Epidemic Modelling: An analysis of the 1861 Hagelloch measles outbreak

G14DIS

Mathematics 4th Year Dissertation

2020/21

School of Mathematical Sciences

University of Nottingham

Dominic Broadbent

Supervisor: Dr. Philip O'Neill

Project code: PDO D1

Assessment type: Review

I have read and understood the School and University guidelines on plagiarism. I confirm that this work is my own, apart from the acknowledged references.

Abstract

We investigate a severe 1861 measles epidemic that affected 188 children in the German village of Hagelloch. Efforts were undertaken to analyse and compare the importance of households and classrooms as transmission grounds for measles. To do this we produced simulations under various stochastic binomial epidemic models, and estimated the probabilities that infections occurred as a result of the various transmission pathways. Model parameters were first estimated using classical means before moving to a Bayesian approach with Markov Chain Monte Carlo (MCMC) methods. It was found that a model accounting for classroom, household and general transmission, allowing for different infection rates in different classrooms, produced the most accurate simulations. The corresponding transmission probabilities led us to the conclusion that classroom transmission was more significant than household transmission in the spread of the epidemic.

Contents

1 Introduction

In 1861 a severe measles epidemic afflicted the children of the German village of Hagelloch. The event was recorded in rich detail by a contemporary epidemiologist Dr Albert Pfeilsticker [1]. In comparison to the frequently very incomplete data sets by which we study modern epidemics, the information we have on this particular outbreak is quite complete. The specifics of the data ranges from the exact day each individual began experiencing different levels of symptoms, to detailed demographic data including the classroom and household each child was a part of.

The main area of interest of this report is in investigating the various ways in which individuals became infected. Specifically, we focus on comparing the relative significance of household and classroom transmission in the spread of the epidemic. We do this in order to better understand the ways in which measles spread through the village and to inform effective infection control strategies.

To accomplish this, after introducing the data and some pathological background on measles in Section 2, we formulate a general stochastic binomial epidemic model and state relevant assumptions in Section 3. Following this, in Section 4, we identify household and classrooms as significant staging grounds for the spread of measles via exploratory analysis on the data set. Then, in Section 5 we fit a basic control model, and state an algorithm for simulating epidemics under this model which we use to assess performance. Once we have a control model which assumes general mixing of the entire population, we formulate further models in Section 6 which account for additional classroom transmission and then simulate epidemics under these models. Then, in section 7, we focus on modelling and simulating the effects of household transmission before introducing the final most complex model in Section 8 which allows for both household and classroom transmission.

Throughout the modelling sections we estimate the probabilities that each infection occurred as a result of the various available transmission pathways and use Maximum Likelihood Estimation (MLE) to estimate model parameters. Asymptotic assumptions related

to the classical approach to parameter estimation cast doubt on the conclusions we could make. Therefore, in Section 9, this motivated a Bayesian approach to statistical inference through the use of Markov Chain Monte Carlo (MCMC) methods which produced similar, but more reliable parameter estimates. Initially, with the use of simulation results and relevant transmission probabilities, we then concluded that classroom transmission was a much more significant source of infection than household transmission. Specifically, the extended classroom model from Section 6.2 produced the most accurate simulation results, as this model allowed for the infection rates in the two classrooms to differ. This was surprising; we expected the classroom-household model from Section 8 to produce the strongest simulations as it allowed for more avenues of infection which more accurately reflects reality. Therefore, in Section 9.3 we used the MCMC framework to produce a sensitivity analysis on our model assumptions in order to check their validity. This analysis exposed that we were using non-optimal values for the length of two stages of the measles infection. Rectifying this, we found that the classroom-household model did then produce the best simulations. Further, the transmission probabilities associated with the model indicated that individuals in the population were on average more likely to be infected due to classroom transmission than household transmission.

2 Background

Before we introduce the epidemic model and perform some exploratory analysis on the observed epidemic, it is pertinent to introduce the data we have available and to understand the history, transmission pathways and clinical features of measles.

2.1 The Data

We have data on the 1861 measles epidemic that severely affected the isolated village of Hagelloch. This epidemic is of particular interest due to the completeness of the data available; information was collected daily by Dr Albert Pfeilsticker and thus the entire outbreak was recorded in detail [1]. Statistical analysis of infectious disease is often hindered by a lack of complete data. Therefore it is important to exploit this dataset in order to better understand transmission pathways and inform infection control strategies for viral outbreaks.

