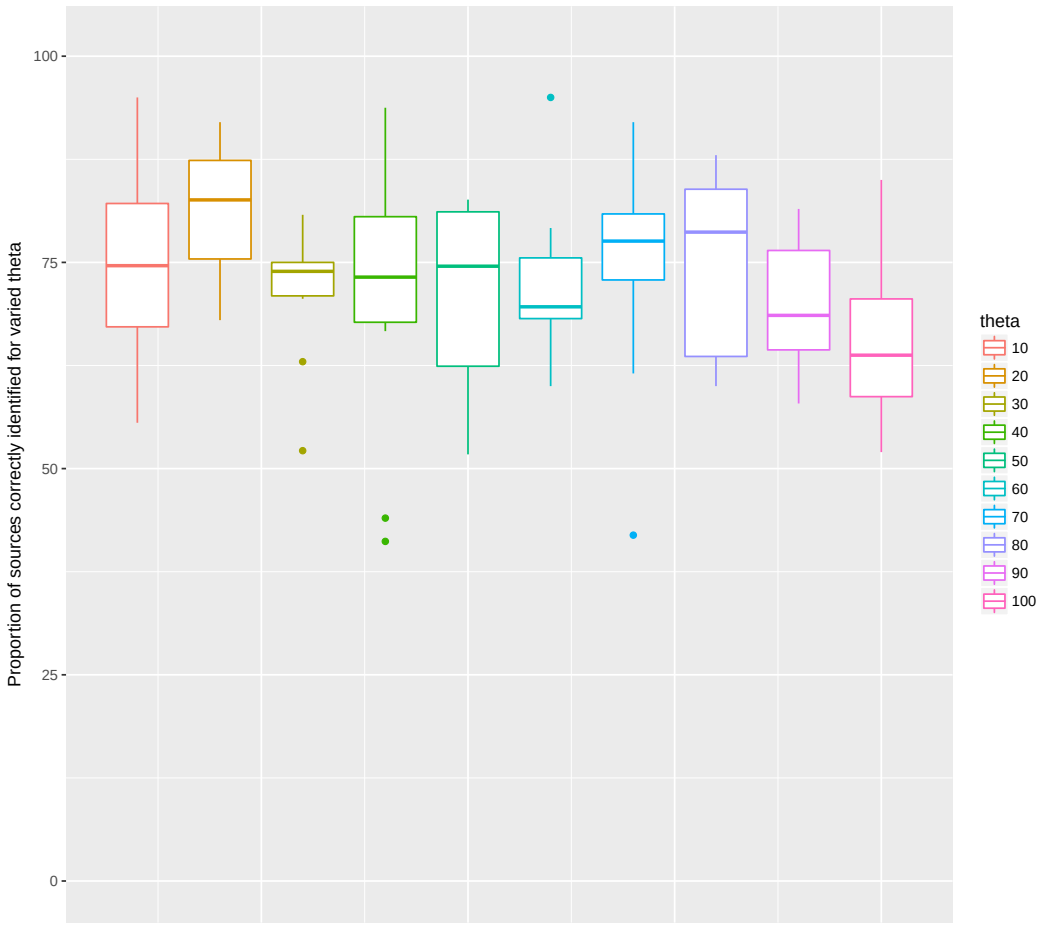


Figure D.4: Boxplots to show the proportion of infection sources for patients recovered correctly for simulations with varied values for parameter θ for the Chain Error model.

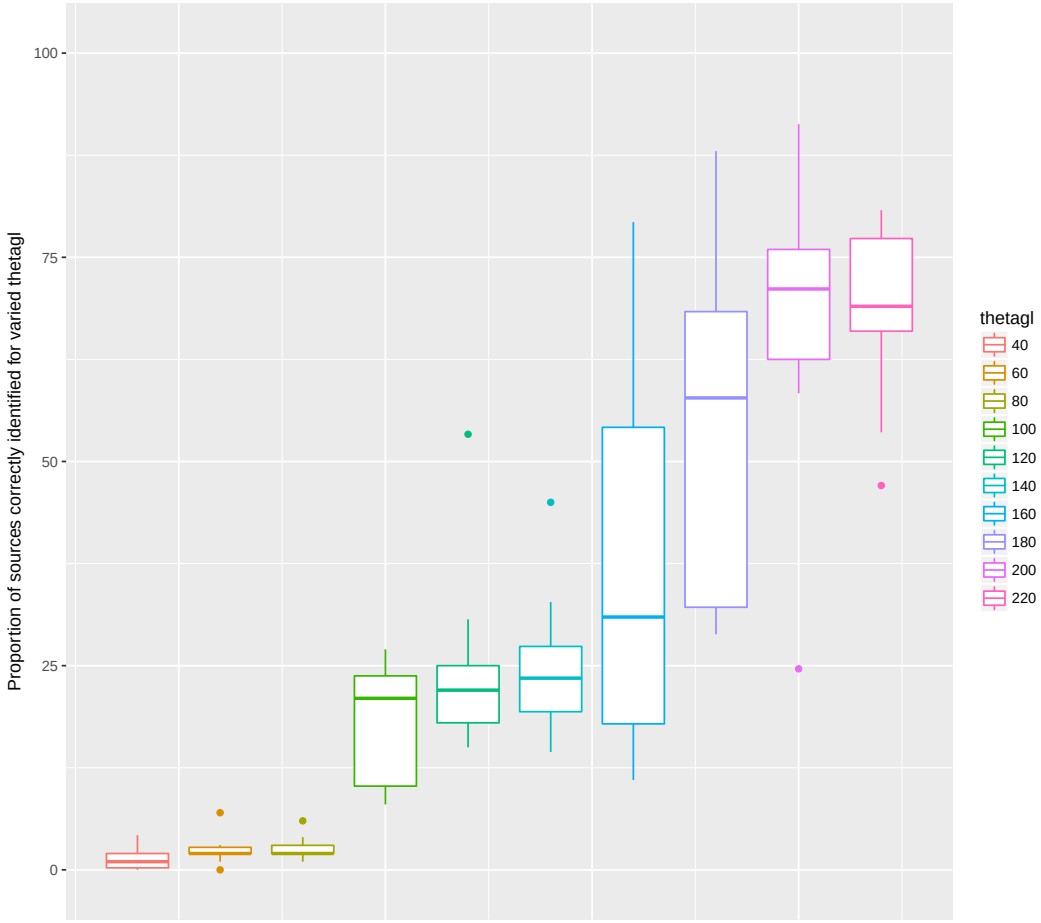


Figure D.5: Boxplots to show the proportion of infection sources for patients recovered correctly for simulations with varied values for parameter θ_{gl} for the Chain Error model.

D.0.2 Chain Poisson model

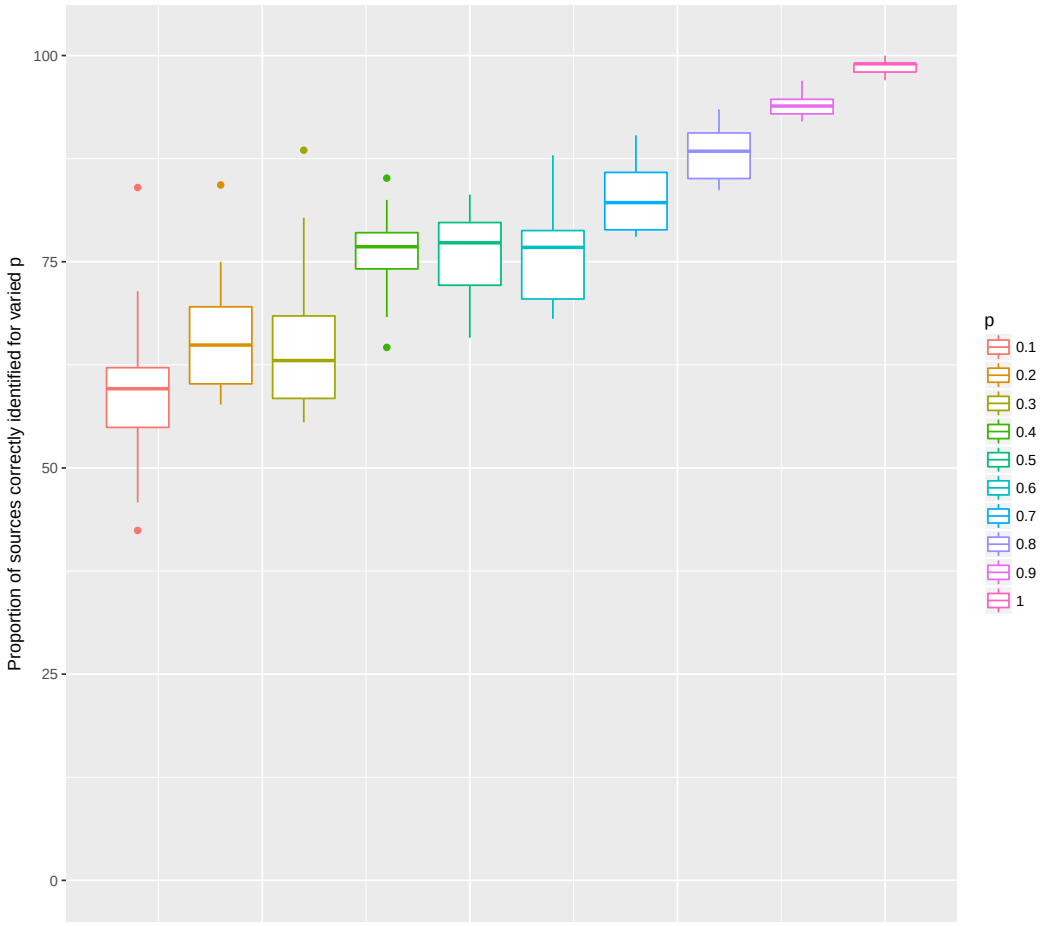


Figure D.6: Boxplots to show the proportion of infection sources for patients recovered correctly for simulations with varied values for parameter p for the Chain Poisson model.

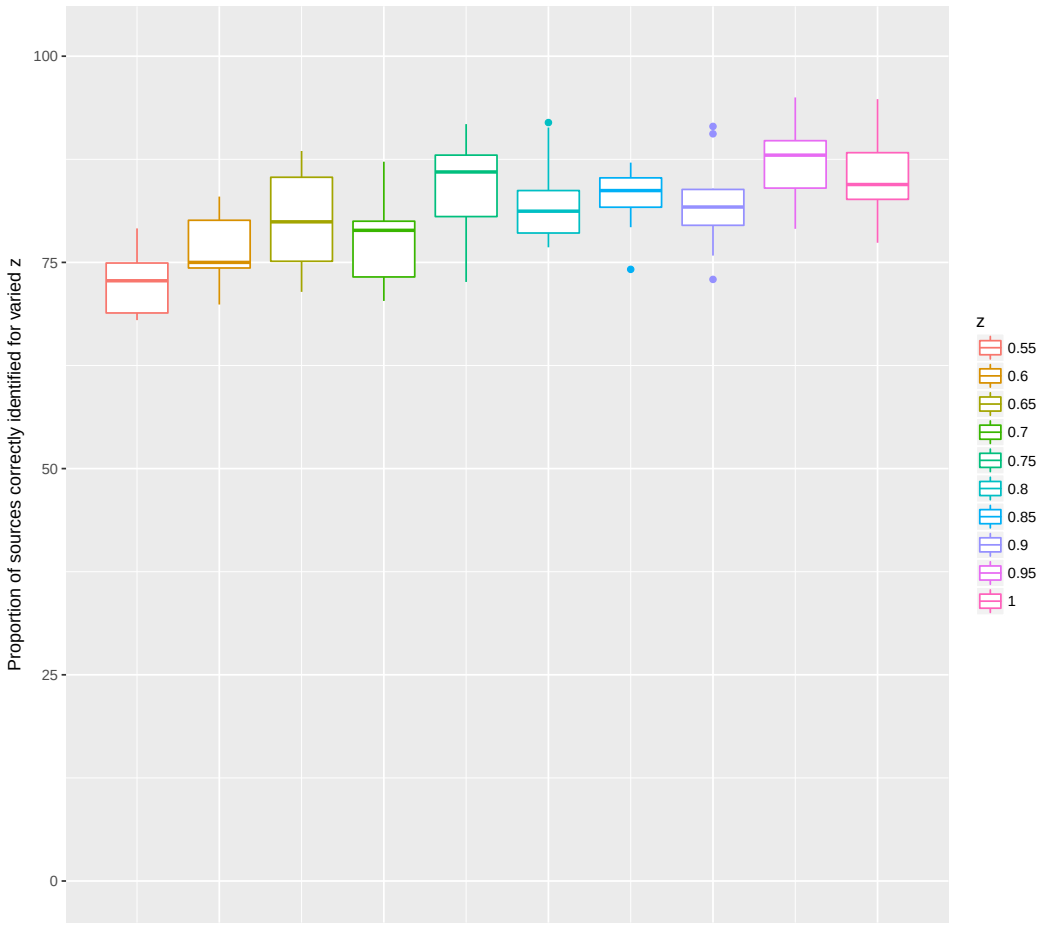


Figure D.7: Boxplots to show the proportion of infection sources for patients recovered correctly for simulations with varied values for parameter z for the Chain Poisson model.

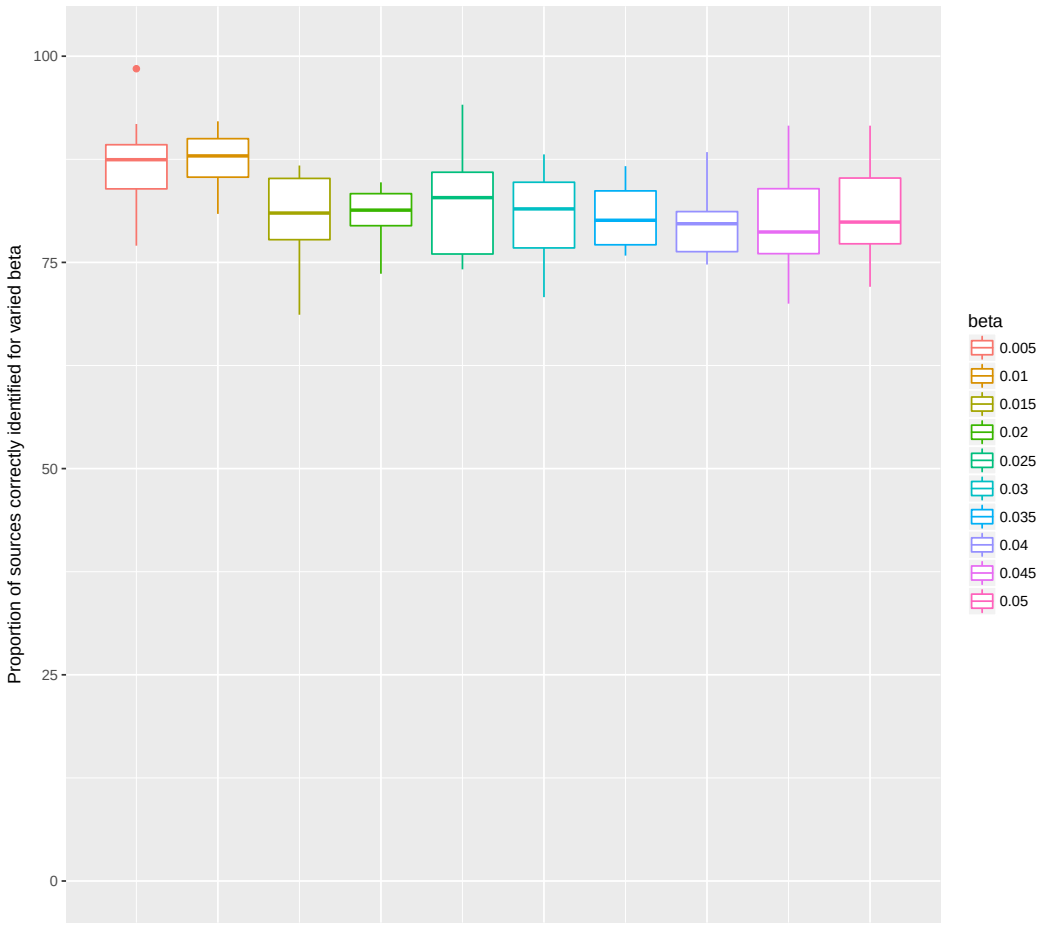


Figure D.8: Boxplots to show the proportion of infection sources for patients recovered correctly for simulations with varied values for parameter β for the Chain Poisson model.

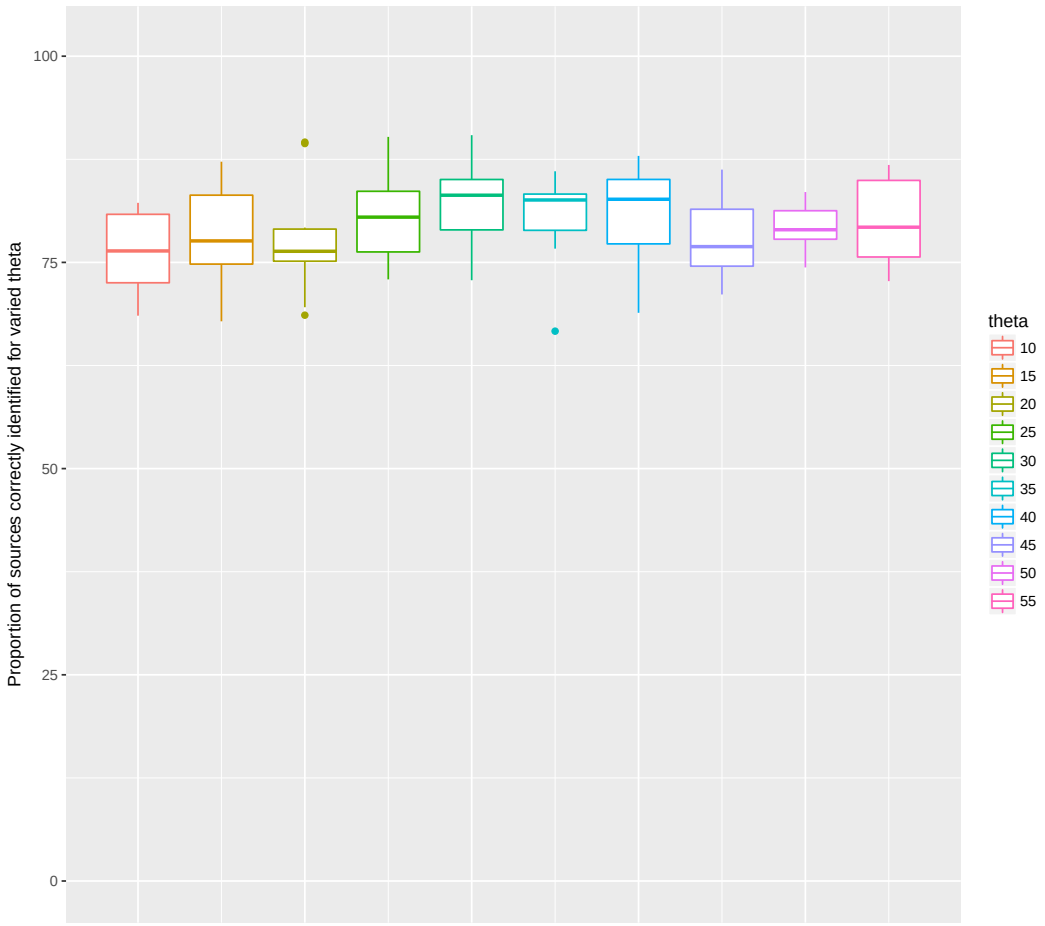


Figure D.9: Boxplots to show the proportion of infection sources for patients recovered correctly for simulations with varied values for parameter θ for the Chain Poisson model.

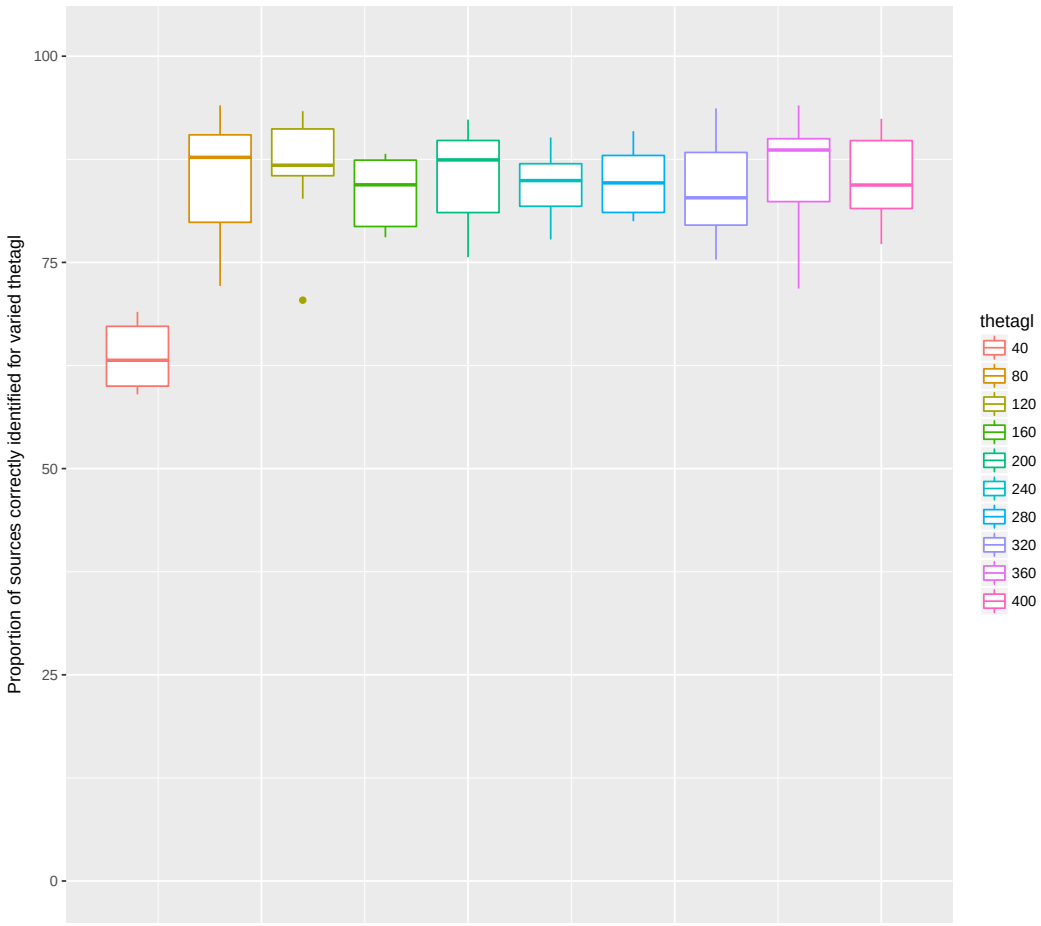


Figure D.10: Boxplots to show the proportion of infection sources for patients recovered correctly for simulations with varied values for parameter θ_{gl} for the Chain Poisson model.