Hagelloch had previously suffered a serious measles outbreak in the winter of 1847, thus it is likely that only those individuals born after 1847 were susceptible to measles with the rest of the population having immunity. This hypothesis is supported by the dataset. The village had a population of 380 inhabitants; 197 were children, 188 of whom became infected over the course of the outbreak. Of these infected individuals, 187 were aged 14 or younger, and thus were born after the 1847 epidemic, while one was aged 15 and had avoided contracting measles previously [2]. Dr Pfeilsticker also provides information about the 12 children under the age of 14 who avoided infection. Seven were infants and presumably had placental immunity [3], four were kept totally isolated throughout the epidemic, and the final child was an immigrant who had previously contracted measles [2]. Thus we assume the entire at-risk population became infected over the course of the epidemic. This is necessary as no information was collated on the uninfected population, the vast majority of whom were very likely immune, or alternatively under strict isolation and thus at no risk of infection. Therefore, we limit our population of interest to those children in the village who did become infected.

We have a highly detailed table of information on our at-risk population of 188 children. Information relevant to our purposes includes the date each individual exhibited a fever, the date of rash eruption and the date of death (if applicable). Dr Pfeilsticker also recorded demographic data such as the individuals name, age, sex, their household number and location, and the classroom they attended.

PN	NAME	HN	AGE	SEX	PRO	ERU	CL	DEAD	HNX	HNY
81	Maurer	11	12.00	1	44	47	2	NA	91	87
128	Schneck	12	5.00	1	46	48	0	NA	96	90
129	Schneck	12	4.00	2	48	50	0	NA	96	90
149	Hipp	13	4.00	1	48	52	0	58	102	92
150	Hipp	13	0.50	1	54	58	0	NA	102	92
161	Hipp	13	6.00	2	55	59	0	NA	102	92

Figure 1: Typical lines from the dataset. PN = number associated with the individual (1-188), NAME = family name, HN = household number (1-56), PRO = date of first fever, ERU = date of rash eruption, CL = classroom number (0,1,2), DEAD = date of death, HNX/HNY = coordinates of the household

Other less relevant data not shown here includes information on complications as a result of other diseases, maximum fever temperature, the most likely individual behind each infection and the day of maximum fever.

Within the village, children over the age of approximately 6 attended school. Those between the rough ages of 7 and 10 attended "classroom 1" while those between the ages of 11 and 15 attended "classroom 2". These numbers are not exact; we do not have a date of birth, rather an age rounded to the nearest 0.5, and thus some 10 year-olds attend classroom 1 while others attend classroom 2, similarly some 6 and 7 year-olds attend classroom 1 while others do not attend school at all. It is clear however that school attendance, and more specifically, which classroom a child attends, is decided by age. Each child is also assigned a household number representing their home; there are 56 households in the village and for each an (x, y) coordinate is given.

2.2 History of Measles

Measles is an acute viral infectious disease which first evolved from rinderpest, a cattle-borne virus that in 2010 became only the second viral disease - after smallpox - to have been globally eradicated [4]. The first systematic description, including its distinction from other viral diseases such as smallpox and chickenpox, is credited to Persian physician Muhammad ibn Zakariya al-Razi who in the 9th century published The Book of Smallpox and Measles [5]. It is believed that, at that time, measles outbreaks were rare and the virus was not entirely adapted to humans. However, by the 12th century measles had fully diverged from rinderpest and had become a distinct virus [6]. Measles is an endemic disease, this means it is continually present in communities with many individuals gaining resistance. However, in populations previously unafflicted by measles, exposure can be deadly[1].

It is estimated that between 1855 and 2005, measles was responsible for the deaths of over 200 million people worldwide [8] with 7-8 million children dying each year before the development of the 1963 MMR vaccine [9]. The number of measles cases and resulting fatalities has dropped significantly in the last century. However, due to reduced vaccination uptake in recent years, the WHO Strategic Advisory Group of Experts on Immunisation concluded that "measles elimination is greatly under threat" and that "the disease has resurged in a number of countries that had achieved, or were close to achieving, elimination" [10]. Measles most severely affects developing countries with more than 95% of measles fatalities occurring in countries with low per capita incomes and weak health infrastructures[2] [10].

[1] In 1529, measles was responsible for killing two thirds of the indigenous Cuban population who had previously survived smallpox [7]

[2] In early 2019, a measles epidemic broke out in the Democratic Republic of Congo and by August of 2020 the outbreak was declared complete with a final death toll of over 7000 people; children under the age of five represented 90% of fatalities [11]

2.3 Transmission Pathways

It is clear that developing effective infection control strategies by analysing measles outbreaks and studying common transmission pathways is of vital importance, especially with vaccination rates in developed countries falling. The first step in this process is understanding how measles is spread. Measles is an airborne disease which spreads quickly from person to person via the exhalations, sneezes and coughs of an infected individual, or by direct contact with their nasal or throat secretions.

Measles is highly contagious with more than 90% of individuals who share living space with an infected person becoming infected themselves. The measles virus remains active and contagious in the air and on infected surfaces for up to 2 hours [10]. However, it does rapidly deactivate when exposed to heat or sunlight [12]. In fact, studies on the SARS-CoV2 virus indicated that infection chance is massively reduced outdoors, in well ventilated spaces, areas where close person-person contact is limited or in low-duration encounters. Conversely, the highest rate of transmission is found indoors or in poor ventilated areas, spaces where people are tightly packed and in encounters occurring over several hours [13]. Fewer direct studies of this type have been undertaken on measles, however the SARS-CoV2 and measles viruses have similar transmission methods and so a comparison can be made with some confidence.

These findings suggest that the children of Hagelloch would have most likely been infected while attending school or at home where they would have spent long periods of time in tightly-packed, poorly ventilated indoor spaces; the perfect environment for measles to be transmitted.

2.4 Clinical Features

Another area of significance when studying and modelling viral transmission is understanding the clinical features of the resulting disease, i.e. what are the properties of the infection life-cycle and when is an infected person contagious.

Once an at-risk individual has become infected by the measles virus, they enter an incubation period for an average of 10-12 days during which they do not experience symptoms. Following this, they enter a *prodromal* phase which lasts an average of 3-4 days, but has a large range of approximately 1-7 days. Prodrome is a medical term which indicates the onset of the early symptoms of an illness, for measles this is characterised by a fever and a cough. The individual will then erupt into the distinctive rash which is characteristic of a measles infection. This can last up to 5-6 days. If the individual recovers from the rash they will gain immunity; it is highly unlikely that a person will be at risk of reinfection by the measles virus [12].

It is important to note that an infected individual is not contagious until one day prior to prodrome. They then remain contagious until 3-4 days into the eruption of the rash, we denote this specific contagious phase as the *eruption* period. We also call the time between infection and when an individual becomes contagious as the *exposed* period.

3 Binomial Epidemic Model

Now that we have an understanding of the history, transmission pathways and clinical features of measles, we can introduce a model for epidemic spread among a population of individuals, state relevant assumptions and define some terminology. Note that our population of interest is closed, i.e. a fixed group with no one leaving or entering. Individuals are categorised in the following way:

Susceptible (S): currently healthy but could be infected,
Exposed (E): currently infected but not contagious,
Infective (I): currently infected and contagious,
Removed (R): currently immune, recovered or deceased.

Models with the progression

$$S \to E \to I \to R$$

are called SEIR models; we only consider the discrete-time variant of these models.

We can use the clinical features discussed in Section 2.4 to assume a typical measles infection timeline. Let E denote the length of the exposed period, x the length of the prodromal period and d the length of the eruption period. We assume the exposed period has a constant length of 10 days ($E = 10$), such that the incubation period (i.e. the time from intitial exposure to prodrome) lasts 11 days. We also assume the eruption period lasts 3 days ($d = 3$) and that the individual is removed immediately after this phase ends. The length of the prodromal period can vary significantly, thus we do not assume a constant length for x. Then, assuming an individual is contagious from one day prior to prodrome until the end of the eruption period, the entire infective period has length $1 + x + d$ days. Under these conditions a typical timeline of a measles infection is as follows.