D.0.3 Time Dependent Distances model

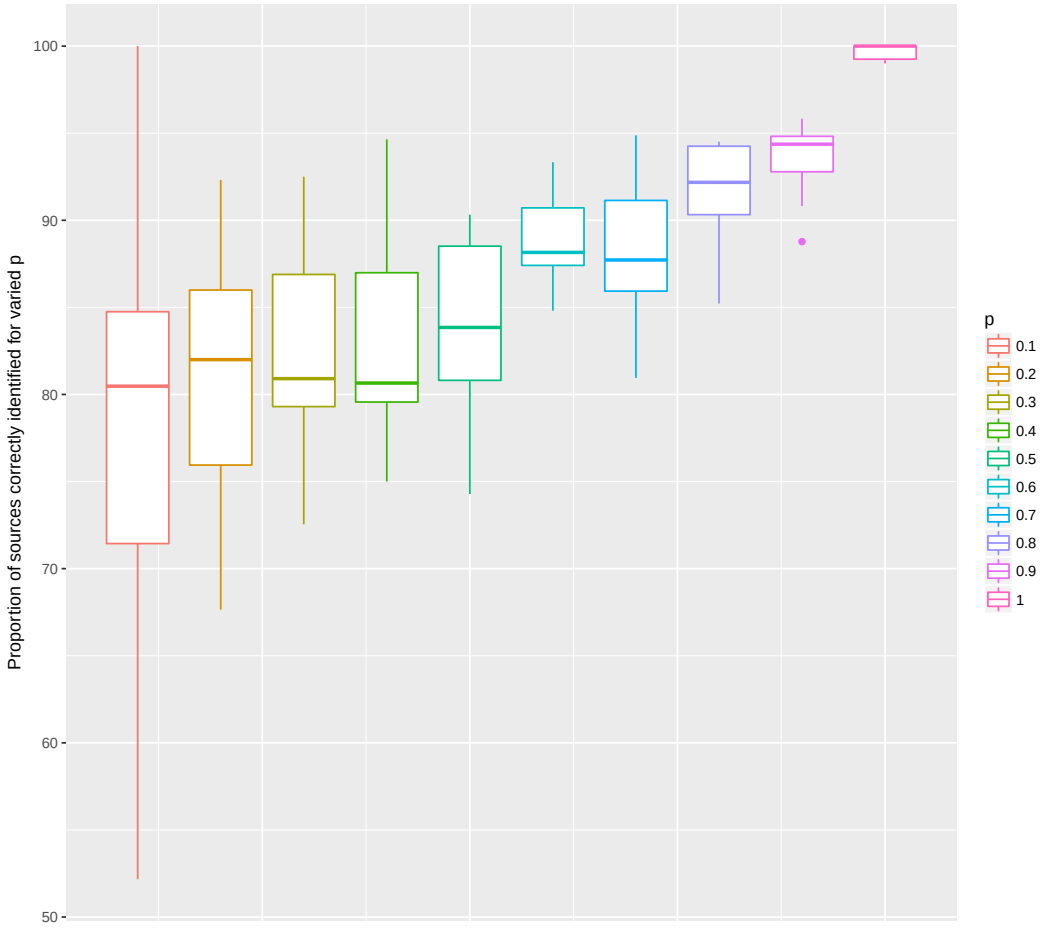


Figure D.11: Boxplots to show the proportion of infection sources for patients recovered correctly for simulations with varied values for parameter p for the Time Dependent Distances model.

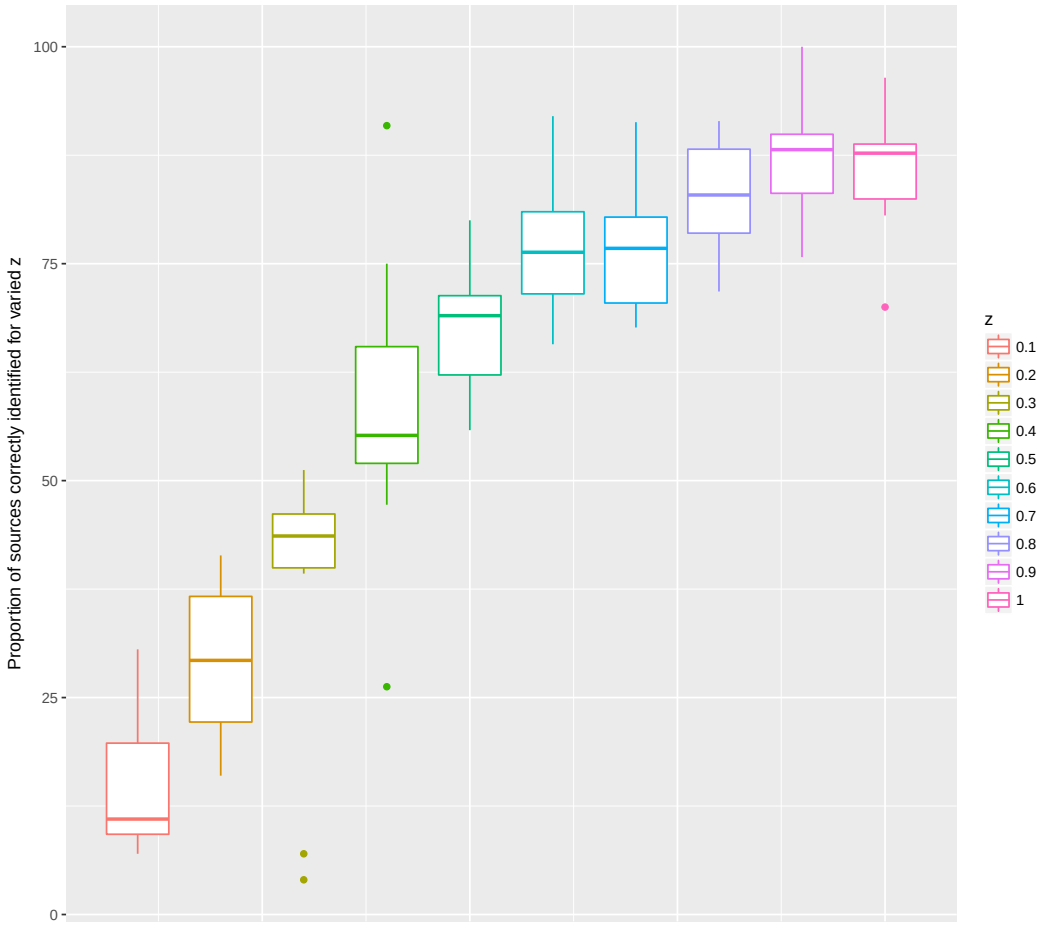


Figure D.12: Boxplots to show the proportion of infection sources for patients recovered correctly for simulations with varied values for parameter z for the Time Dependent Distances model.

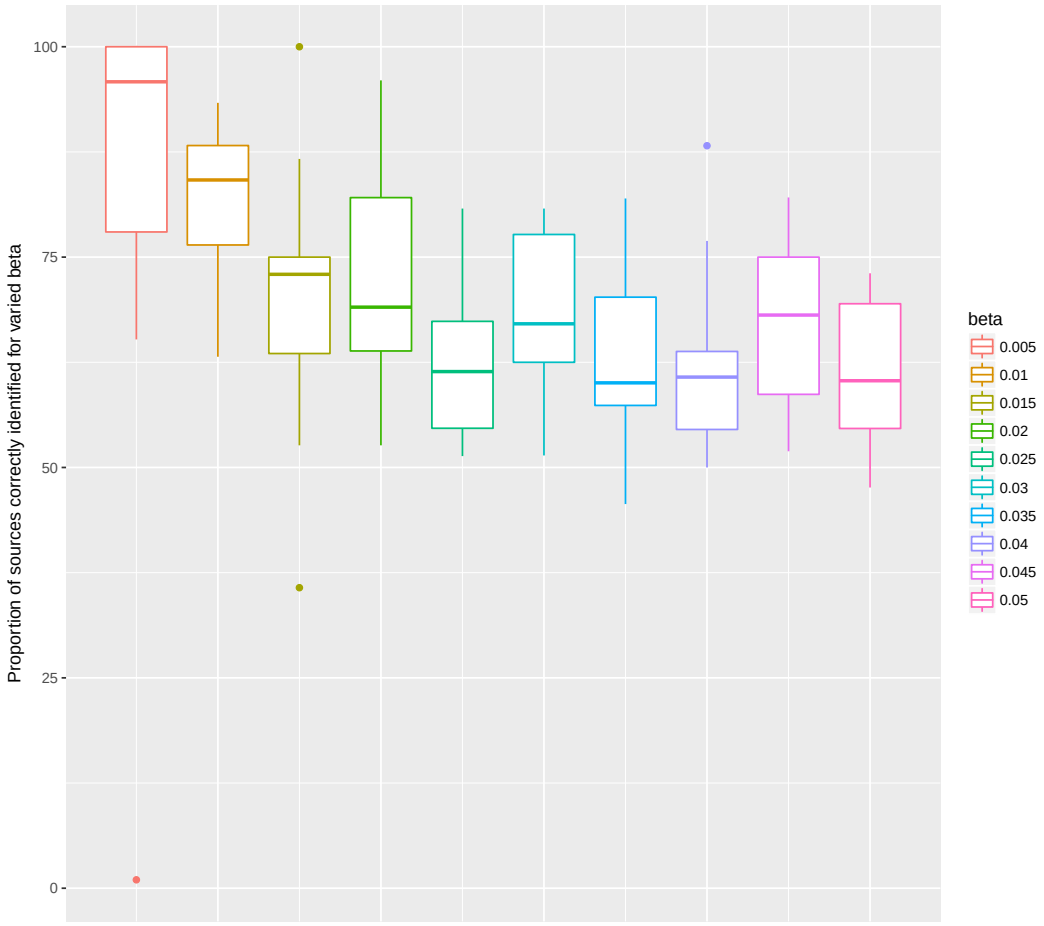


Figure D.13: Boxplots to show the proportion of infection sources for patients recovered correctly for simulations with varied values for parameter β for the Time Dependent Distances model.

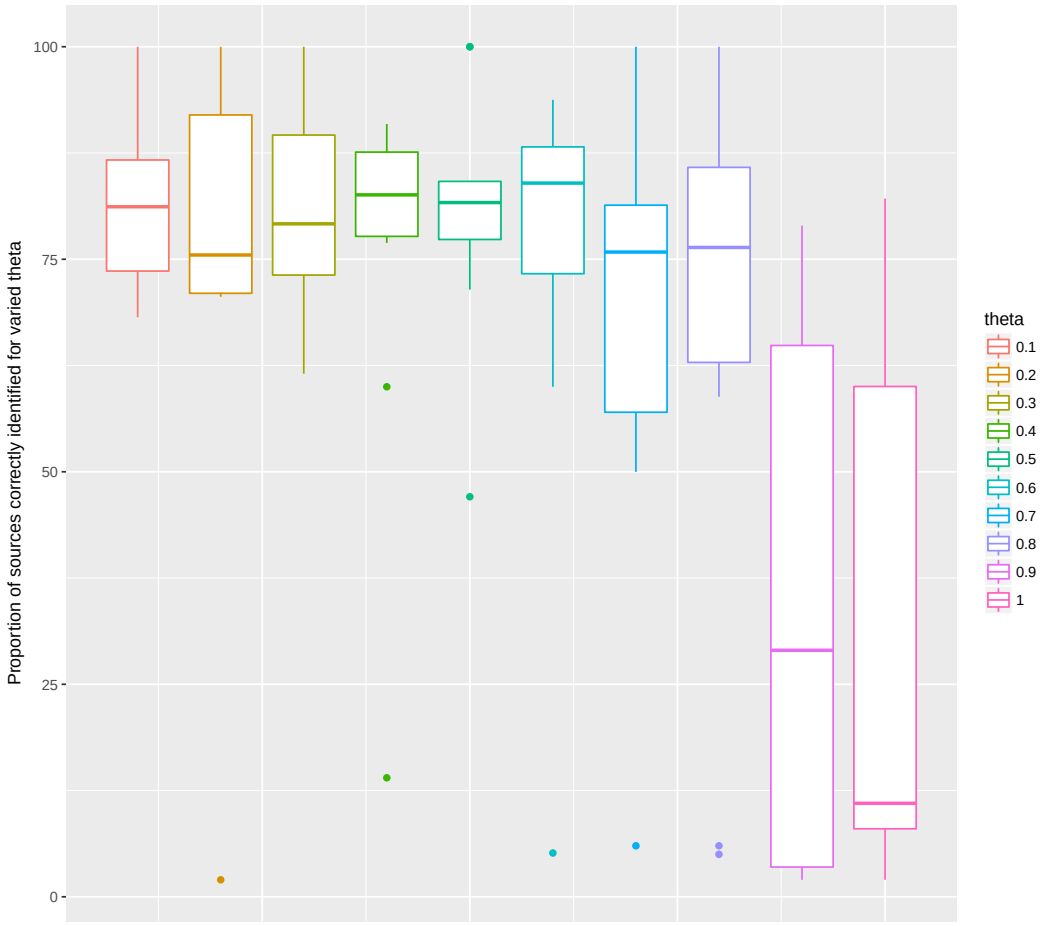


Figure D.14: Boxplots to show the proportion of infection sources for patients recovered correctly for simulations with varied values for parameter θ for the Time Dependent Distances model.

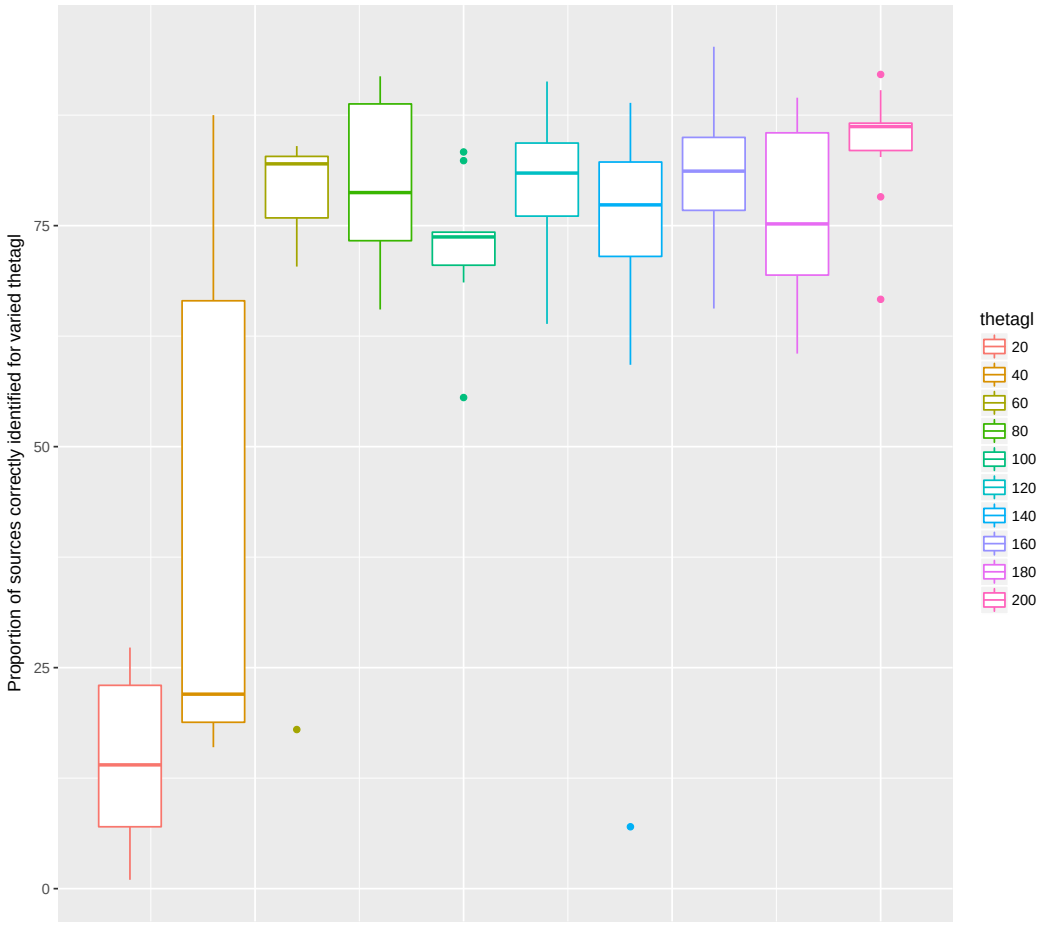


Figure D.15: Boxplots to show the proportion of infection sources for patients recovered correctly for simulations with varied values for parameter θ_{gl} for the Time Dependent Distances model.

Appendix E

Proposal ratios for augmented data step in the MCMC algorithm for avian influenza models

E.1 Changing genetic distances

If we propose to change the genetic distances from one farm to each other farm the proposal ratio is

$$q_{T,T^*} = \frac{\prod\limits_{j:\, i \neq j} \mathrm{P}\left[\Psi_{i,j} = \Psi^{a}_{i,j} | \Theta\right]}{\prod\limits_{j:\, i \neq j} \mathrm{P}\left[\Psi_{i,j} = \Psi^{a*}_{i,j} | \Theta\right]}.$$

addedThis is simply the probability of choosing the current genetic distances divided by the probability of choosing the proposed genetic distances. The probability of choosing the particular farm is the same for both this move and the reverse move and so cancels out.

E.2 Updating an infection time and resampling the sources

If we propose to change the infection time of one farm and resample the sources of infection of all farms the proposal ratio is

$$q_{T,T^*} = \frac{\mathrm{e}^{\left(\gamma\left(I_i - I_i^*\right)\right)} \prod_{i=1}^{n_I} \frac{\beta_{i,s(i)}}{\sum_L \beta_{i,j}}}{\prod_{i=1}^{n_I} \frac{\beta_{i,s^*(i)}}{\sum_{L^*} \beta_{i,j}}}.$$

The probability of proposing this move is

- the probability of choosing this particular farm from the set of all farms which ever get infected, which is $1/n_I$,

- multiplied by the probability of choosing the proposed infection time, I_i^*, of the farmby subtracting a draw from an exponential distribution with parameter γ from the latest point at which the farm may have been susceptible, f_i. This probability is $\gamma \exp(-\gamma(f_i - I_i^*))$,
- multiplied by the probability of choosing each source of infection for each farm. Since each farm's source is chosen with probability weight $\frac{\beta_{i,j}}{\sum_{j:I_j<E_i^*<R_j}\beta_{i,j}}$ this probability will be the product $\prod_{i=1}^{n_I}\frac{\beta_{i,s^*(i)}}{\sum_{L^*}\beta_{i,j}}$.

The probability of proposing the reverse move, which is changing the proposed infection time of this farm to the current infection time and resampling all the sources, is

- the probability of choosing this particular farm from the set of all farms which ever get infected, which is $1/n_I$,
- multiplied by the probability of choosing the current infection time, I_i, of the farm by subtracting a draw from an exponential distribution with parameter γ from the latest time at which the farm may have been susceptible, f_i. This probability is $\gamma \exp(-\gamma(f_i - I_i))$,
- multiplied by the probability of choosing each source of infection for each farm. Since each farm's source is chosen with probability weight $\frac{\beta_{i,j}}{\sum_{j:I_j<E_i<R_j}\beta_{i,j}}$ this probability will be the product $\prod_{i=1}^{n_I}\frac{\beta_{i,s(i)}}{\sum_{L}\beta_{i,j}}$.

The probability of choosing the farm cancels out, and $\frac{\gamma\exp(-\gamma(f_i-I_i))}{\gamma\exp(-\gamma(f_i-I_i^*))} = \exp(\gamma((f_i - I_i^*) - (f_i - I_i))) = \exp(\gamma(I_i - I_i^*)$.

E.3 Changing the infection time and the source of one farm

The proposal ratio for the move in which we pick one farm which is not the initial infective and change their infection time and source of infection is

$$q_{T,T^*} = \frac{\mathrm{e}^{\left(\gamma\left(I_i - I_i^*\right)\right)}\frac{\beta_{i,s(i)}}{\sum_L \beta_{i,j}}}{\frac{\beta_{i,s^*(i)}}{\sum_{L^*}\beta_{i,j}}}.$$

The probability of proposing this move is

- the probability of choosing the farm from the set of all infected farm excluding the initial infective, which is $1/(n_I - 1)$,

- multiplied by the probability of setting the new infection time, I_i^*, by subtracting a drawn from a truncated exponential distribution (truncated at $(f_i - t_1)$ so that the time can not be prior or equal to the infection time of the initial infective) from the latest time at which the farm may have been susceptible, f_i. This probability is $\frac{\gamma \mathrm{e}^{(-\gamma(f_i - I_i^*))}}{1-\mathrm{e}^{(-\gamma(f_1 - t_1))}}$,
- multiplied by the probability of choosing the farm's new source of infection from the farms which were infectious at I_i^*. This probability is $\frac{\beta_{i,s^*(i)}}{\sum_{L^*} \beta_{i,j}}$.

The probability of making the reverse move is

- the probability of choosing the farm from the set of all infected farm excluding the initial infective, which is $1/(n_I - 1)$,
- multiplied by the probability of setting the current infection time, I_i, by subtracting a drawn from a truncated exponential distribution (truncated at $(f_i - t_1)$ so that the time can not be prior or equal to the infection time of the initial infective) from the latest time at which the farm may have been susceptible, f_i. This probability is $\frac{\gamma \mathrm{e}^{(-\gamma(f_i - I_i))}}{1-\mathrm{e}^{(-\gamma(f_1 - t_1))}}$,
- multiplied by the probability of choosing the farm's current source of infection from the farms which were infectious at I_i. This probability is $\frac{\beta_{i,s(i)}}{\sum_{L} \beta_{i,j}}$.

The probability of choosing the farm cancels out, and $\left(\frac{\gamma \mathrm{e}^{(-\gamma(f_i - I_i^*))}}{1-\mathrm{e}^{(-\gamma(f_1 - t_1))}} \Big/ \frac{\gamma \mathrm{e}^{(-\gamma(f_i - I_i))}}{1-\mathrm{e}^{(-\gamma(f_1 - t_1))}} \right) = \mathrm{e}^{(\gamma(I_i - I_i^*))}$.

E.4 Change the time of the initial infection

If we propose to set a new time for the initial infection by subtracting a draw from an exponential distribution with parameter γ from the last time at which the farm could have been susceptible, f_i, the proposal ratio is simply

$$q_{T,T^*} = \mathrm{e}^{(\gamma(I_l - I_l^*))}$$

where I_l^* is the proposed initial infection time of initial infective l, and I_l is the current initial infection time of the initial infective, since $\frac{\gamma \exp(-\gamma(f_i - I_i))}{\gamma \exp(-\gamma(f_i - I_i^*))} = \exp(\gamma((f_i - I_i^*) - (f_i - I_i))) = \exp(\gamma(I_i - I_i^*)$.

E.5 Block update of an infection time, all sources, and the genetic parameters

This move is the same as the move described in section E.2 and additionally includes an update of the genetics parameters. Therefore the proposal ratio is equal to that described in section E.2 multiplied by the proposal ratio of the parameter updates. Since the parameters θ^*, θ^*_{gl} are drawn from gamma distributions this gives

$$q_{T,T^*} = \frac{\mathrm{e}^{\left(\gamma\left(I_i - I_i^*\right)\right)} \prod_{i=1}^{n_I} \frac{\beta_{i,s(i)}}{\sum_L \beta_{i,j}}}{\prod_{i=1}^{n_I} \frac{\beta_{i,s^*(i)}}{\sum_{L^*} \beta_{i,j}}} \times \frac{\frac{\zeta_\theta^{\mu_\theta}}{\Gamma(\mu_\theta)} \theta^{\mu_\theta - 1} \mathrm{e}^{-\zeta_\theta \theta}}{\frac{\zeta_\theta^{*(\mu_\theta^*)}}{\Gamma(\mu_\theta^*)} \theta^{*(\mu_\theta^* - 1)} \mathrm{e}^{-\zeta_\theta^* \theta^*}} \times \frac{\frac{\zeta_{\theta_{gl}}^{\mu_{\theta_{gl}}}}{\Gamma(\mu_{\theta_{gl}})} \theta_{gl}^{\mu_{\theta_{gl}} - 1} \mathrm{e}^{-\zeta_{\theta_{gl}} \theta_{gl}}}{\frac{\zeta_{\theta_{gl}}^{*(\mu_{\theta_{gl}}^*)}}{\Gamma(\mu_{\theta_{gl}}^*)} \theta_{gl}^{*(\mu_{\theta_{gl}}^* - 1)} \mathrm{e}^{-\zeta_{\theta_{gl}}^* \theta_{gl}^*}}.$$

Appendix F

Graphs for estimation of simulation parameters for avian influenza

F.1 Chain Error model

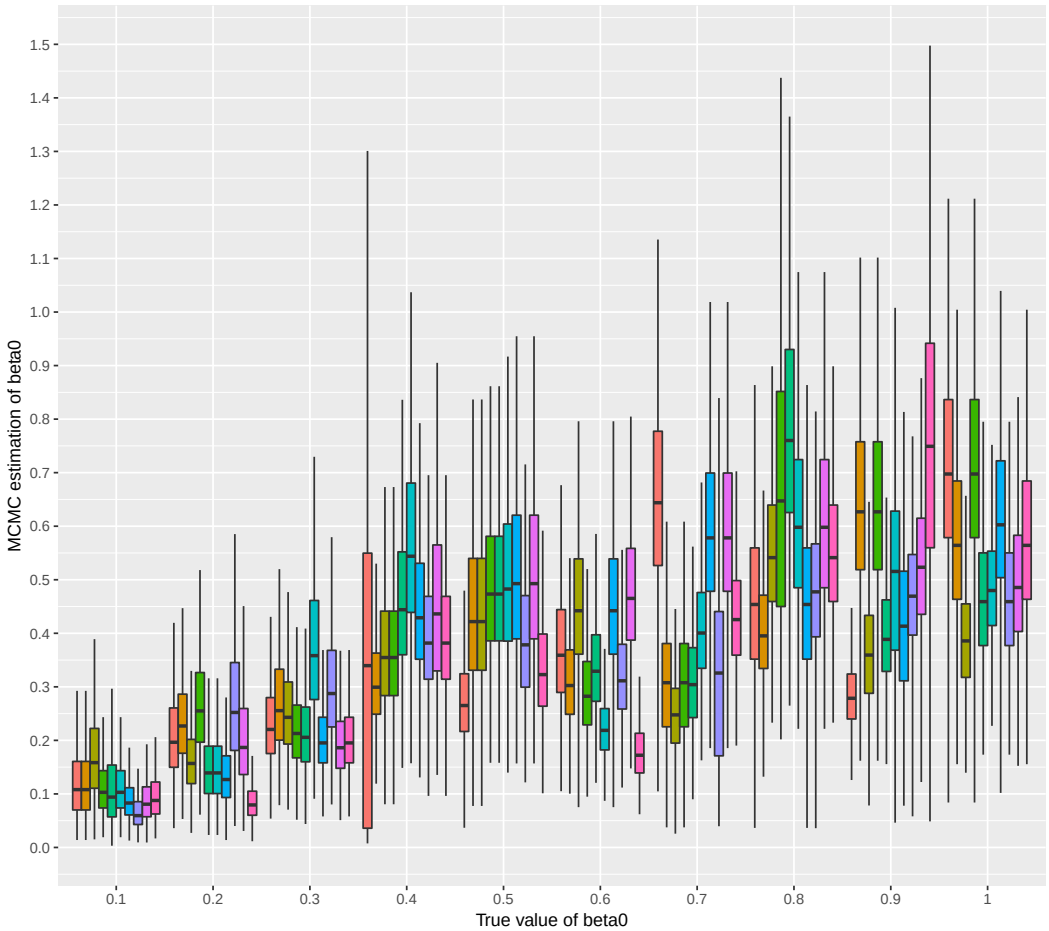


Figure F.1: The posterior estimates of parameter β_0 from fitting the Chain Error model to 100 simulated datasets from the same model with varied input value for β_0, with 10 simulations for each value.