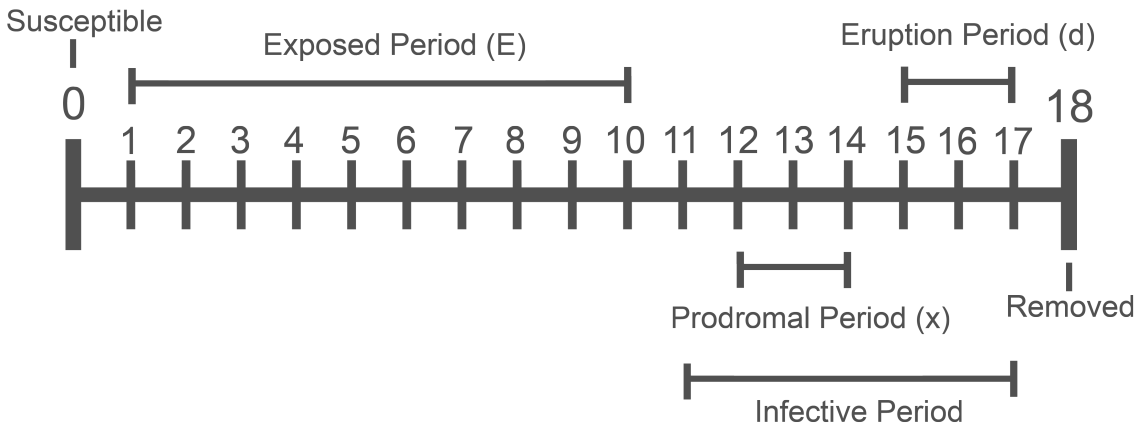


Figure 2: A typical timeline of a measles infection, under our model using discrete-time, with $x = 3$, $E = 10$ and $d = 3$. Note that each tick-mark represents a full 24-hour period

Because we are working with discrete-time, each tick mark in Figure 2 represents a full day or 24-hour span. Therefore, a prodromal period of length 3, such as the one we see in Figure 2, represents a full inclusive 3-day period. Informally, we are imagining that this individual began to suffer from a fever at midnight on day 11 until midnight on day 14.

Under our model we assume that each individual is equally infectious throughout their infectious period. In reality, this assumption is not entirely founded. Research into viral

transmission suggests that an individual's infectivity varies during the infectious period [14] and that an individual can even be more or less infective depending upon the viral load they are exposed to [15]. However, making the assumption of constant and equal infectivity of every individual greatly simplifies the model.

Now, consider a discrete-time model ($t = 0, 1, 2, 3...$) of a measles epidemic in a population of N individuals. For the observed epidemic we have $N = 188$. Then, at time t, we define the following population statistics,

$S(t)$ = number of susceptible individuals,
$E(t)$ = number of exposed individuals
$I(t)$ = number of infective individuals,
$R(t)$ = number of removed individuals.

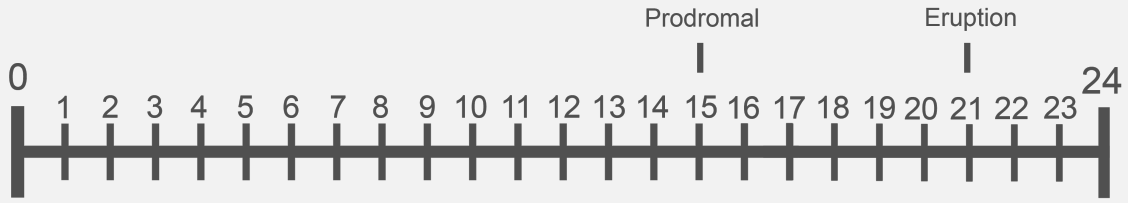
Note that that $S(t)+E(t)+I(t)+R(t) = 188$ for all t. Each susceptible at time t, remains susceptible, i.e. avoids infection at time $t+1$ with some *avoidance* probability we denote as $Q(I(t))$. That is, each at risk person avoids being infected with a probability that is a function of the number of contagious people in the population. Then, under these model assumptions, we have that

$$S(t+1) \sim \text{Binomial}(S(t), Q(I(t)),$$

We can calculate the population statistics $S(t), E(t), I(t)$ and $R(t)$ at each time t using the data and our model assumptions. To see this, first recall that we know the date each individual first exhibited a fever (i.e. prodrome) and the date that their rash erupted. Then, using this knowledge and the assumption of constant exposed and eruption period lengths (10 and 3 respectively) we can calculate the exact times when each individual was susceptible, exposed, latent and removed.

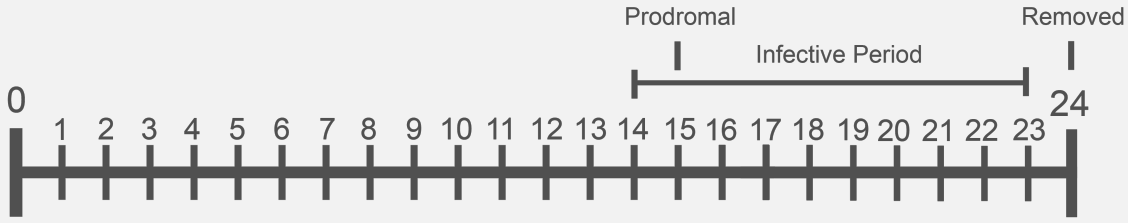
Example:

Consider an individual who began prodrome on day 15 and whose rash erupted on day 21.



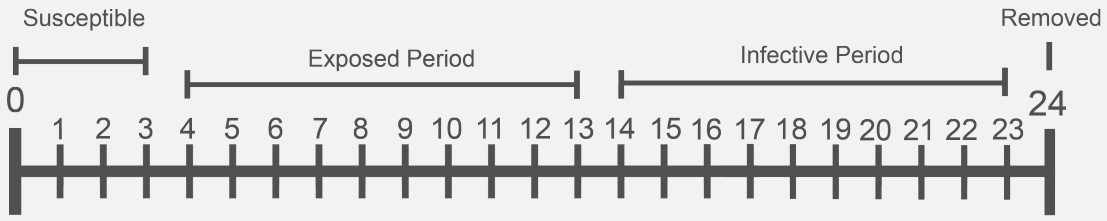
We assume that the eruption period has length 3 and immediately following this the individual is removed.

Then, we assume each person is infective from one day prior to prodrome until the end of the eruption period. Therefore this individual becomes infective on day 14 and stays infective until day 23.



Example (cont):

Additionally, we assume that each individual undergoes an exposed period of length 10 days which ends the day before they become infective. Thus the exposed period must end on day 13 and begin on day 4. We also have that each individual is susceptible before becoming exposed.



We have successfully calculated the entire timeline of this infection. The individuals is susceptible until day 3, they are exposed on day 4 before becoming infective on day 14. This contagious period lasts until day 23 and the following day they recover and are considered removed.

By performing the above calculation for each individual in the data set we can calculate our population statistics at each time t. Note that for each individual we make an assumption of the date they recover and become removed. However, 12 of the 188 individuals died during the course of the epidemic, these deaths all occurred following the eruption of the rash. Thus, for these individuals we replace our assumed removal date with the date of death. Now, all that is left to do is specify the avoidance probability $Q(I(t))$. There are two common approaches to this.

Reed-Frost Model: Set $Q(I(t)) = q^{I(t)}$. This model assumes each susceptible independently avoids infection from each infective with probability q.

Greenwood Model: Set $Q(I(t)) = \begin{cases} q, & I(t) \geq 1 \\ 1, & I(t) = 0 \end{cases}$. This model assumes that each susceptible avoids infection with probability q as long as there is one infective in the pop-

ulation.

Intuition suggests that the more infectives you interact with, the more likely you are to be infected, thus we proceed with the Reed-Frost model. Then, the binomial epidemic model takes the form,

$$S(t+1) \sim \text{Binomial}(S(t), q^{I(t)}),$$

We can interpret q in the context of the epidemic as a parameter that represents general mixing or interaction between individuals in the population. The smaller q is, the more likely individuals will be infected, and vice versa [3]. A toy example of this model in practice can be seen below.

Example: Let $I(0) = 3$ and $S(0) = 2$. Then, we are interested in $S(1)$, the number of susceptibles at time 1. Using the model specified above, we have that $S(1) \sim \text{Binomial}(S(0), q^{I(0)}) = \text{Binomial}(2, q^3)$. Therefore,

$$P(S(1) = k) = \binom{S(0)}{k} (q^3)^k (1-q^3)^{S(0)-k} = \binom{2}{k} q^{3k} (1-q^3)^{2-k}.$$

Then,

$$\begin{aligned}
P(S(1) = 2) &= \binom{2}{2} q^{3*2} (1-q^3)^{2-2} = q^6, \\
P(S(1) = 1) &= \binom{2}{1} q^{3*1} (1-q^3)^{2-1} = 2q^3(1-q^3), \\
P(S(1) = 0) &= \binom{2}{0} q^{3*0} (1-q^3)^{2-0} = (1-q^3)^2.
\end{aligned}$$

The only unknown left in this model is the parameter q. We want to estimate this parameter so that we can perform epidemic simulations under this model. In this way we can assess the accuracy and performance of the model by comparing the simulations to the observed epidemic. However, before we proceed with this, it is of interest to perform some exploratory analysis on the data set to gain further understanding of the epidemic

[3] Note that under the Reed-Frost model, $p = 1 - q =$ infection probability.

data under these new model assumptions and to see what our simulations should look like.