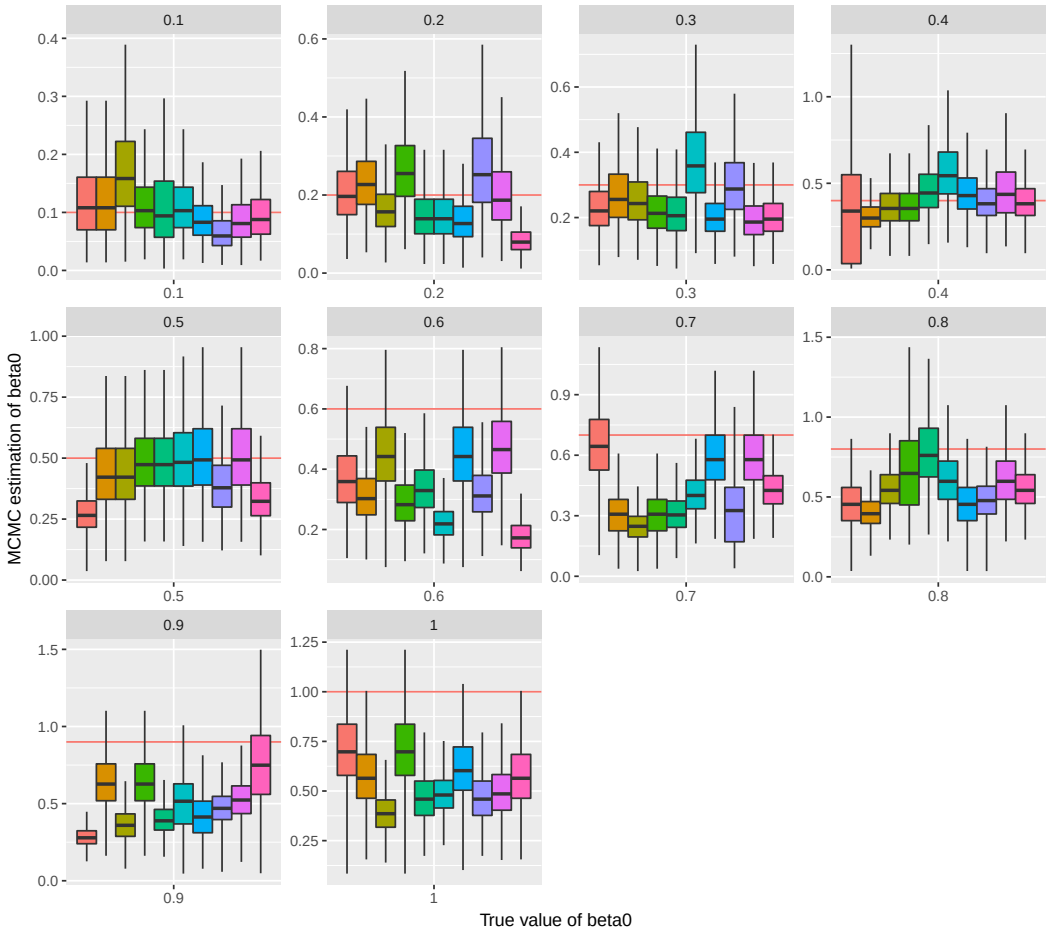


Figure F.2: The posterior estimates of parameter β_0 from fitting the Chain Error model to 100 simulated datasets from the same model with varied input value for β_0, with 10 simulations for each value.

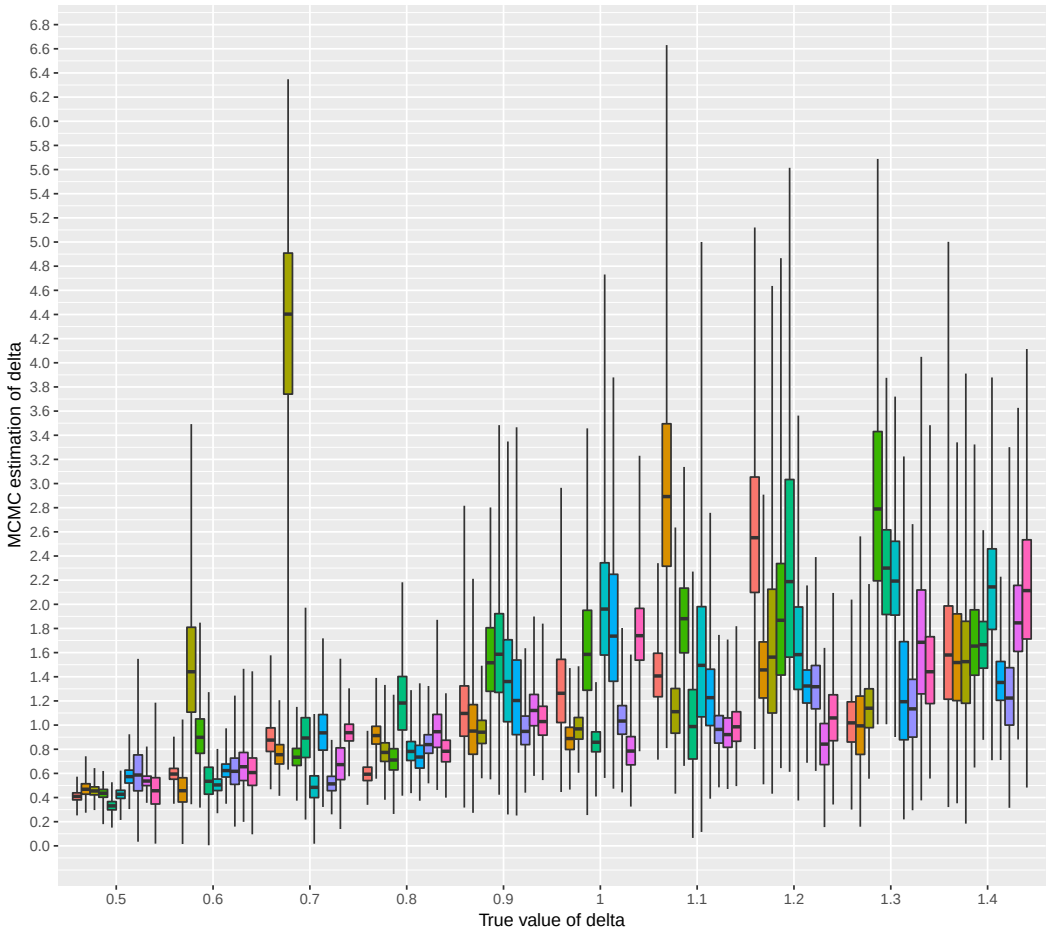


Figure F.3: The posterior estimates of parameter δ from fitting the Chain Error model to 100 simulated datasets from the same model with varied input value for δ, with 10 simulations for each value.

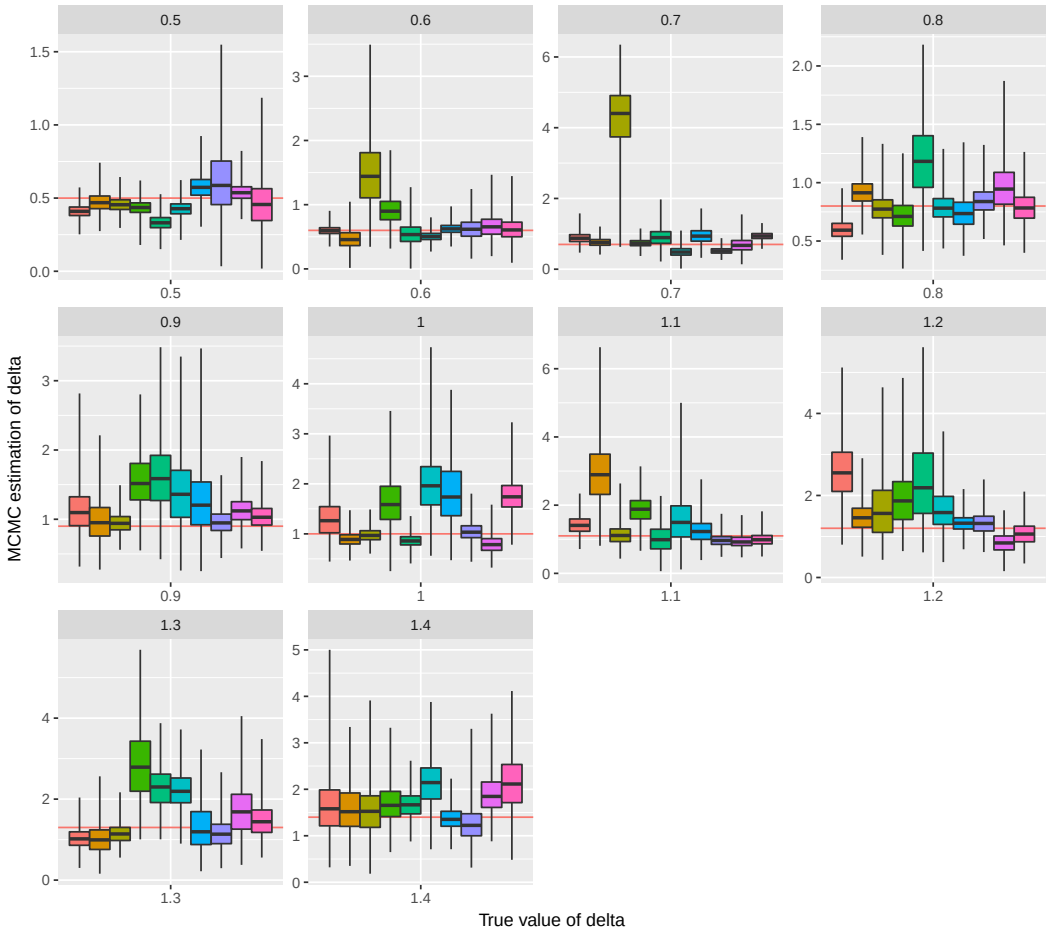


Figure F.4: The posterior estimates of parameter δ from fitting the Chain Error model to 100 simulated datasets from the same model with varied input value for δ, with 10 simulations for each value.

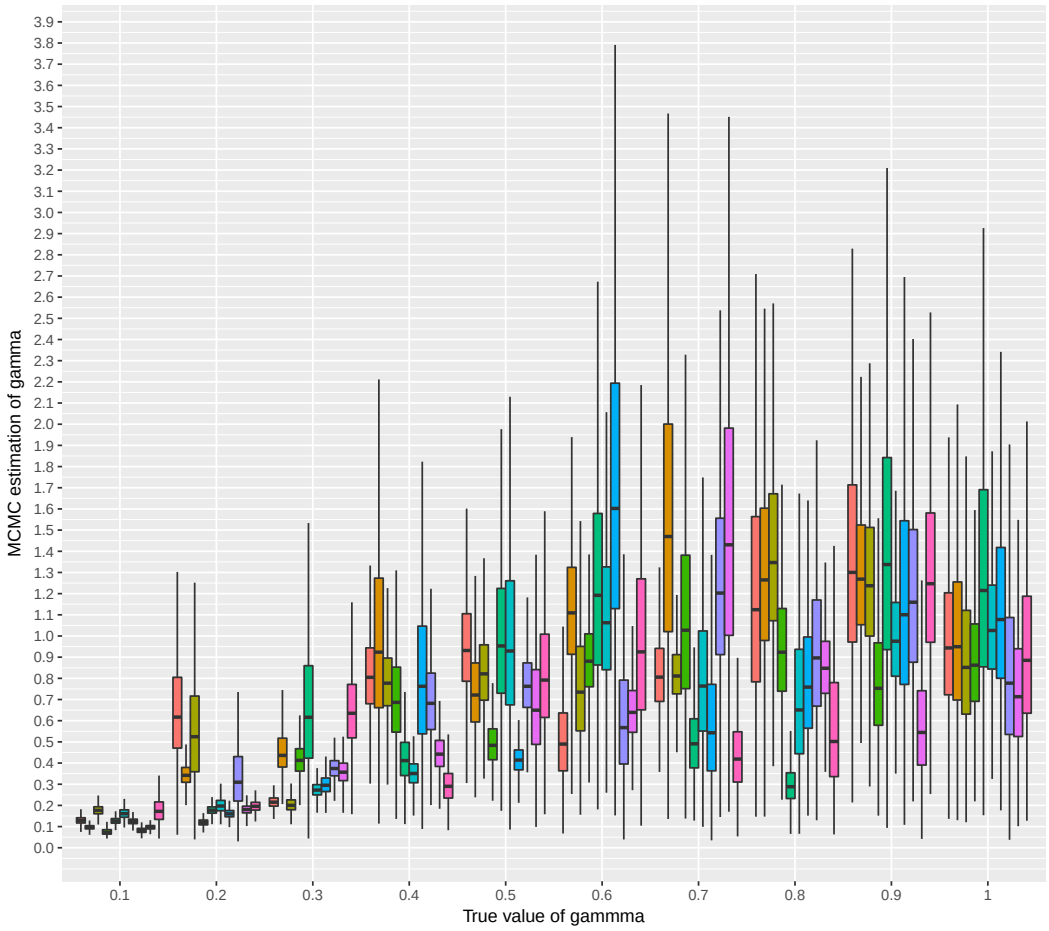


Figure F.5: The posterior estimates of parameter γ from fitting the Chain Error model to 100 simulated datasets from the same model with varied input value for γ, with 10 simulations for each value.

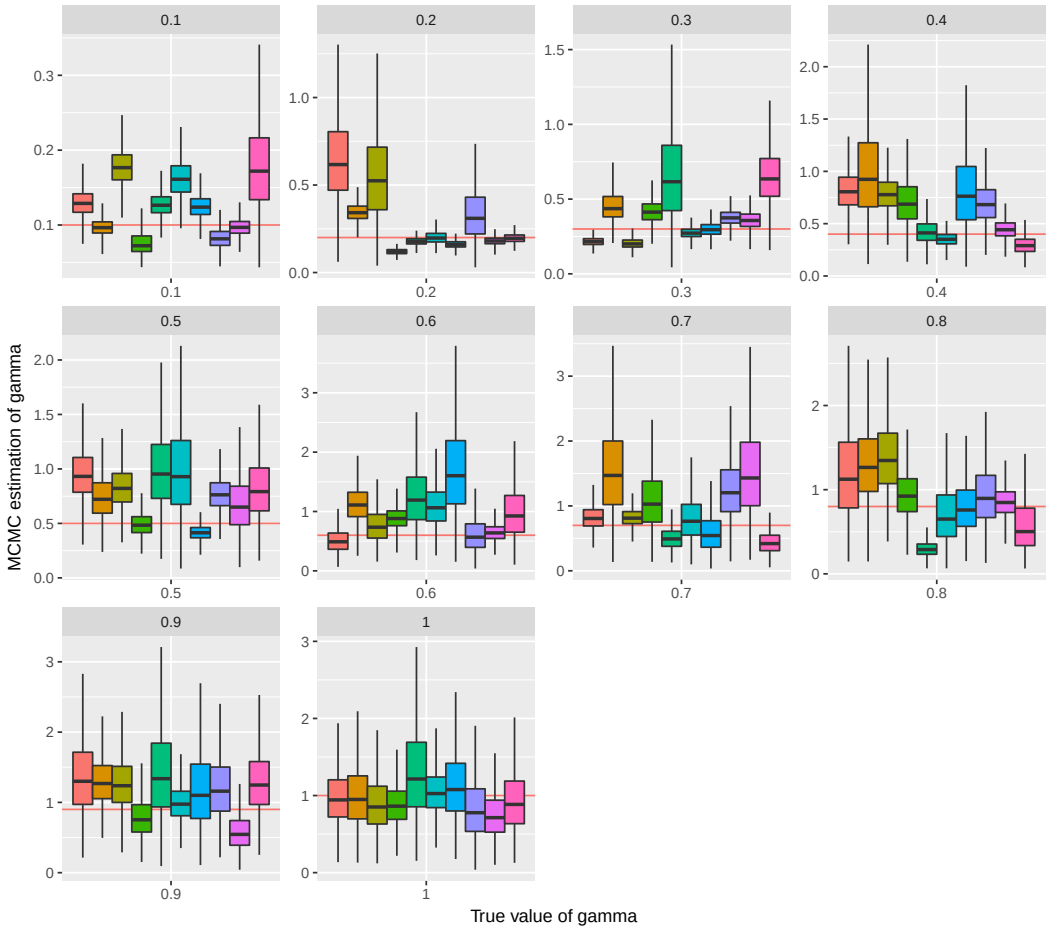


Figure F.6: The posterior estimates of parameter γ from fitting the Chain Error model to 100 simulated datasets from the same model with varied input value for γ, with 10 simulations for each value.

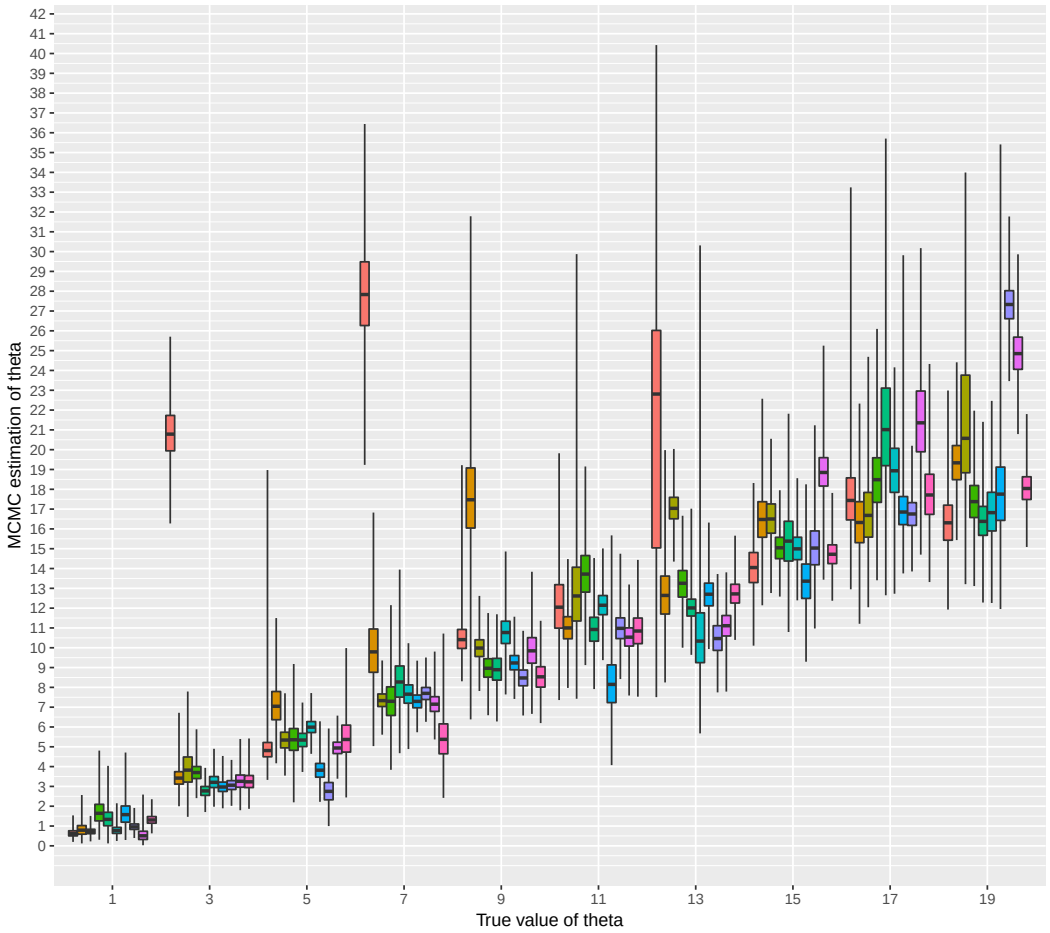


Figure F.7: The posterior estimates of parameter θ from fitting the Chain Error model to 100 simulated datasets from the same model with varied input value for θ, with 10 simulations for each value.

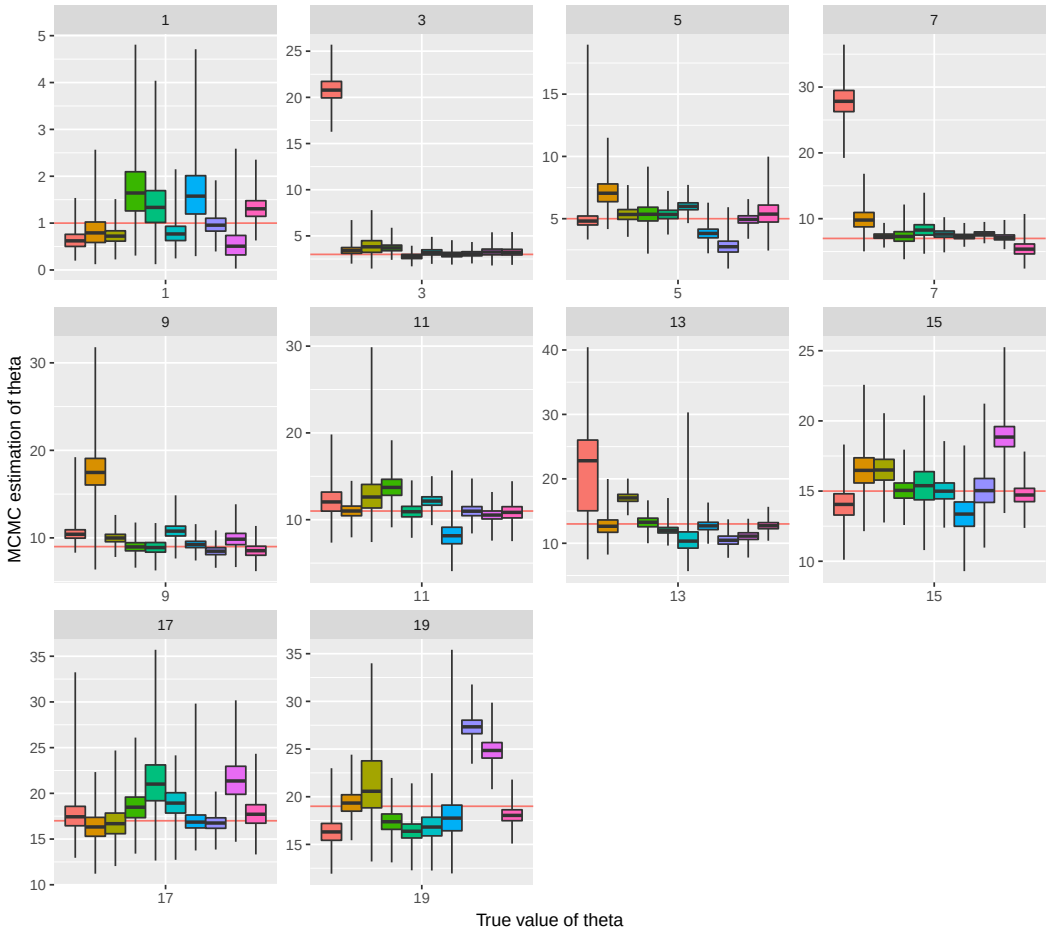


Figure F.8: The posterior estimates of parameter θ from fitting the Chain Error model to 100 simulated datasets from the same model with varied input value for θ, with 10 simulations for each value.

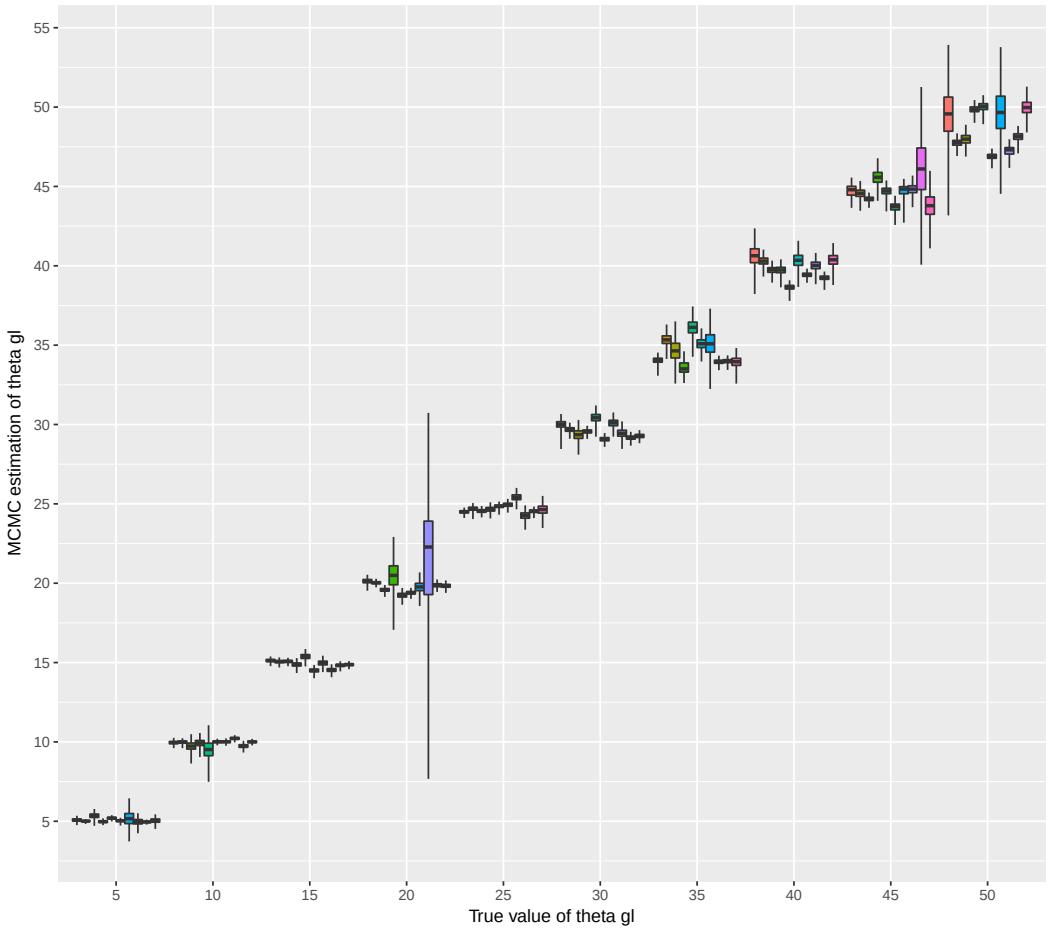


Figure F.9: The posterior estimates of parameter θ_{gl} from fitting the Chain Error model to 100 simulated datasets from the same model with varied input value for θ_{gl}, with 10 simulations for each value.

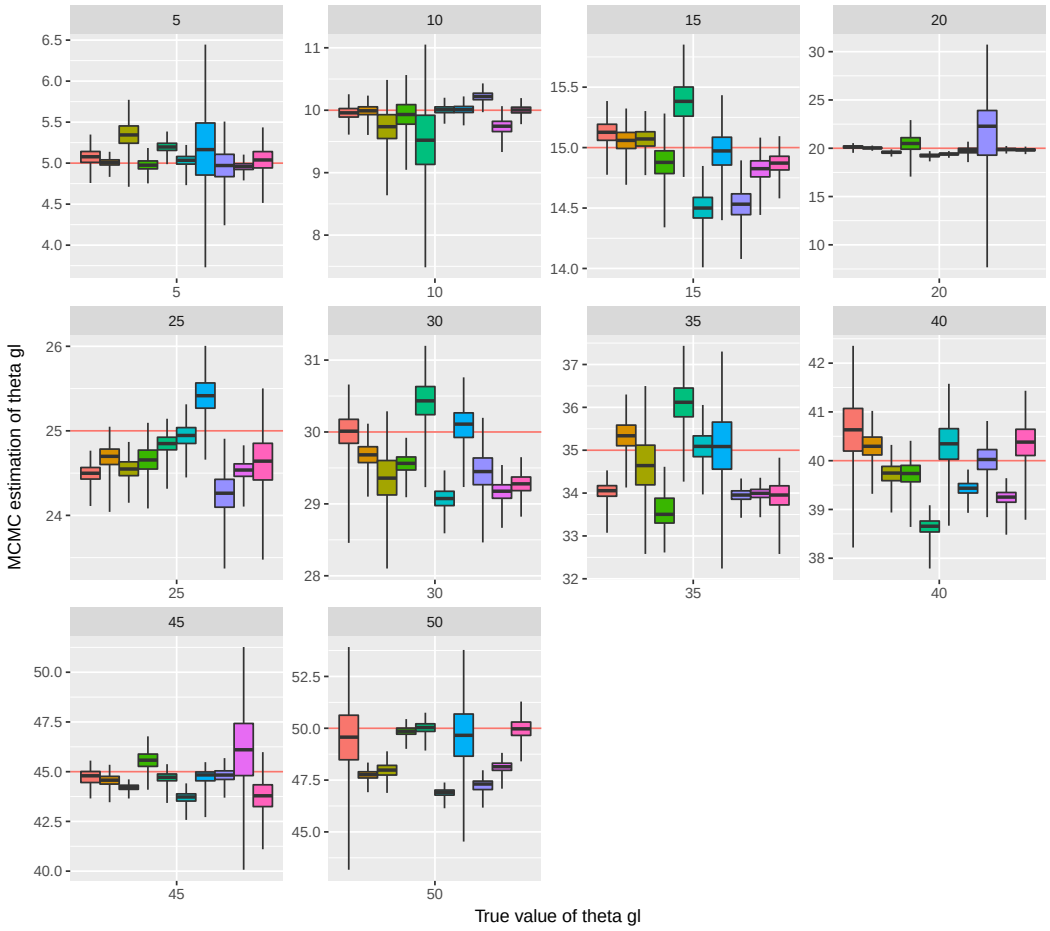


Figure F.10: The posterior estimates of parameter θ_{gl} from fitting the Chain Error model to 100 simulated datasets from the same model with varied input value for θ_{gl}, with 10 simulations for each value.

F.2 Chain Poisson model

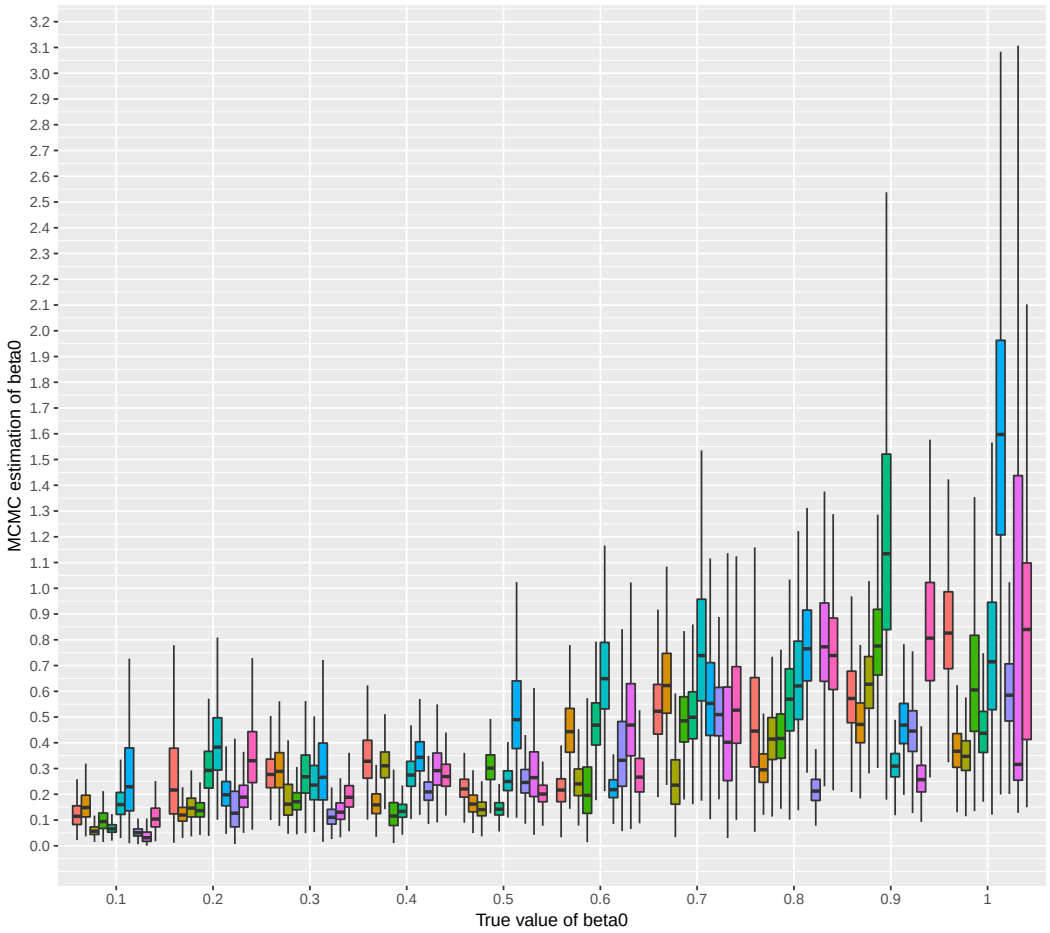


Figure F.11: The posterior estimates of parameter β_0 from fitting the Chain Poisson model to 100 simulated datasets from the same model with varied input value for β_0, with 10 simulations for each value.

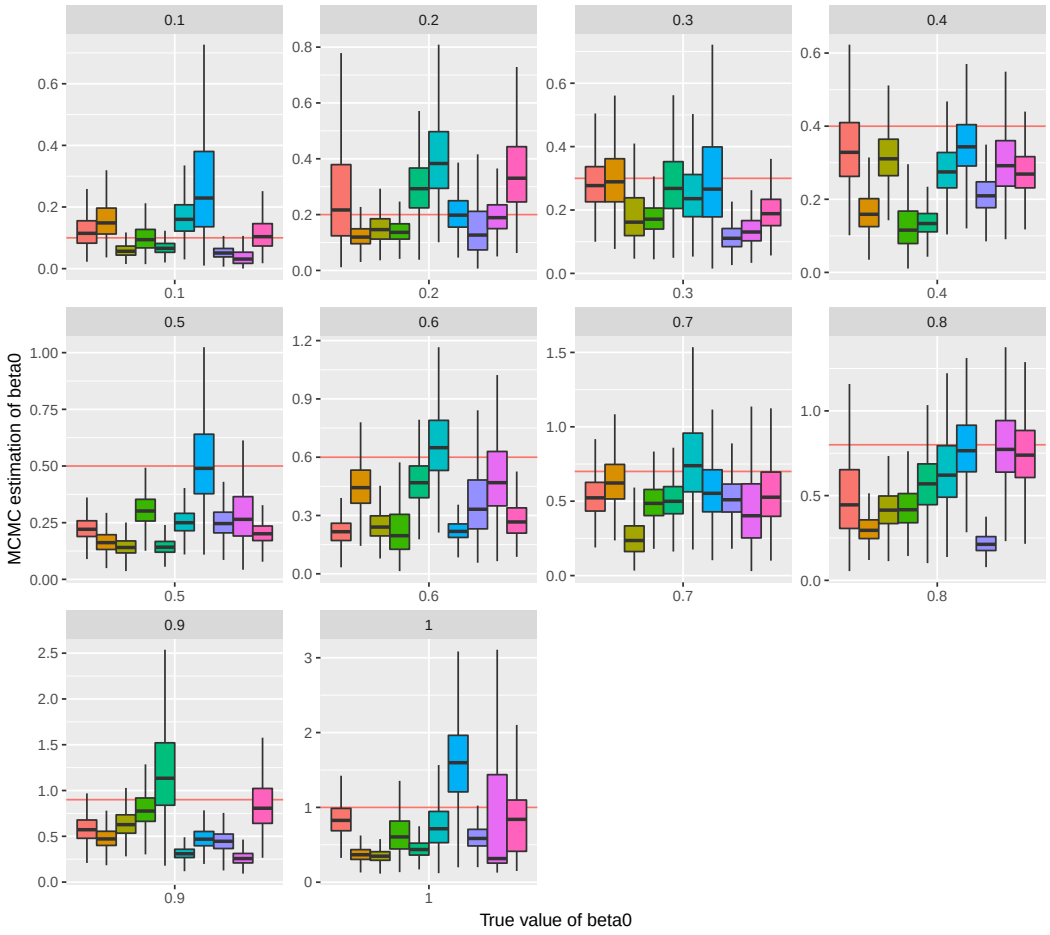


Figure F.12: The posterior estimates of parameter β_0 from fitting the Chain Poisson model to 100 simulated datasets from the same model with varied input value for β_0, with 10 simulations for each value.

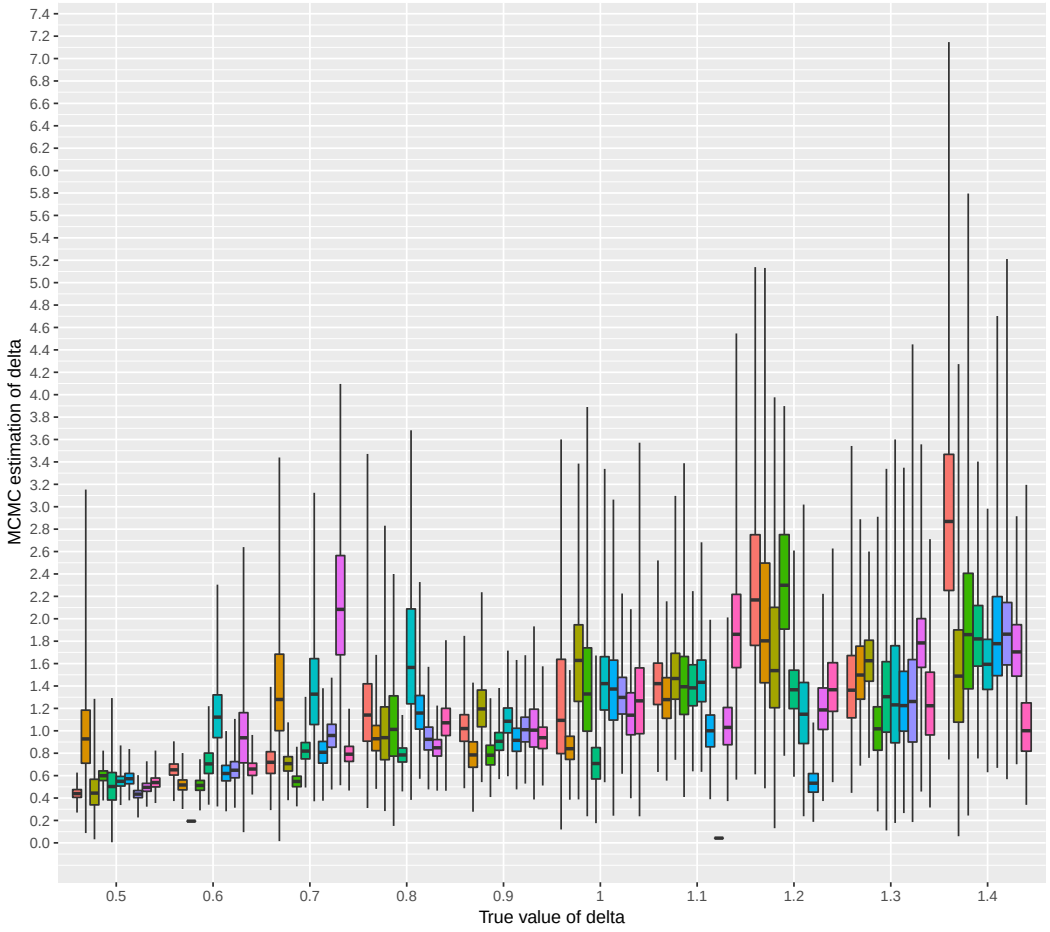


Figure F.13: The posterior estimates of parameter δ from fitting the Chain Poisson model to 100 simulated datasets from the same model with varied input value for δ, with 10 simulations for each value.

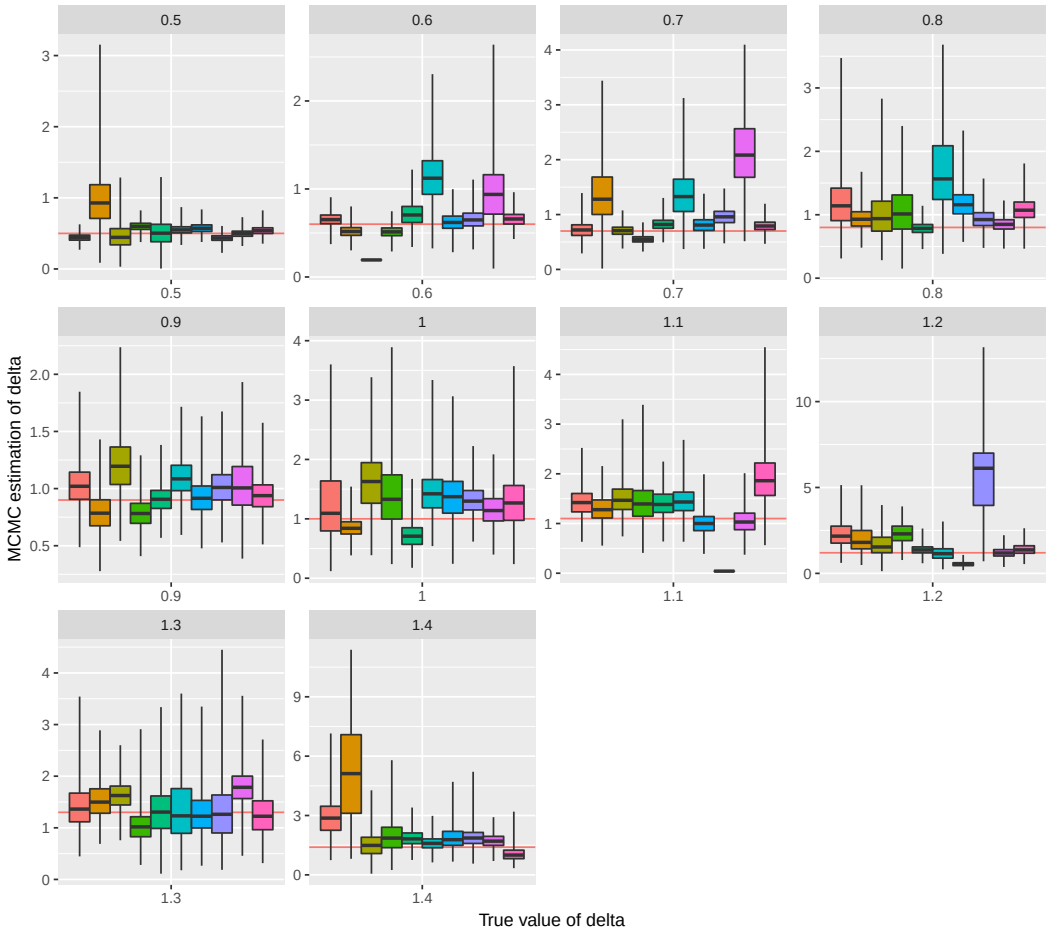


Figure F.14: The posterior estimates of parameter δ from fitting the Chain Poisson model to 100 simulated datasets from the same model with varied input value for δ, with 10 simulations for each value.

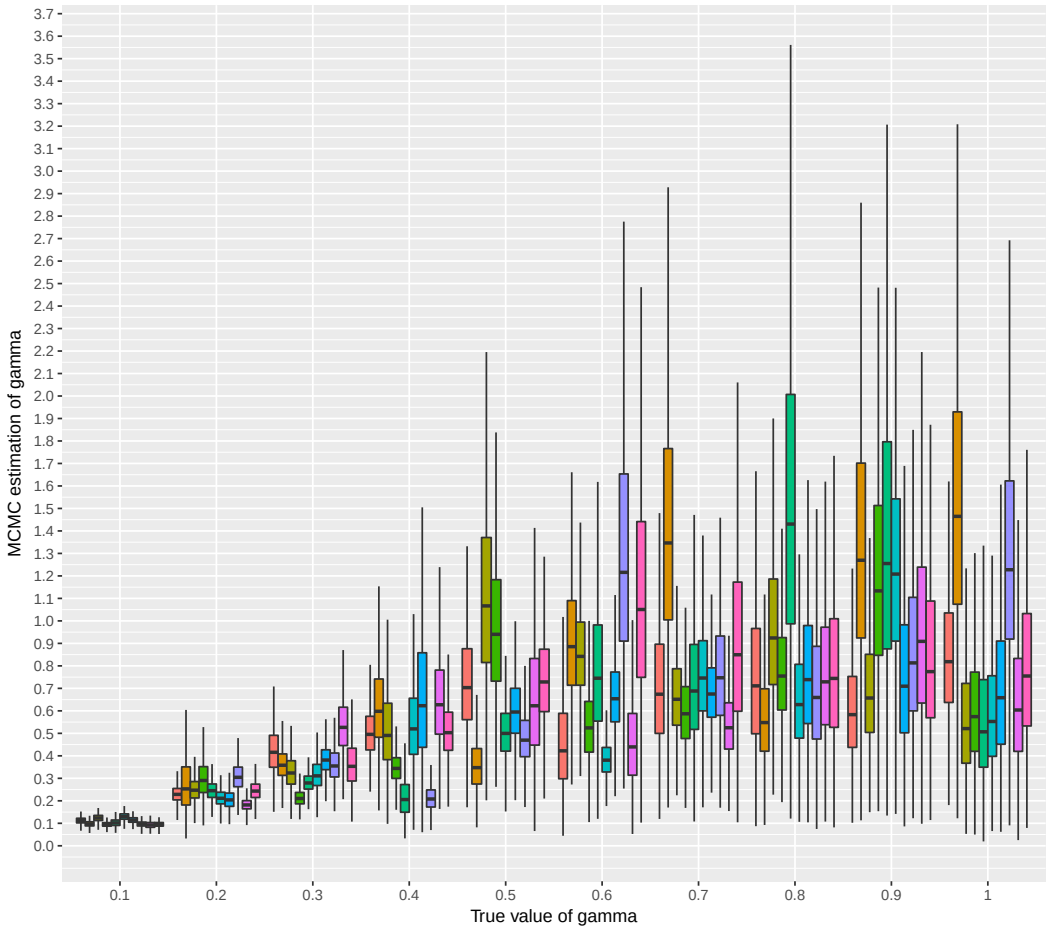


Figure F.15: The posterior estimates of parameter γ from fitting the Chain Poisson model to 100 simulated datasets from the same model with varied input value for γ, with 10 simulations for each value.

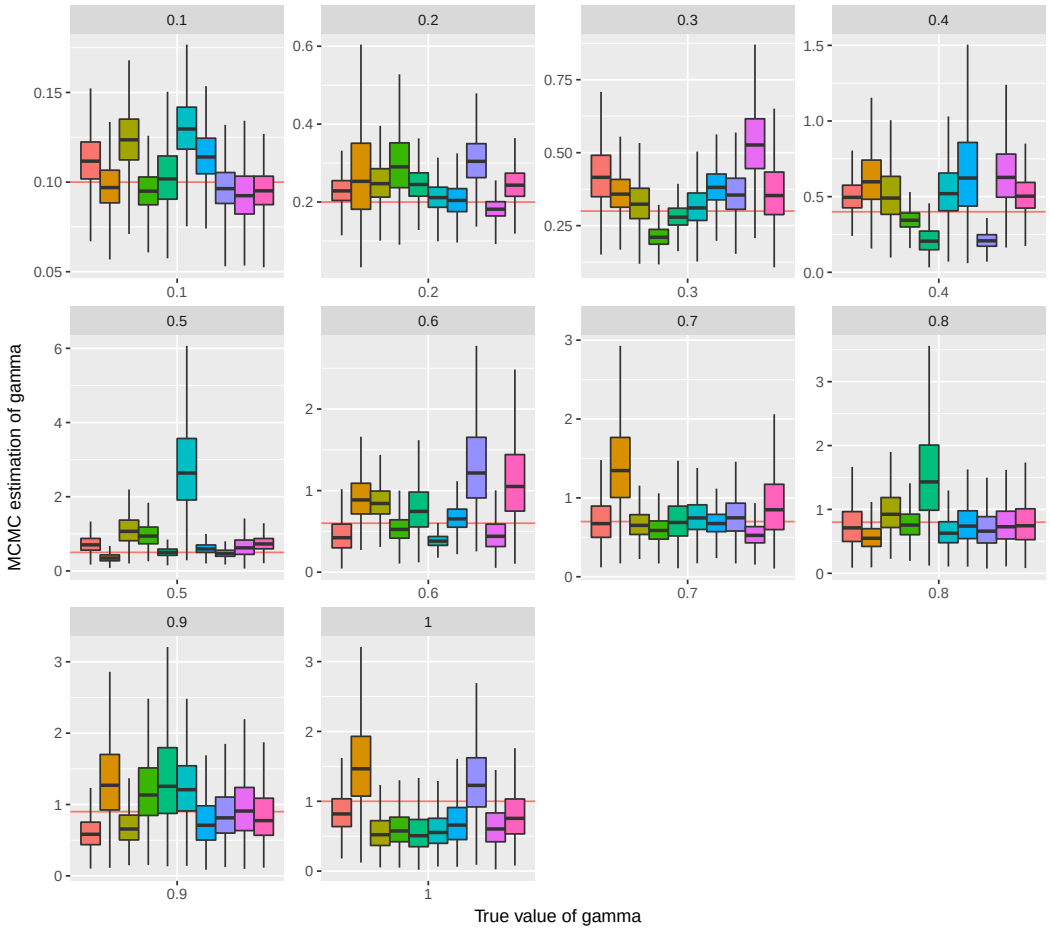


Figure F.16: The posterior estimates of parameter γ from fitting the Chain Poisson model to 100 simulated datasets from the same model with varied input value for γ, with 10 simulations for each value.

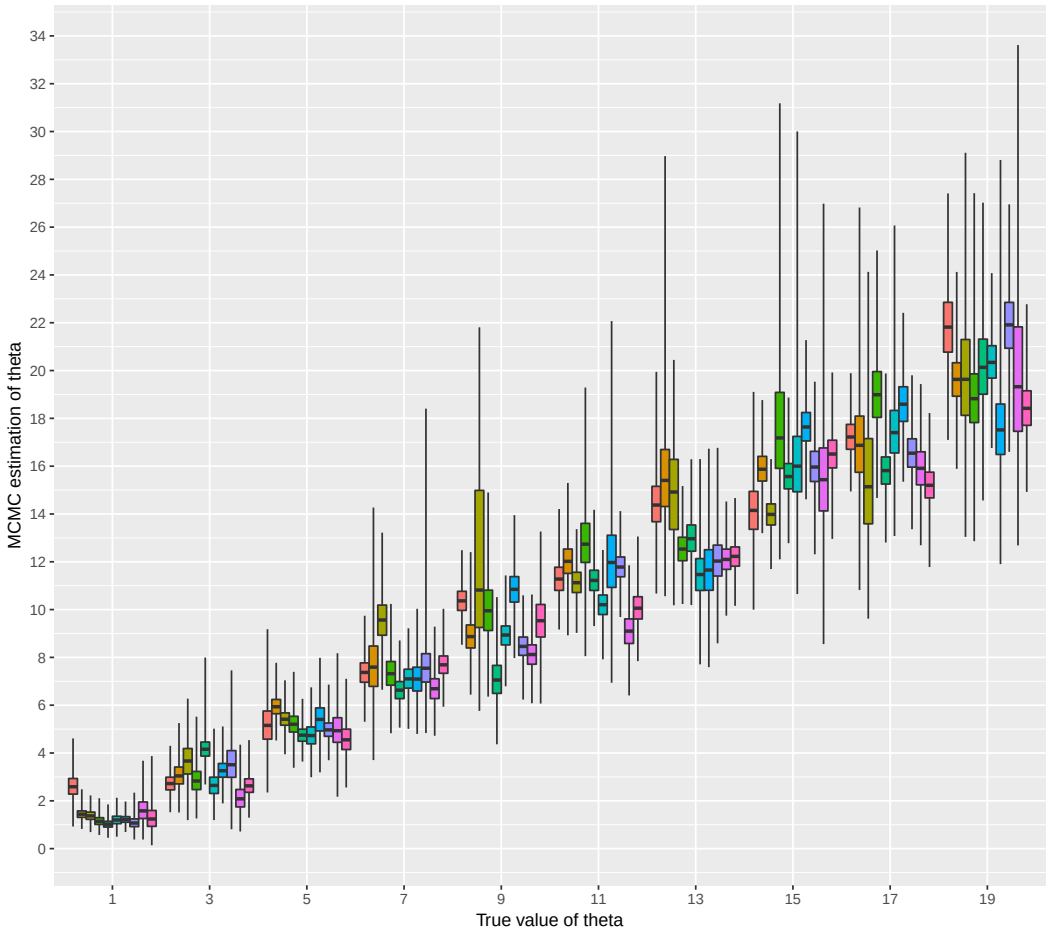


Figure F.17: The posterior estimates of parameter θ from fitting the Chain Poisson model to 100 simulated datasets from the same model with varied input value for θ, with 10 simulations for each value.

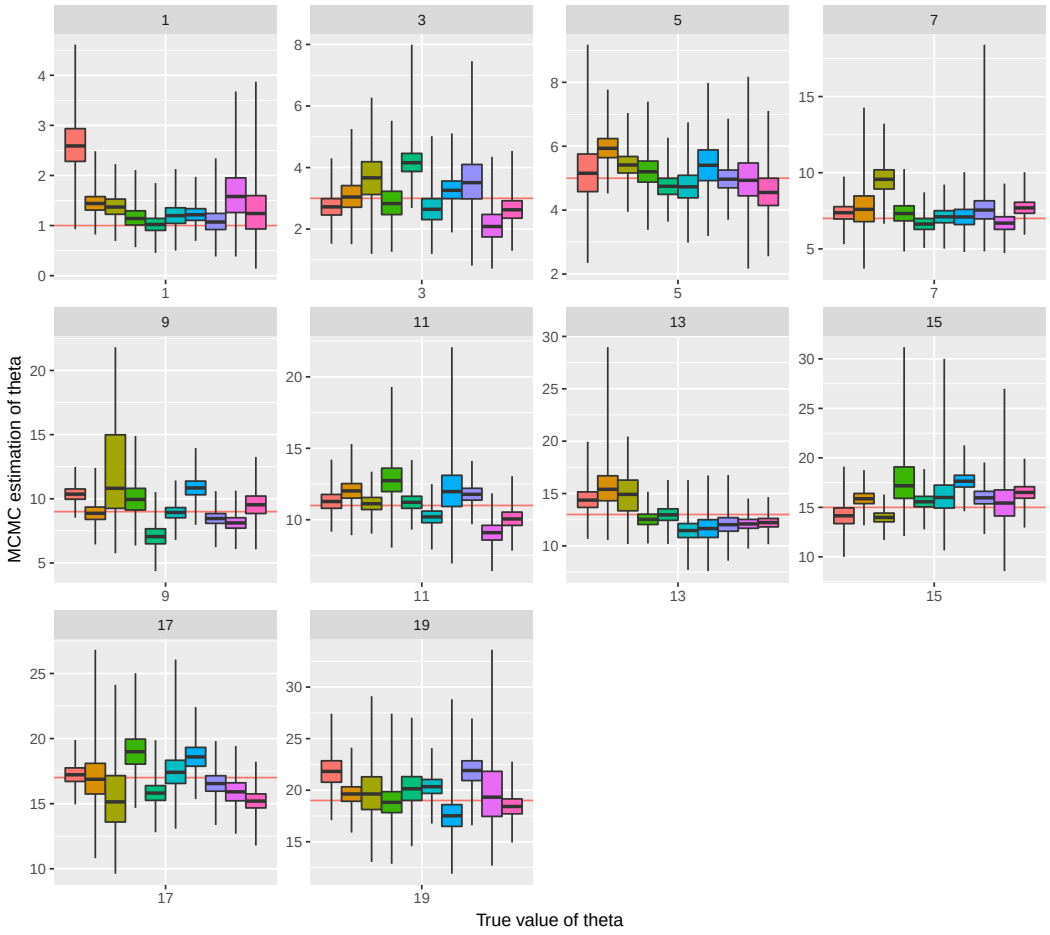


Figure F.18: The posterior estimates of parameter θ from fitting the Chain Poisson model to 100 simulated datasets from the same model with varied input value for θ, with 10 simulations for each value.

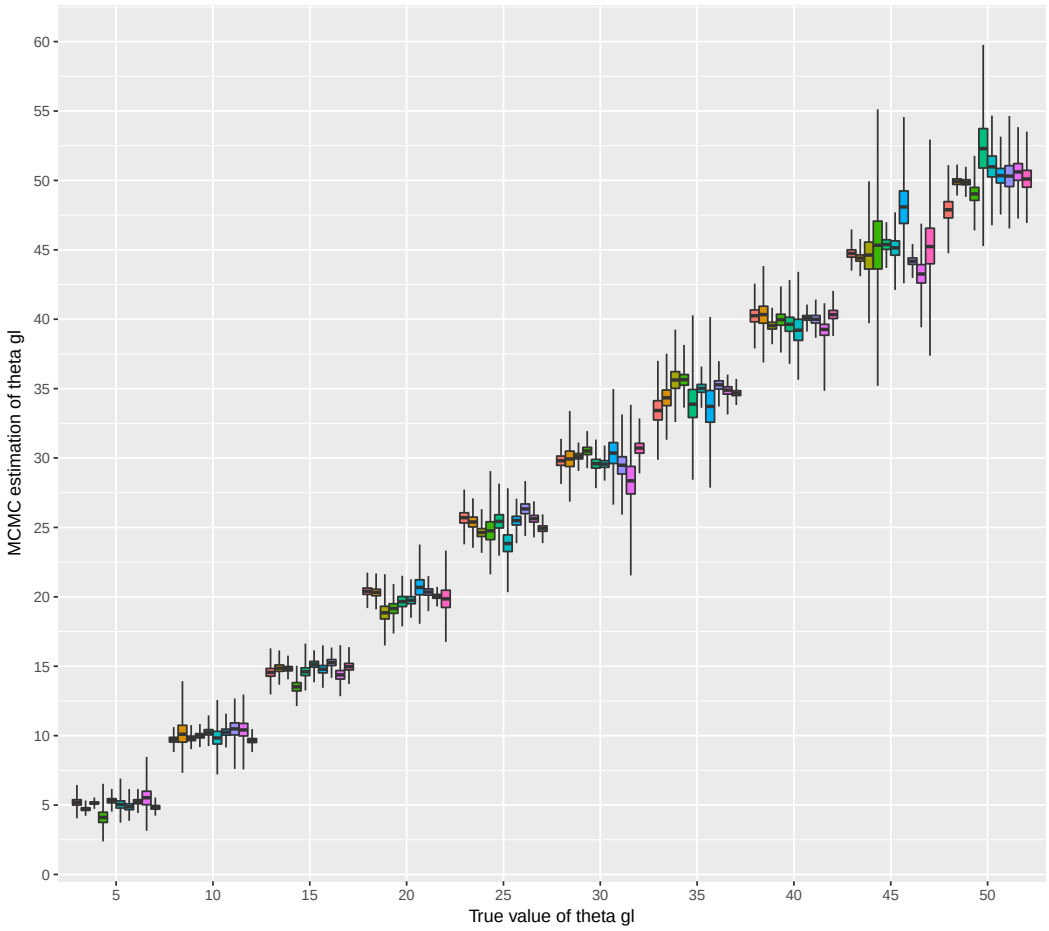


Figure F.19: The posterior estimates of parameter θ_{gl} from fitting the Chain Poisson model to 100 simulated datasets from the same model with varied input value for θ_{gl}, with 10 simulations for each value.

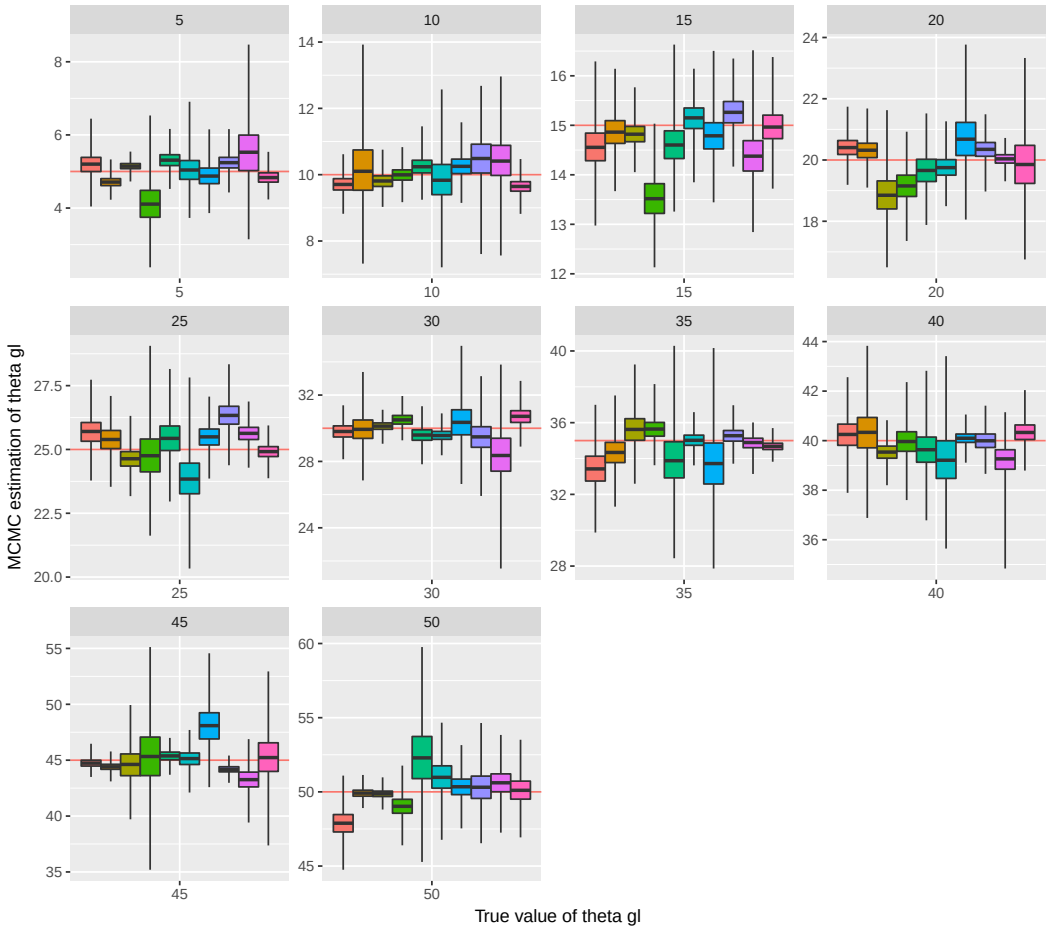


Figure F.20: The posterior estimates of parameter θ_{gl} from fitting the Chain Poisson model to 100 simulated datasets from the same model with varied input value for θ_{gl}, with 10 simulations for each value.

F.3 Time Dependent Distances model

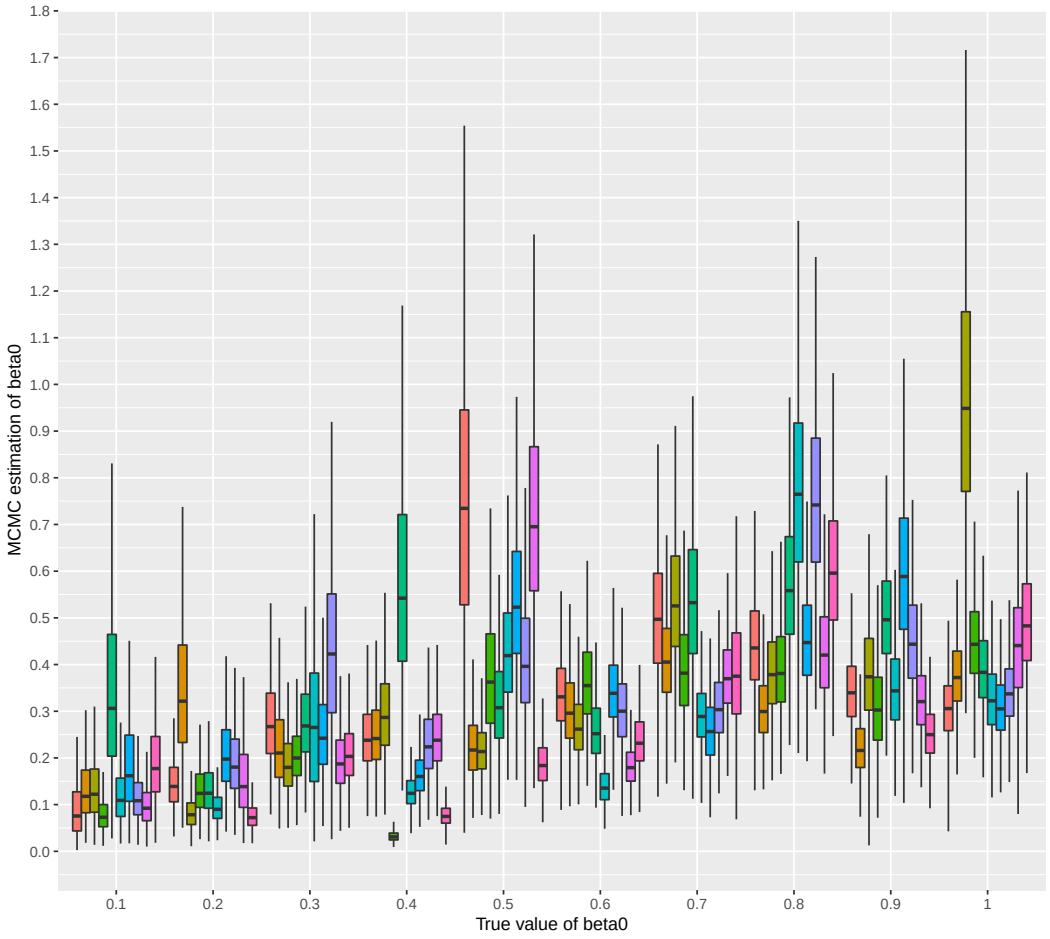


Figure F.21: The posterior estimates of parameter β_0 from fitting the Time Dependent Distances model to 100 simulated datasets from the same model with varied input value for β_0, with 10 simulations for each value.

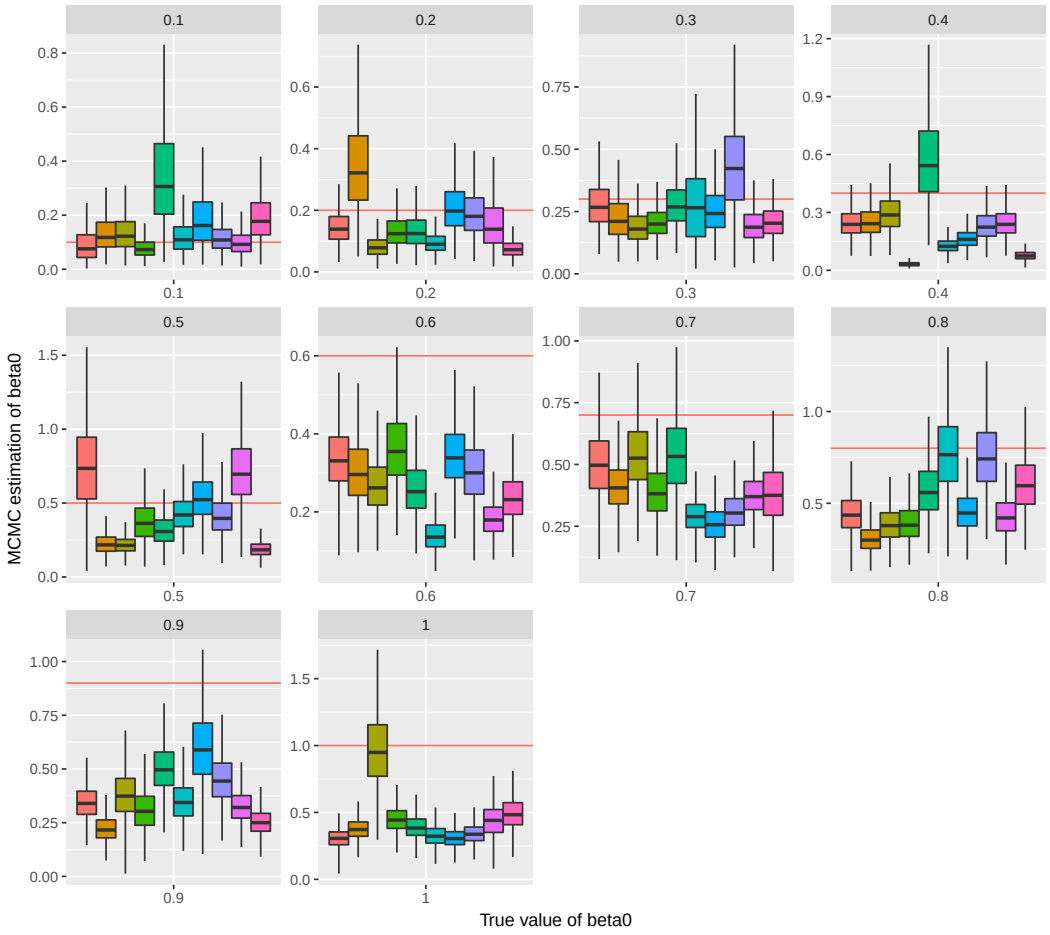


Figure F.22: The posterior estimates of parameter β_0 from fitting the Time Dependent Distances model to 100 simulated datasets from the same model with varied input value for β_0, with 10 simulations for each value.

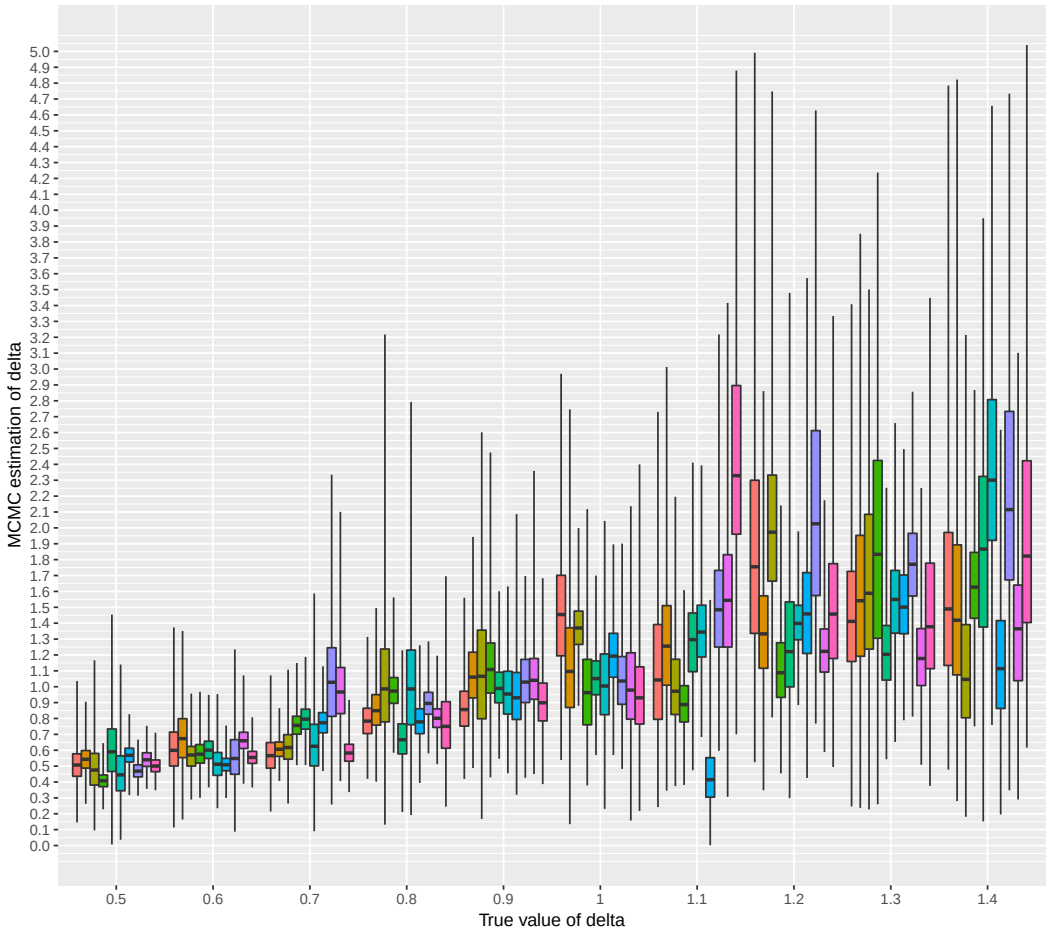


Figure F.23: The posterior estimates of parameter δ from fitting the Time Dependent Distances model to 100 simulated datasets from the same model with varied input value for δ, with 10 simulations for each value.

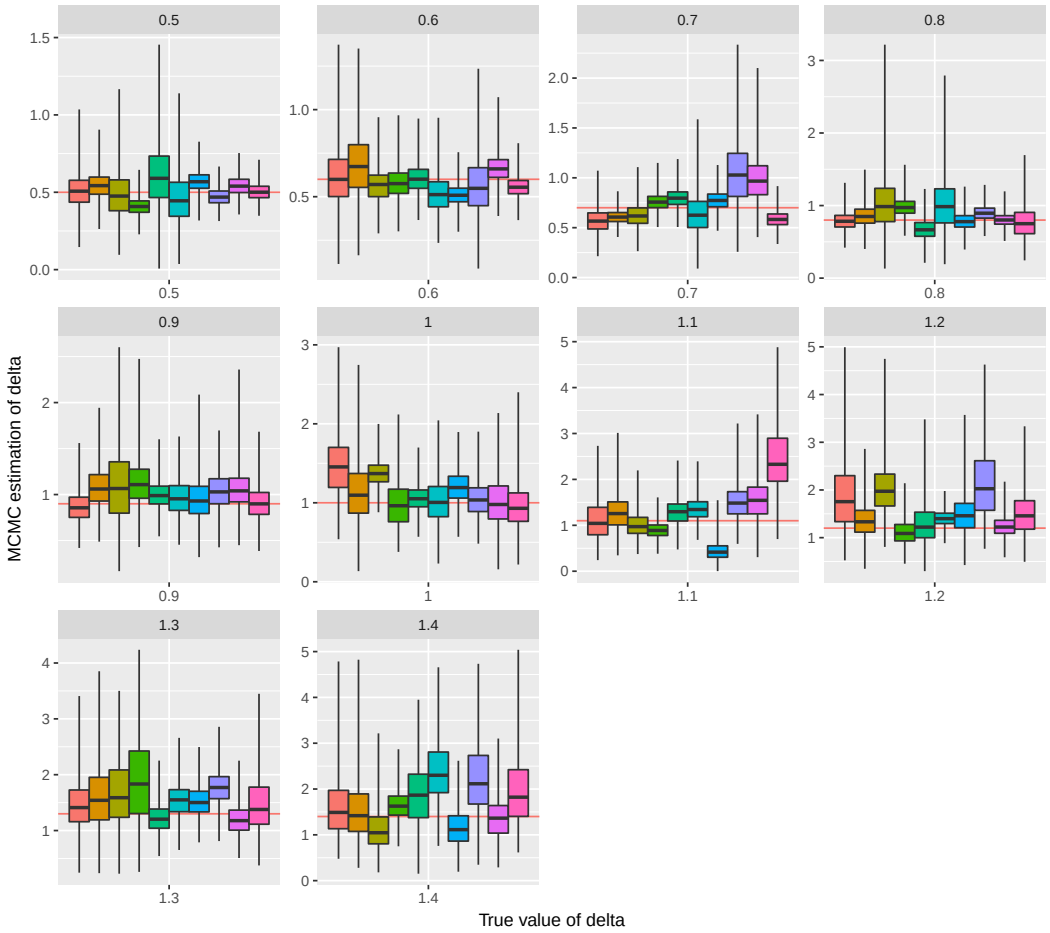


Figure F.24: The posterior estimates of parameter δ from fitting the Time Dependent Distances model to 100 simulated datasets from the same model with varied input value for δ, with 10 simulations for each value.

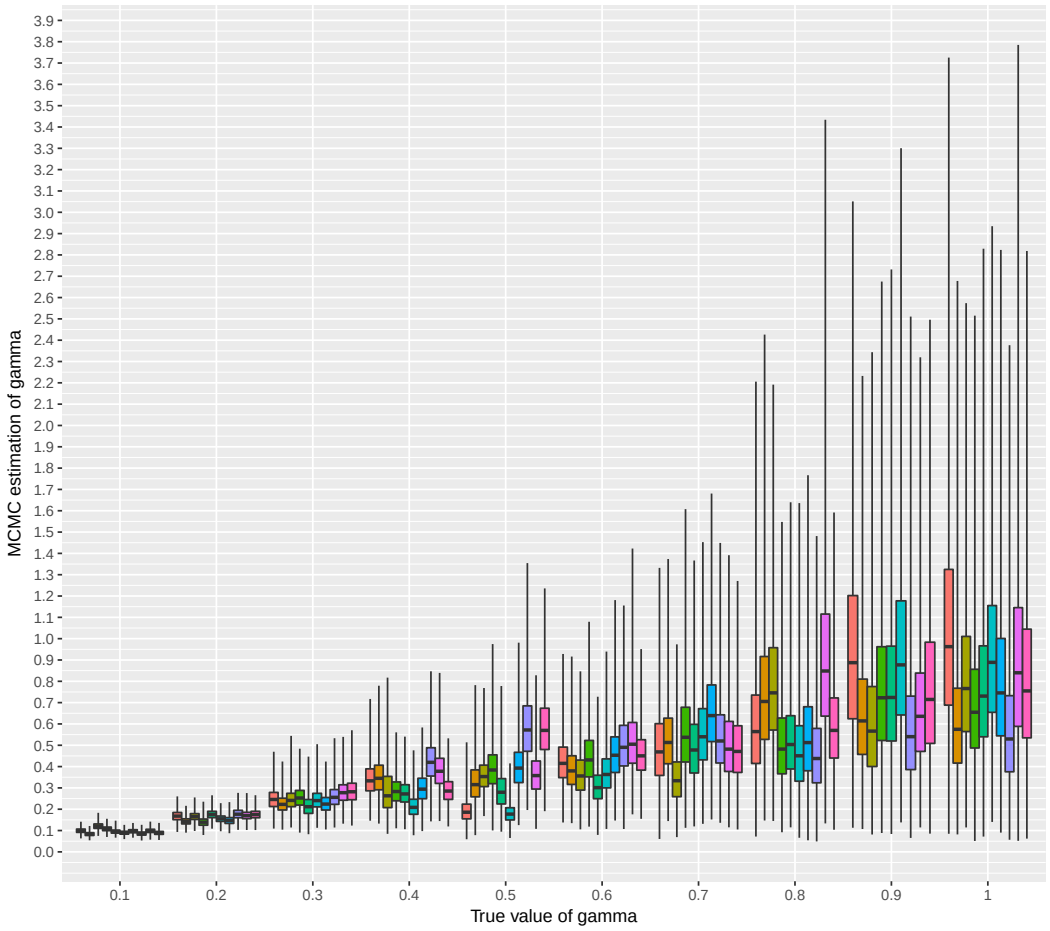


Figure F.25: The posterior estimates of parameter γ from fitting the Time Dependent Distances model to 100 simulated datasets from the same model with varied input value for γ, with 10 simulations for each value.

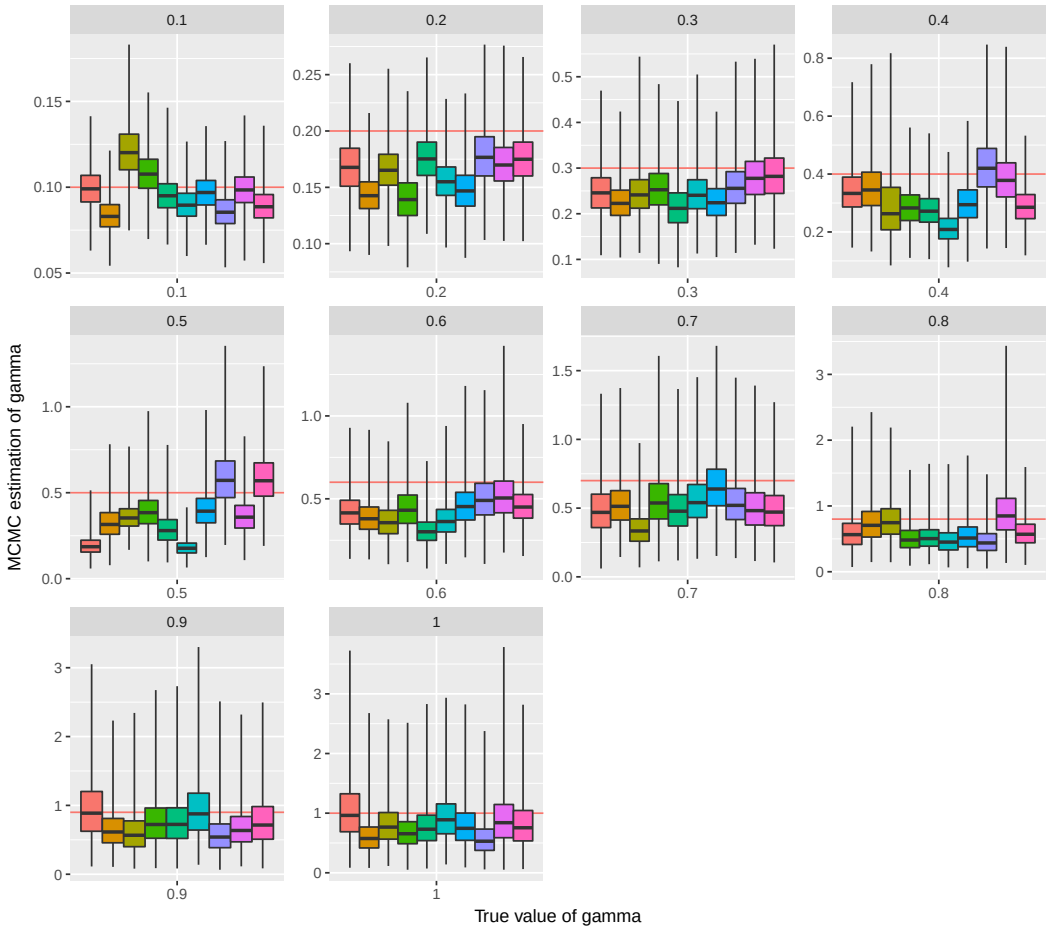


Figure F.26: The posterior estimates of parameter γ from fitting the Time Dependent Distances model to 100 simulated datasets from the same model with varied input value for γ, with 10 simulations for each value.

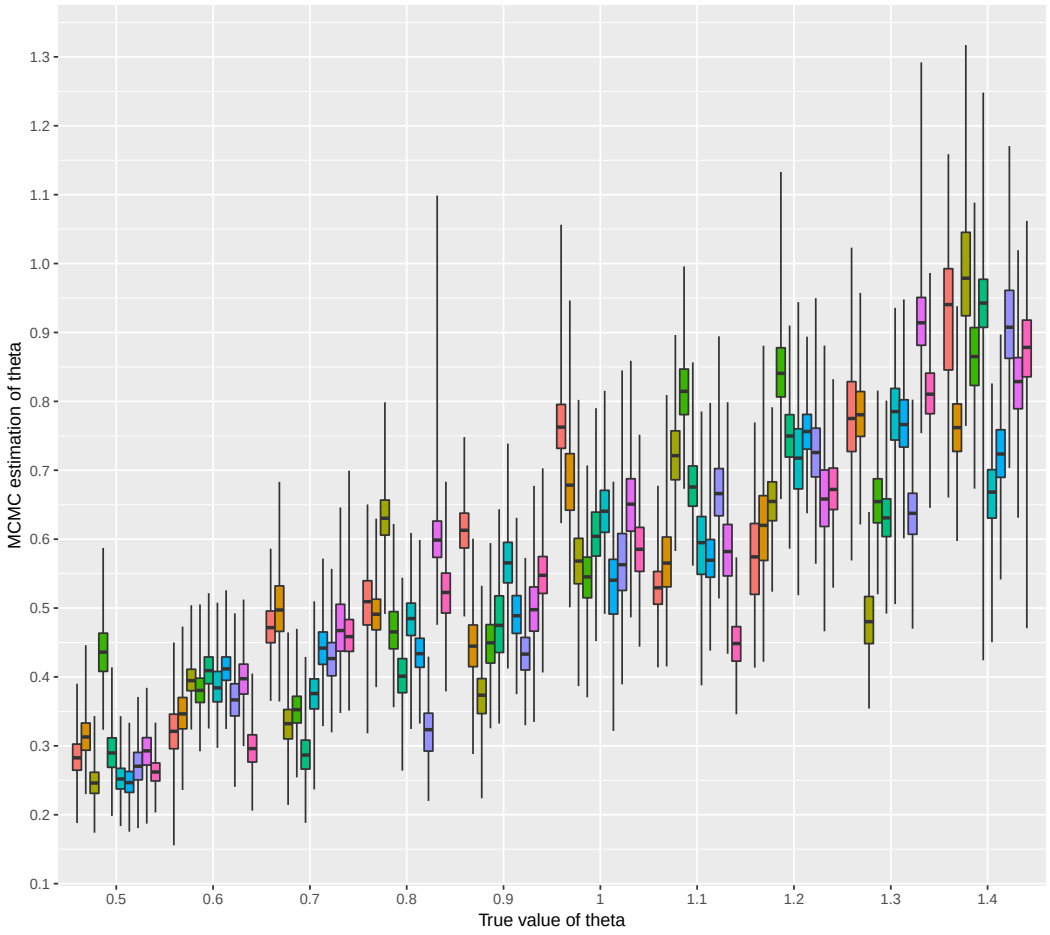


Figure F.27: The posterior estimates of parameter θ from fitting the Time Dependent Distances model to 100 simulated datasets from the same model with varied input value for θ, with 10 simulations for each value.

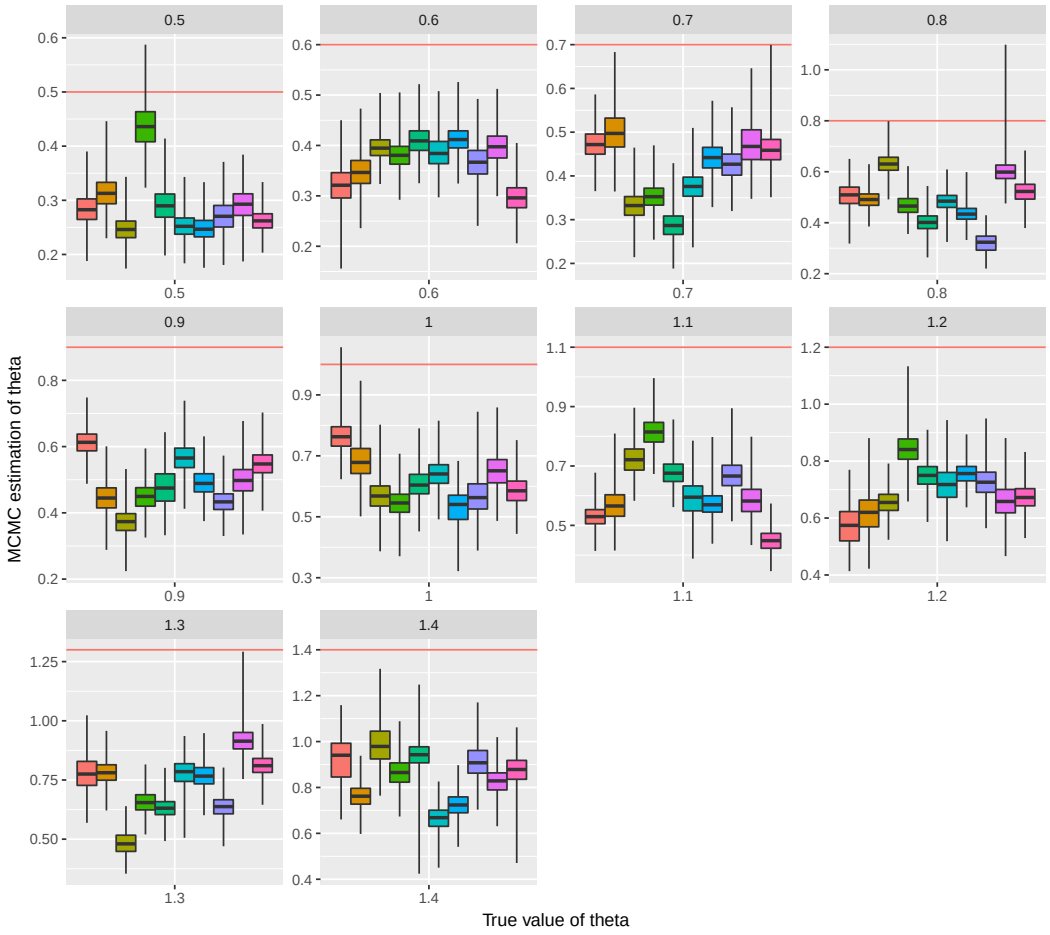


Figure F.28: The posterior estimates of parameter θ from fitting the Time Dependent Distances model to 100 simulated datasets from the same model with varied input value for θ, with 10 simulations for each value.

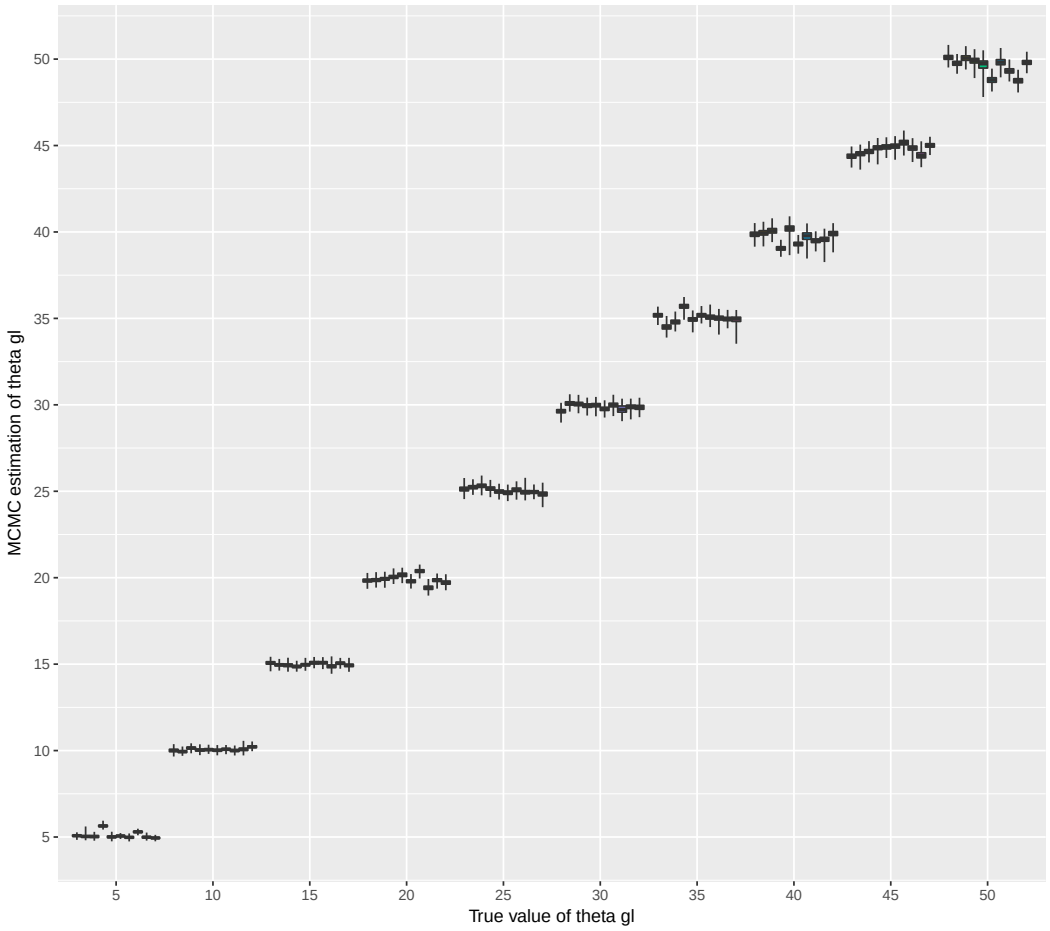


Figure F.29: The posterior estimates of parameter θ_{gl} from fitting the Time Dependent Distances model to 100 simulated datasets from the same model with varied input value for θ_{gl}, with 10 simulations for each value.

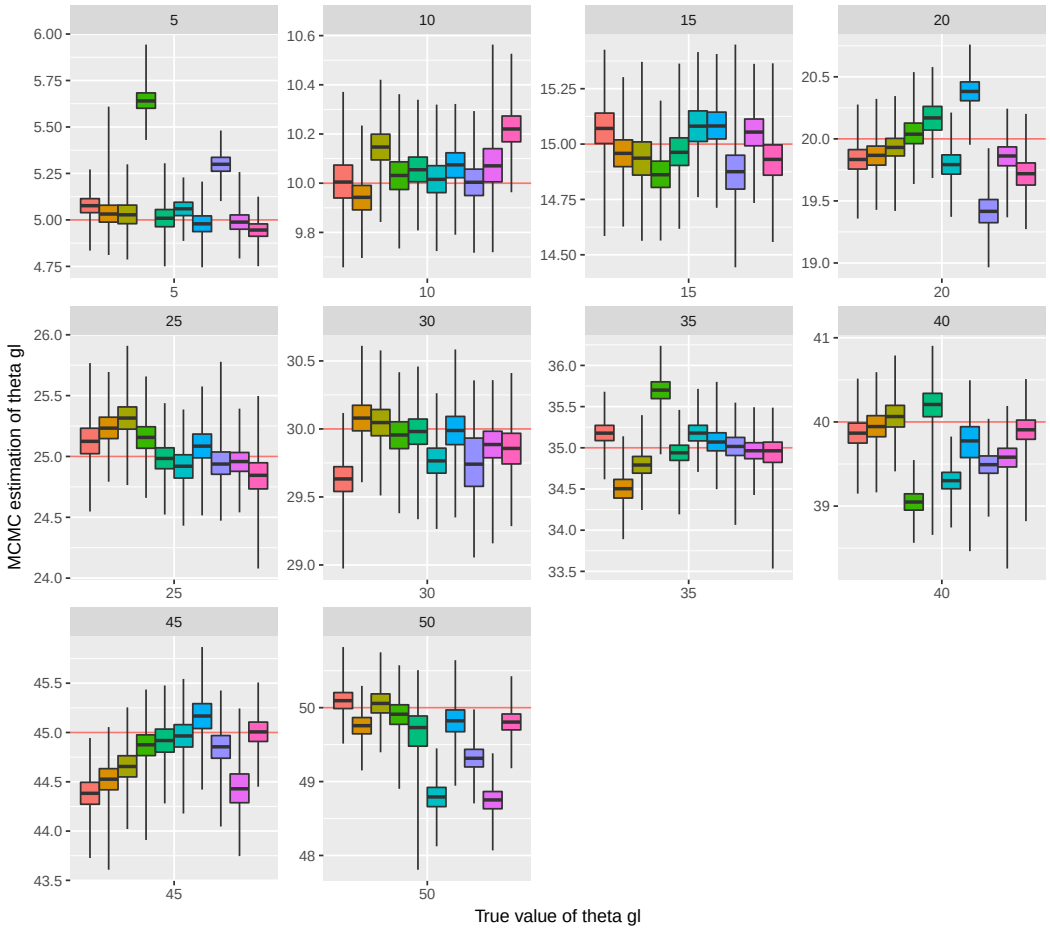


Figure F.30: The posterior estimates of parameter θ_{gl} from fitting the Time Dependent Distances model to 100 simulated datasets from the same model with varied input value for θ_{gl}, with 10 simulations for each value.

Appendix G

Graphs for network reconstruction for simulations for avian influenza

G.0.1 Chain Error model

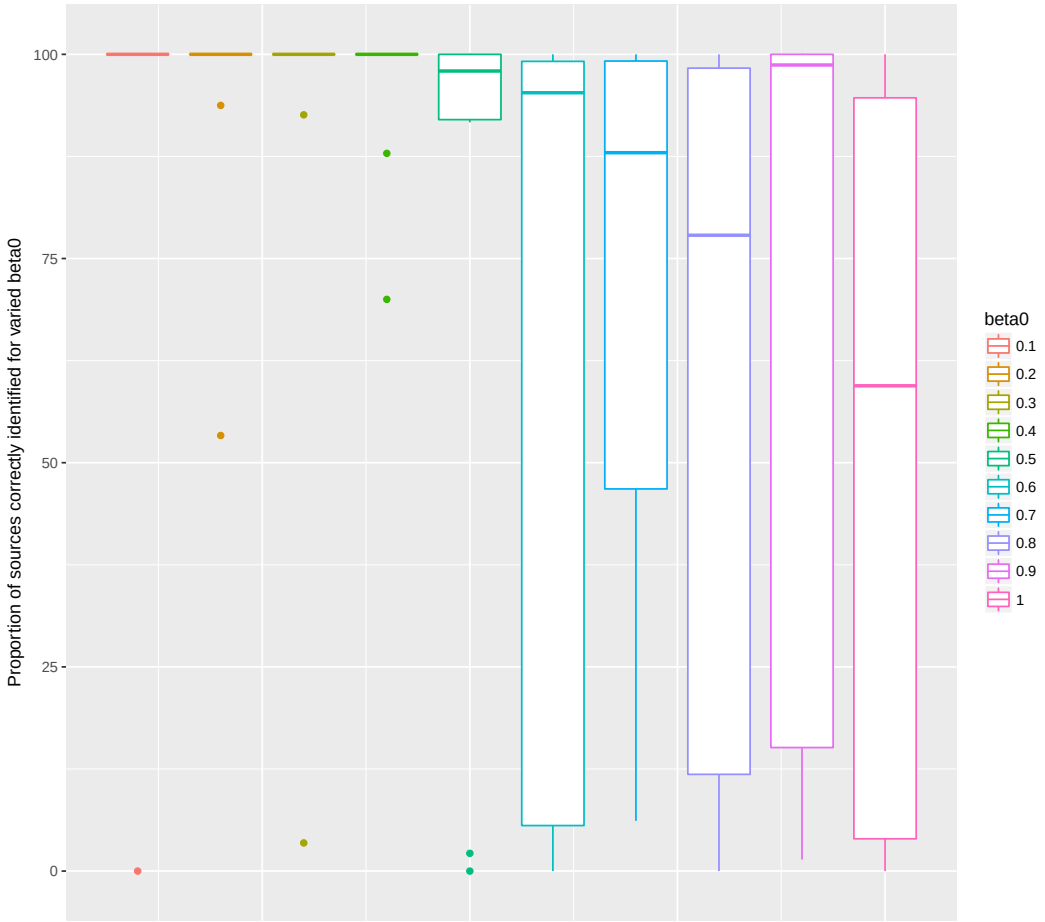


Figure G.1: Boxplots to show the proportion of infection sources for farms recovered correctly for simulations with varied values for parameter β_0 for the Chain Error model.

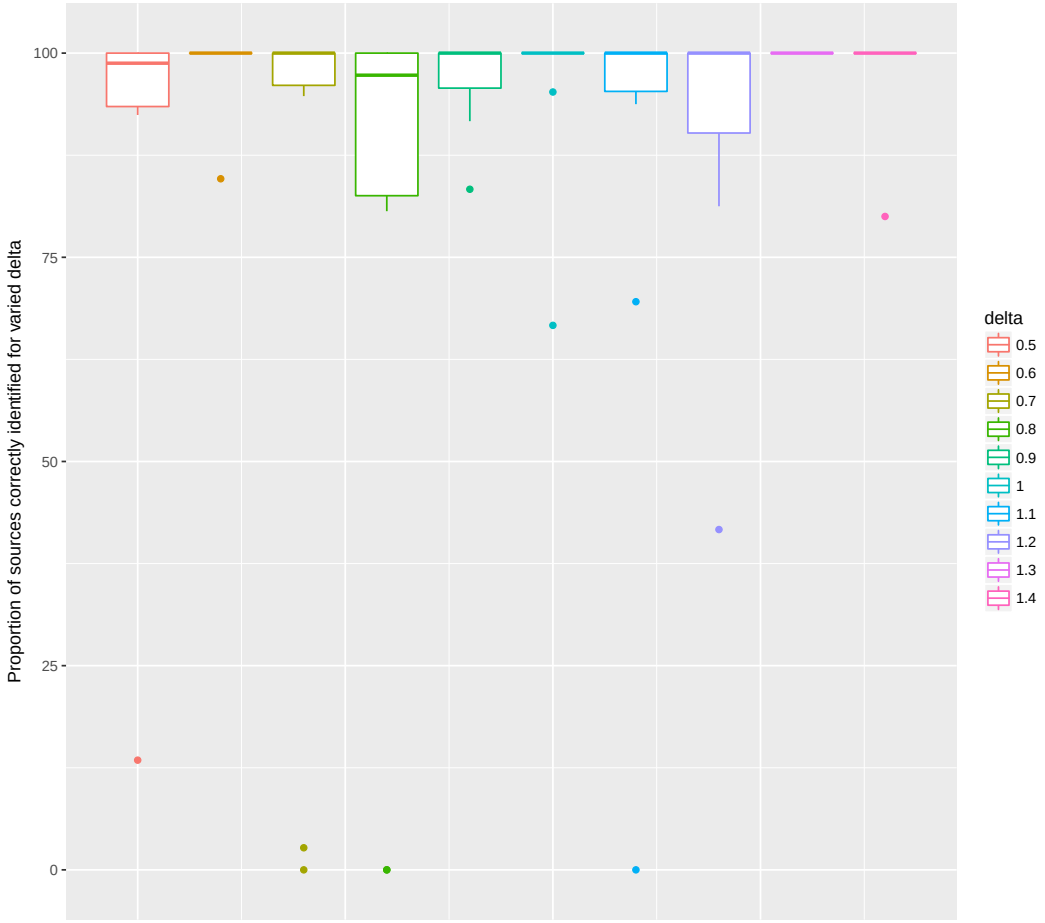


Figure G.2: Boxplots to show the proportion of infection sources for farms recovered correctly for simulations with varied values for parameter δ for the Chain Error model.

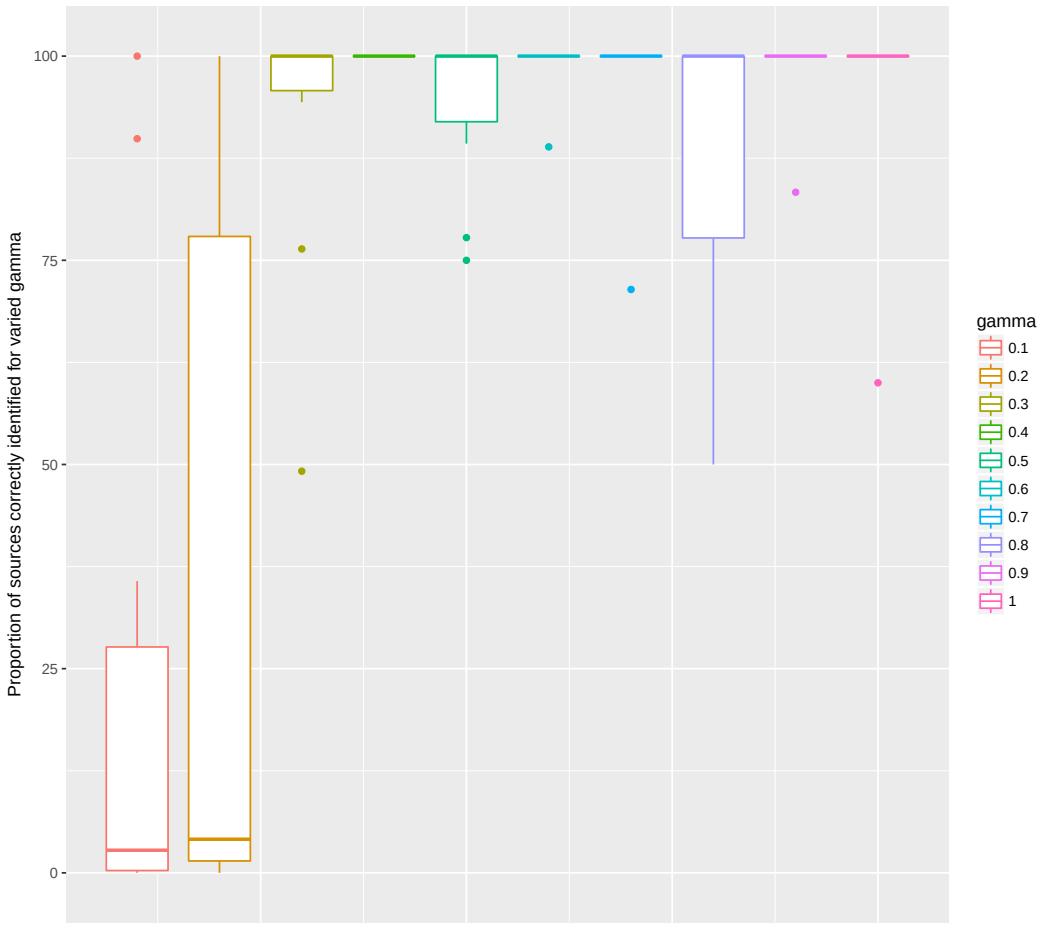


Figure G.3: Boxplots to show the proportion of infection sources for farms recovered correctly for simulations with varied values for parameter γ for the Chain Error model.

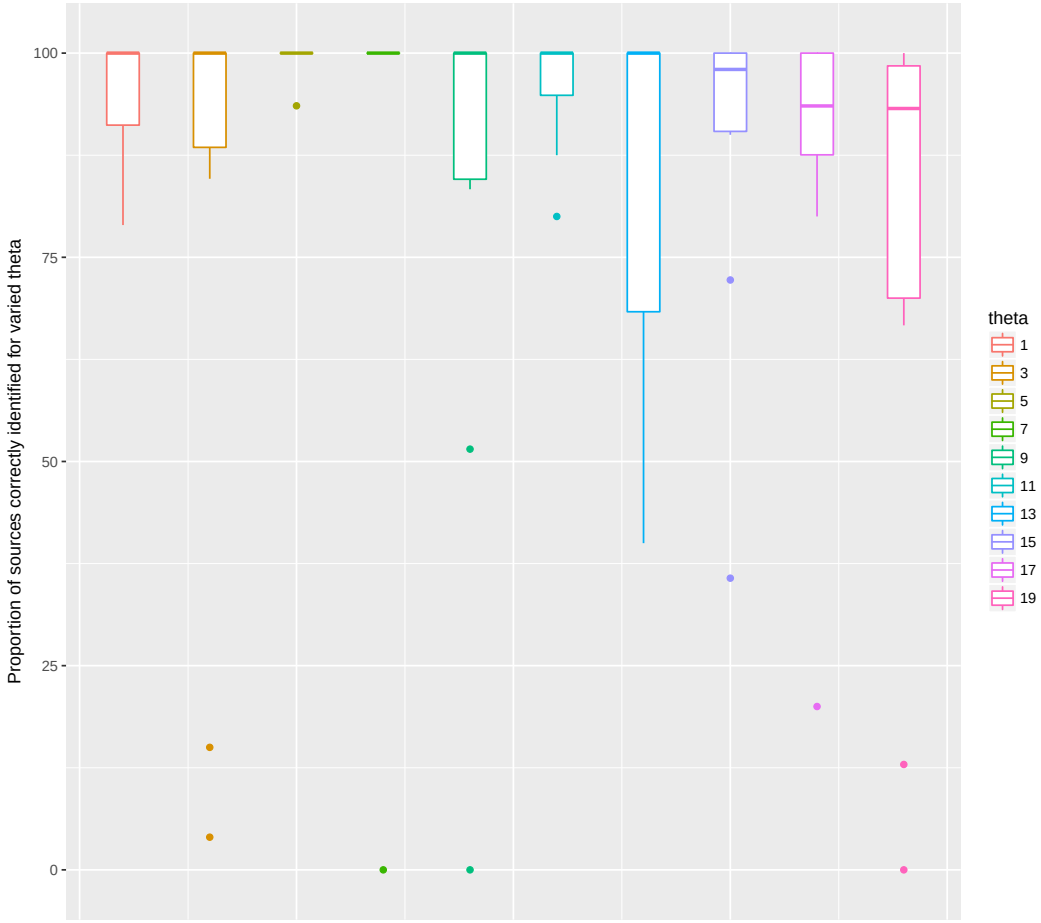


Figure G.4: Boxplots to show the proportion of infection sources for farms recovered correctly for simulations with varied values for parameter θ for the Chain Error model.

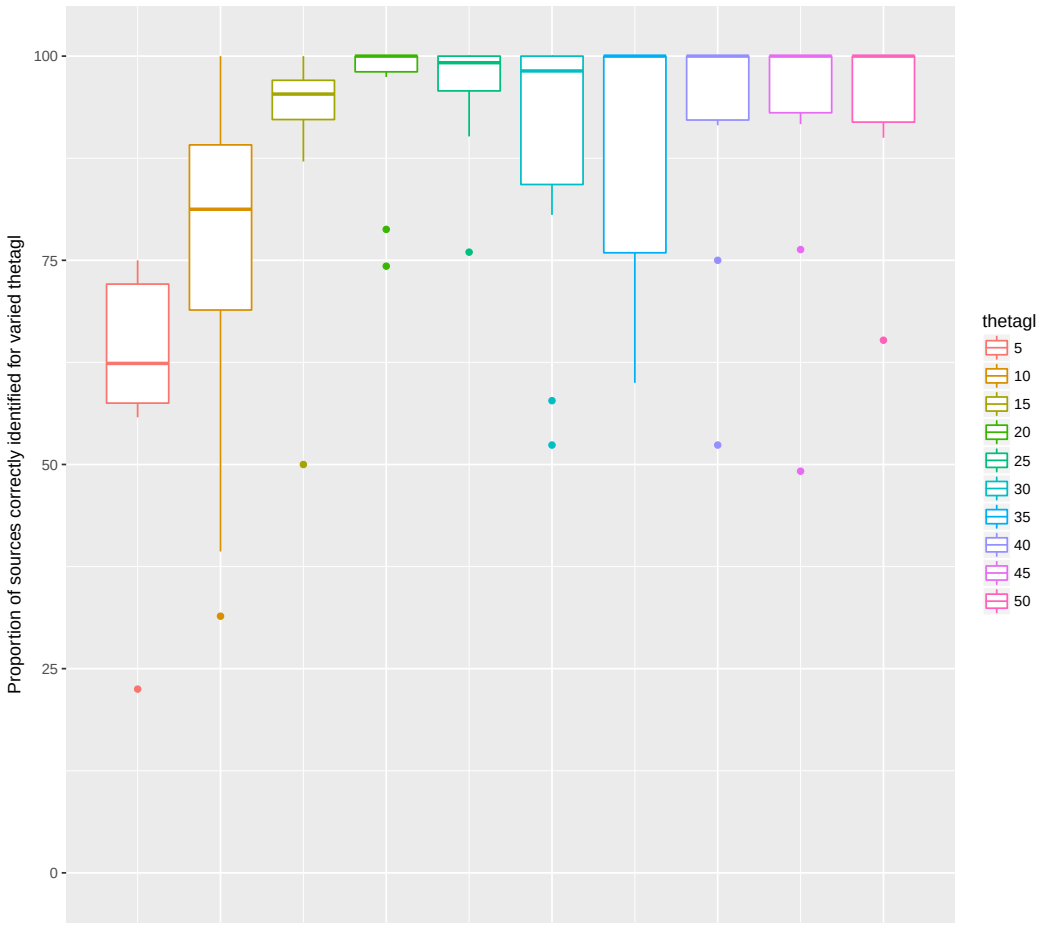


Figure G.5: Boxplots to show the proportion of infection sources for farms recovered correctly for simulations with varied values for parameter θ_{gl} for the Chain Error model.

G.0.2 Chain Poisson model

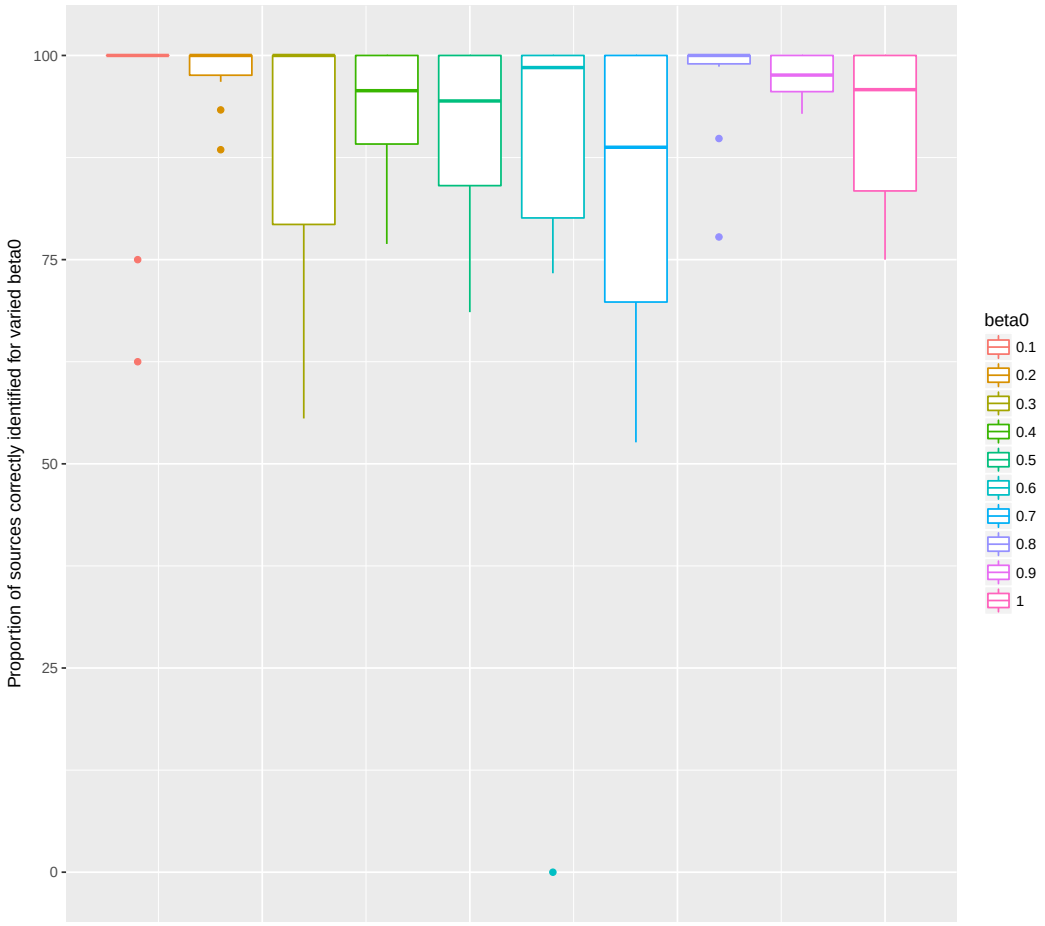


Figure G.6: Boxplots to show the proportion of infection sources for farms recovered correctly for simulations with varied values for parameter β_0 for the Chain Poisson model.

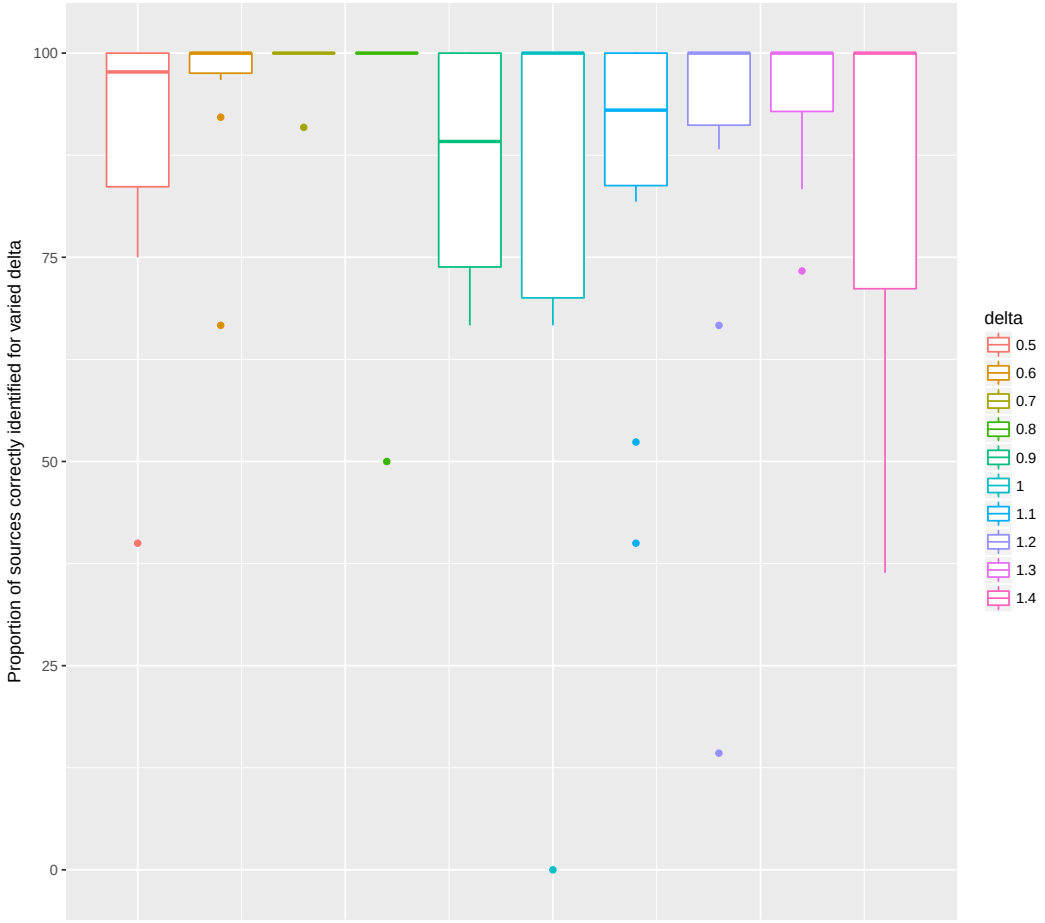


Figure G.7: Boxplots to show the proportion of infection sources for farms recovered correctly for simulations with varied values for parameter δ for the Chain Poisson model.

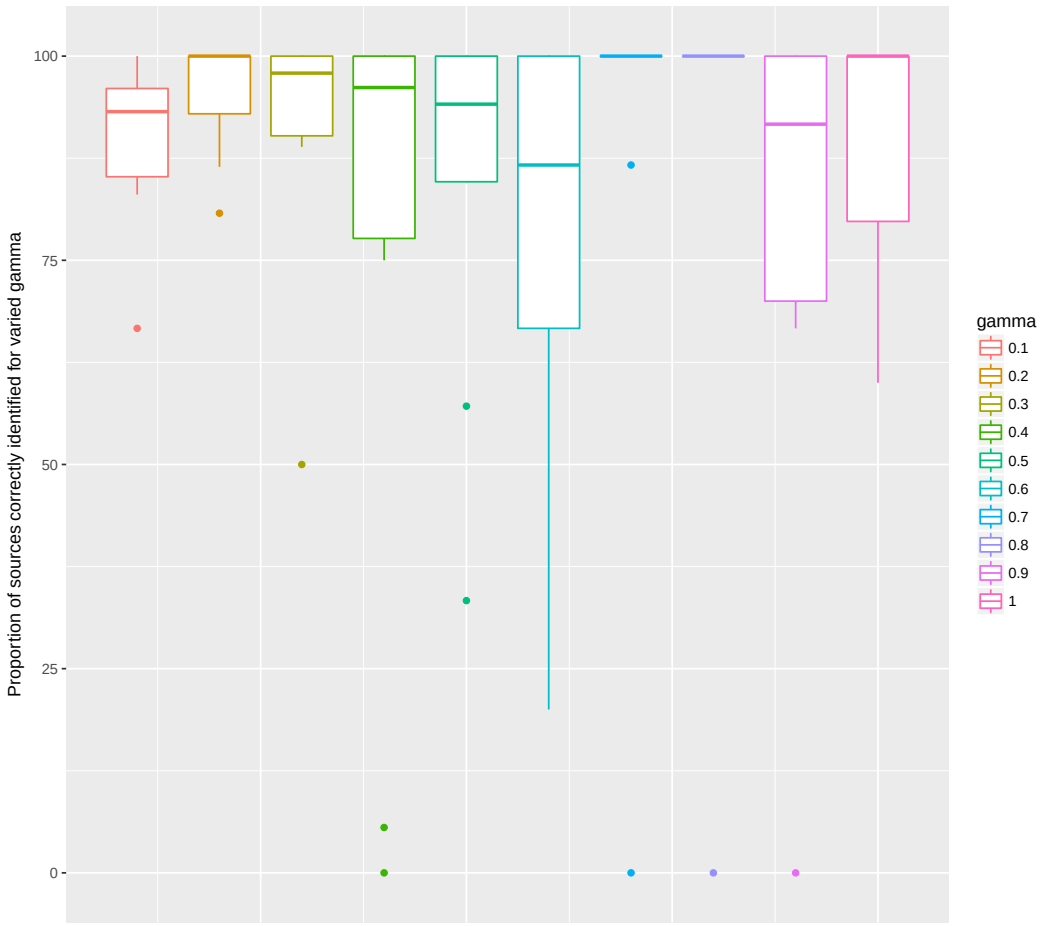


Figure G.8: Boxplots to show the proportion of infection sources for farms recovered correctly for simulations with varied values for parameter γ for the Chain Poisson model.

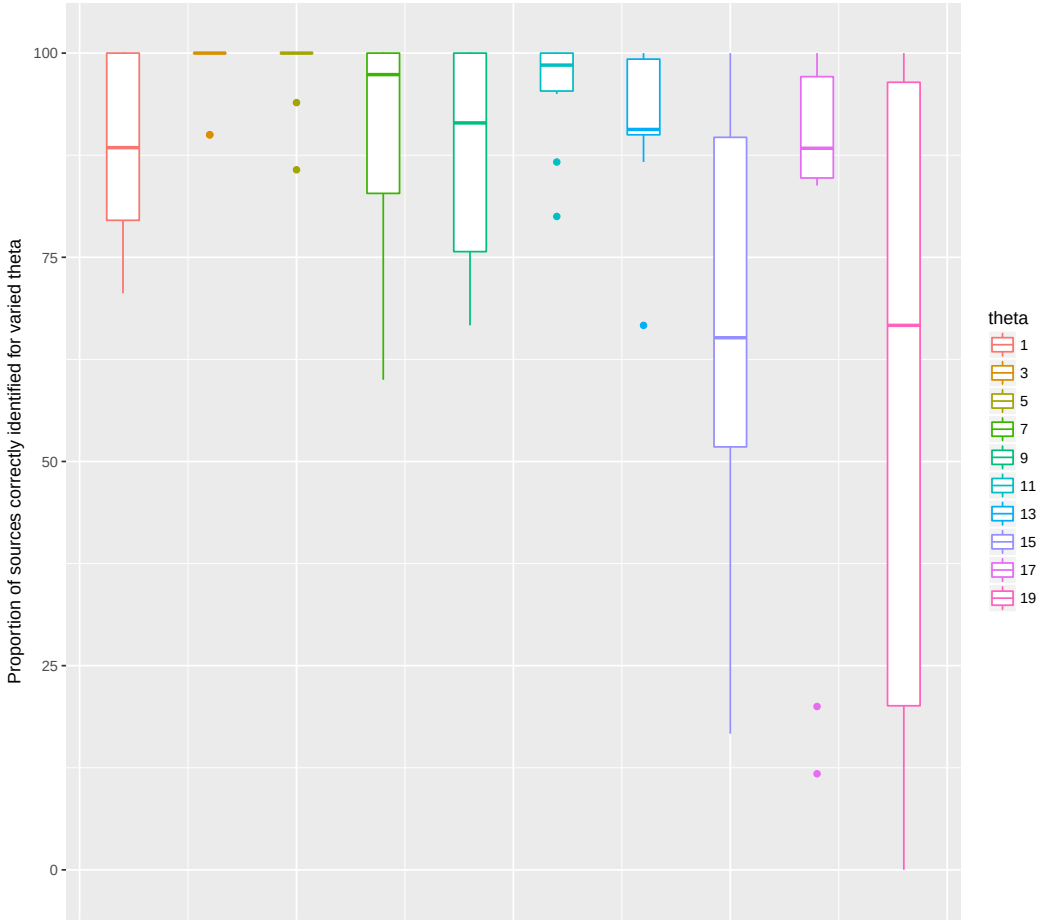


Figure G.9: Boxplots to show the proportion of infection sources for farms recovered correctly for simulations with varied values for parameter θ for the Chain Poisson model.

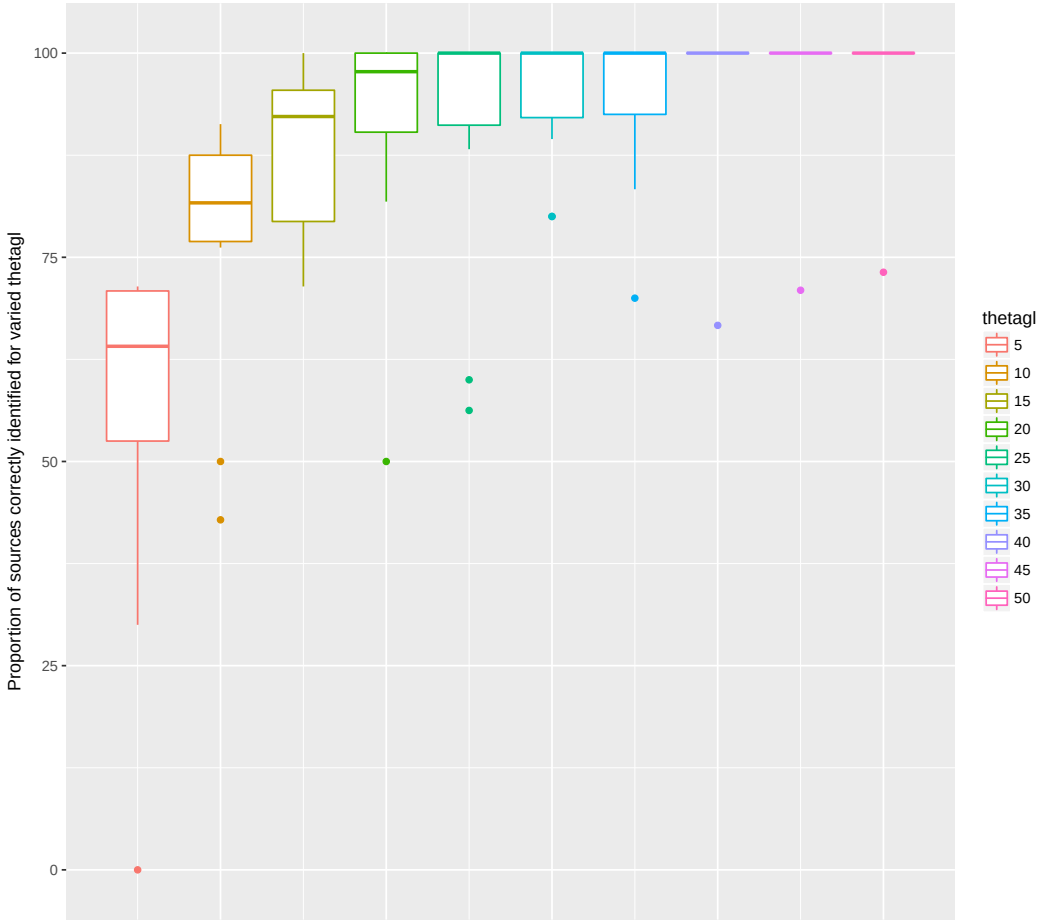


Figure G.10: Boxplots to show the proportion of infection sources for farms recovered correctly for simulations with varied values for parameter θ_{gl} for the Chain Poisson model.

G.0.3 Time Dependent Distances model

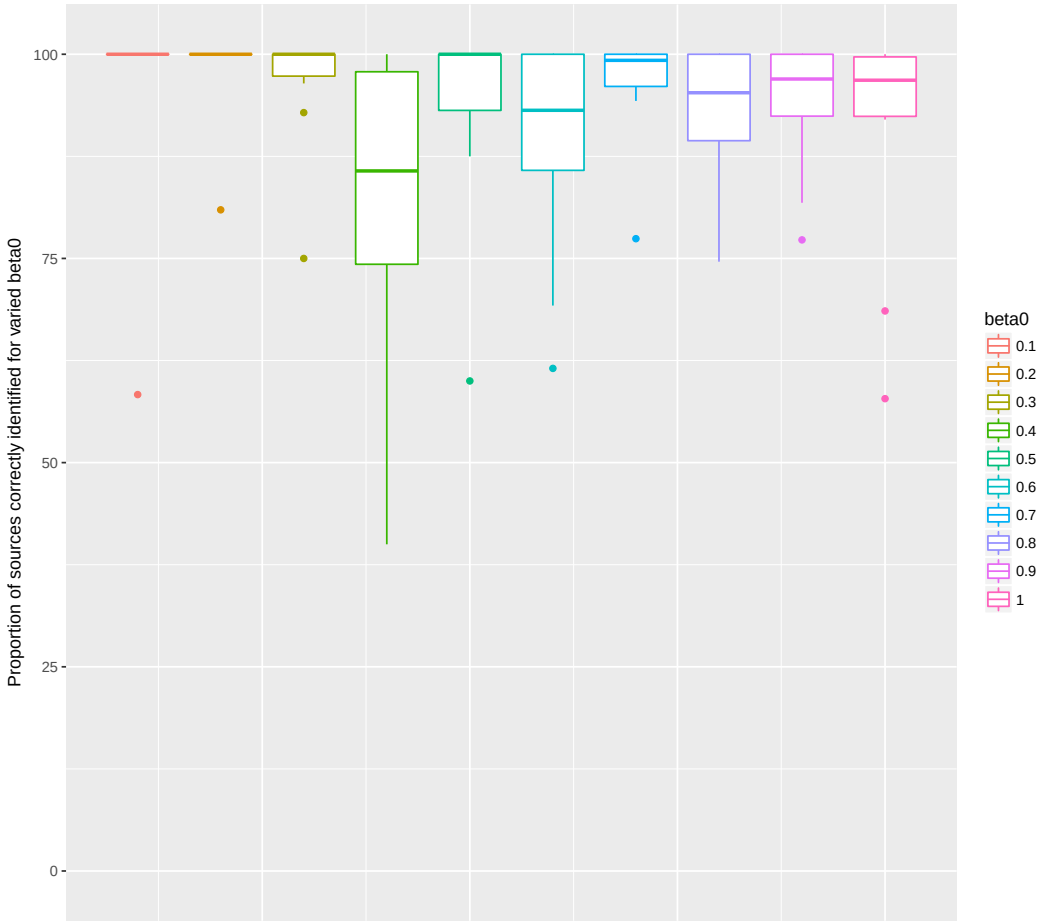


Figure G.11: Boxplots to show the proportion of infection sources for farms recovered correctly for simulations with varied values for parameter β_0 for the Time Dependent Distances model.

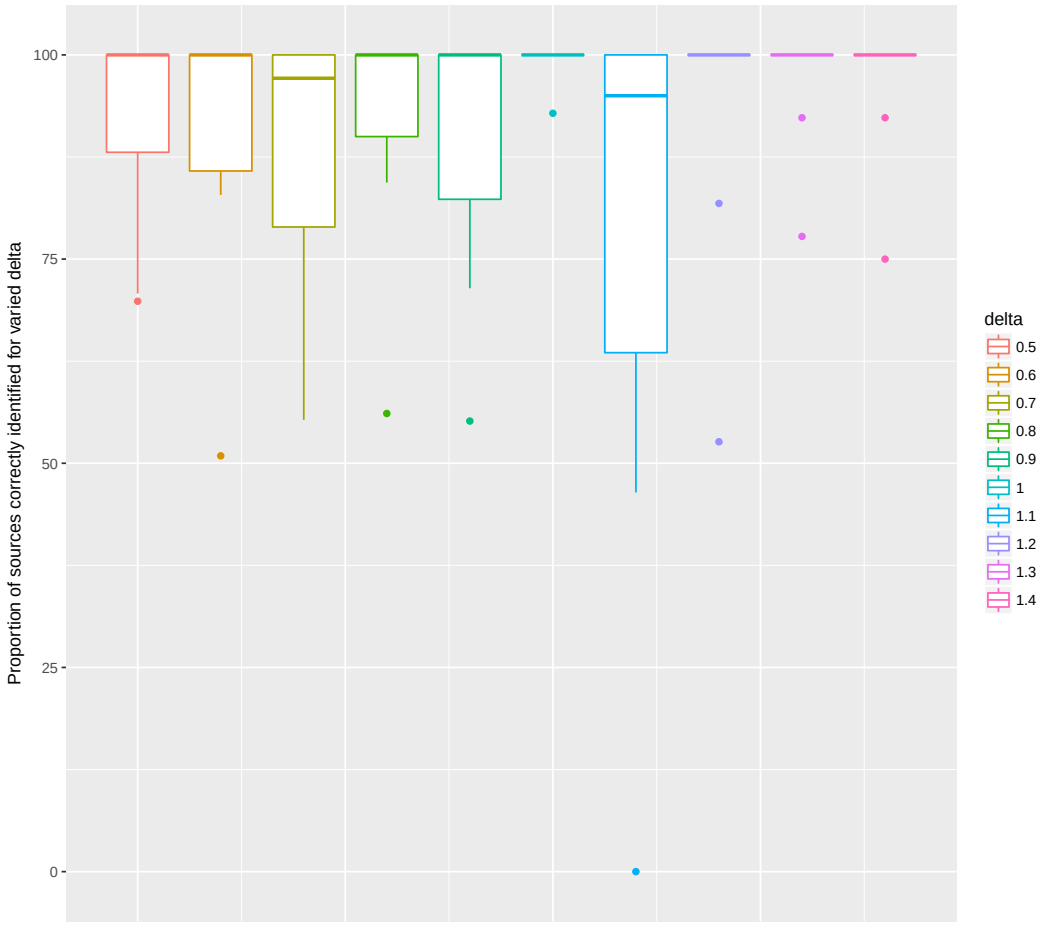


Figure G.12: Boxplots to show the proportion of infection sources for farms recovered correctly for simulations with varied values for parameter δ for the Time Dependent Distances model.

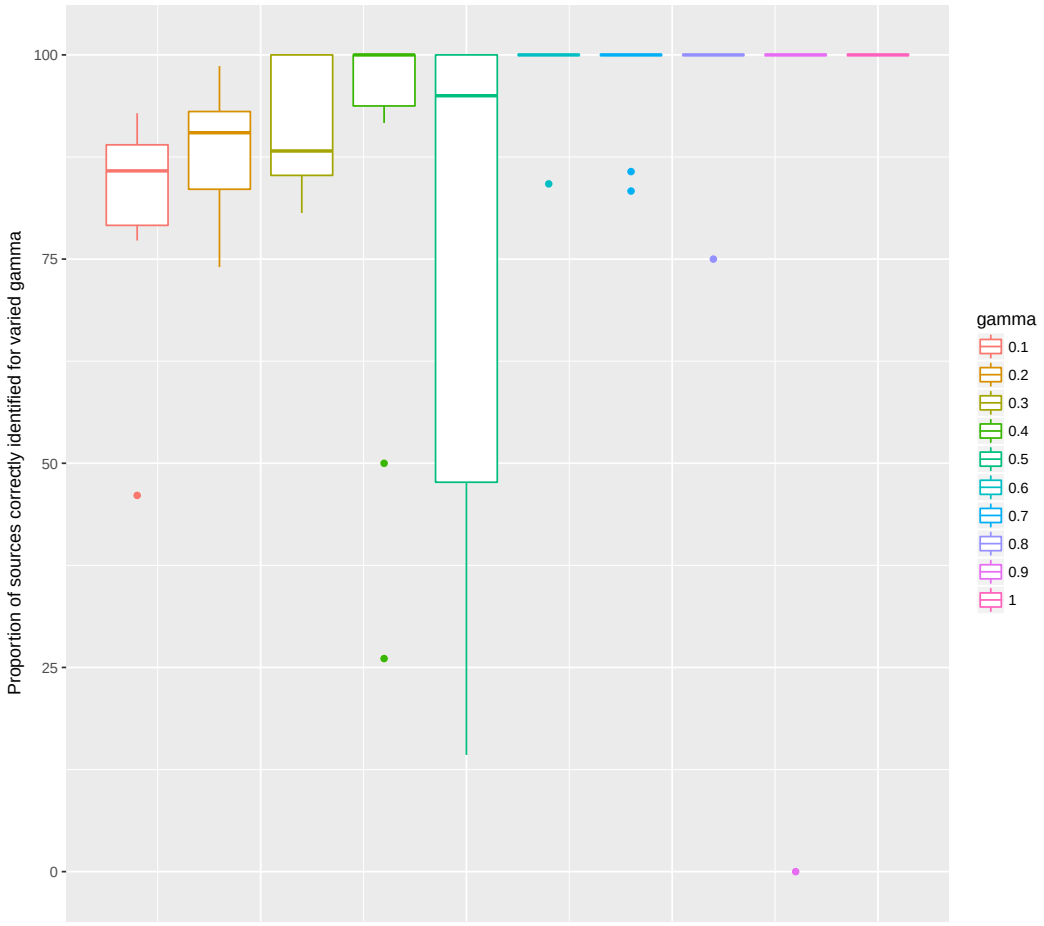


Figure G.13: Boxplots to show the proportion of infection sources for farms recovered correctly for simulations with varied values for parameter γ for the Time Dependent Distances model.

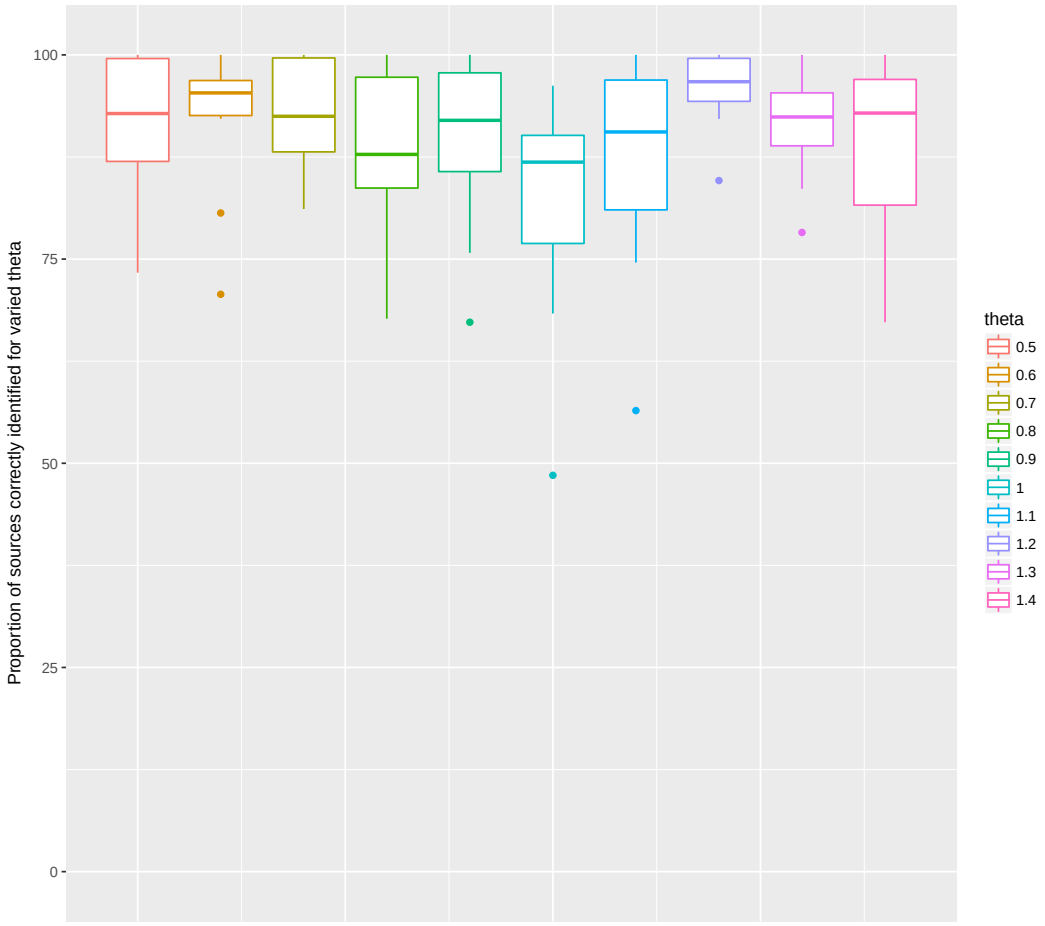


Figure G.14: Boxplots to show the proportion of infection sources for farms recovered correctly for simulations with varied values for parameter θ for the Time Dependent Distances model.

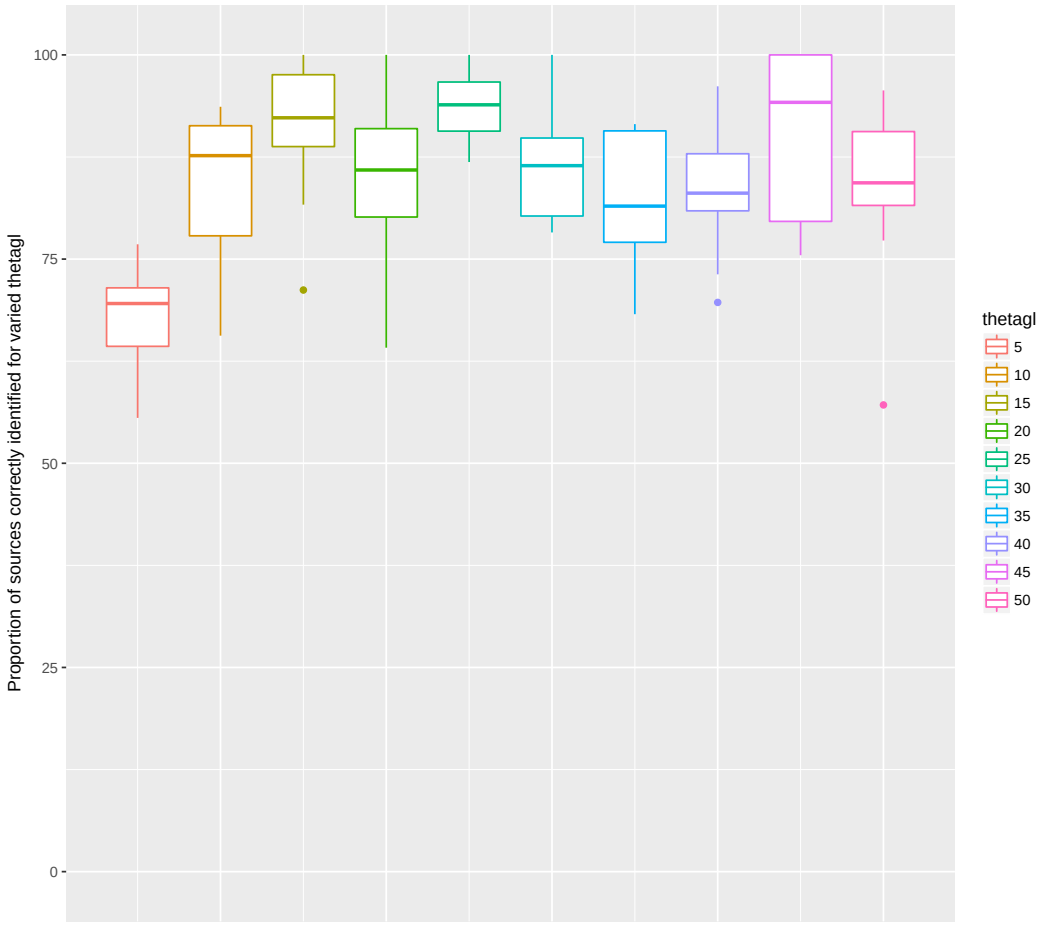


Figure G.15: Boxplots to show the proportion of infection sources for farms recovered correctly for simulations with varied values for parameter θ_{gl} for the Time Dependent Distances model.