4 Exploratory Analysis

We began by producing a plot of the population statistics under the model from Section 3 in order to get a sense of how the epidemic progressed.

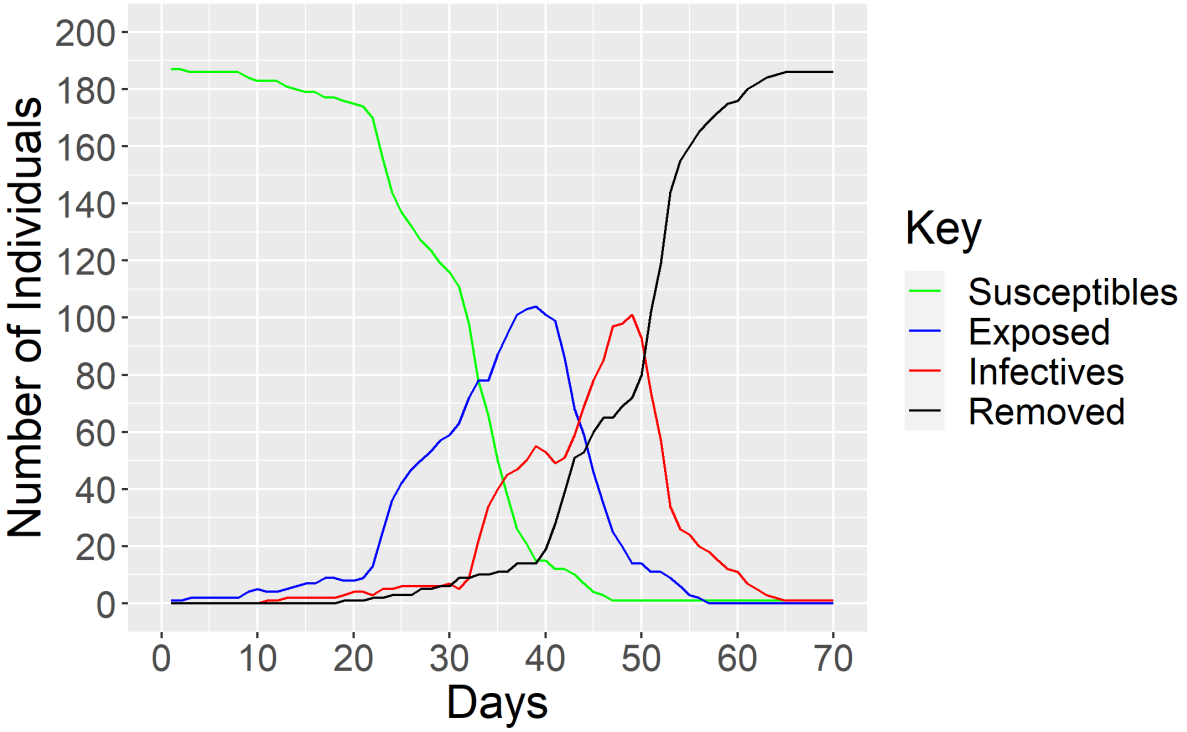


Figure 3: Plot of the population statistics $S(t), E(t), I(t)$ and $R(t)$ throughout the epidemic, assuming $E = 10, d = 3$

Figure 3 shows the progression of the epidemic over time in terms of our model population statistics $S(t), E(t), I(t)$ and $R(t)$. By analysing the green line representing the number of susceptibles, we can see that the epidemic initially spread slowly. By day 25 measles had begun spreading much more quickly, with most of the population having been infected by approximately day 38. From this point onwards the remaining susceptibles were infected at a much lower rate. We can also see, by looking at the red line indicating infective individuals, that the epidemic appeared to progress in two waves. From day 30 to 40 many

individuals became infective, this number then dropped slightly for a few days before once again rising quickly to a peak around day 50. We want our model to produce simulations that closely match the shape of these observed population statistics.

In our model we make no assumptions of the length of the prodromal period x, instead we we use our other assumptions and the observed prodromal and eruption dates to construct infection timelines. However, when simulating an epidemic we will only know when each individual was first infected. Therefore, in order to build the timeline, we need to assign a prodromal period length for each infection. We could do this by assuming a constant length, however, unlike the eruption and exposed periods, the length of the prodromal period of measles can vary significantly (See Section 2.4).

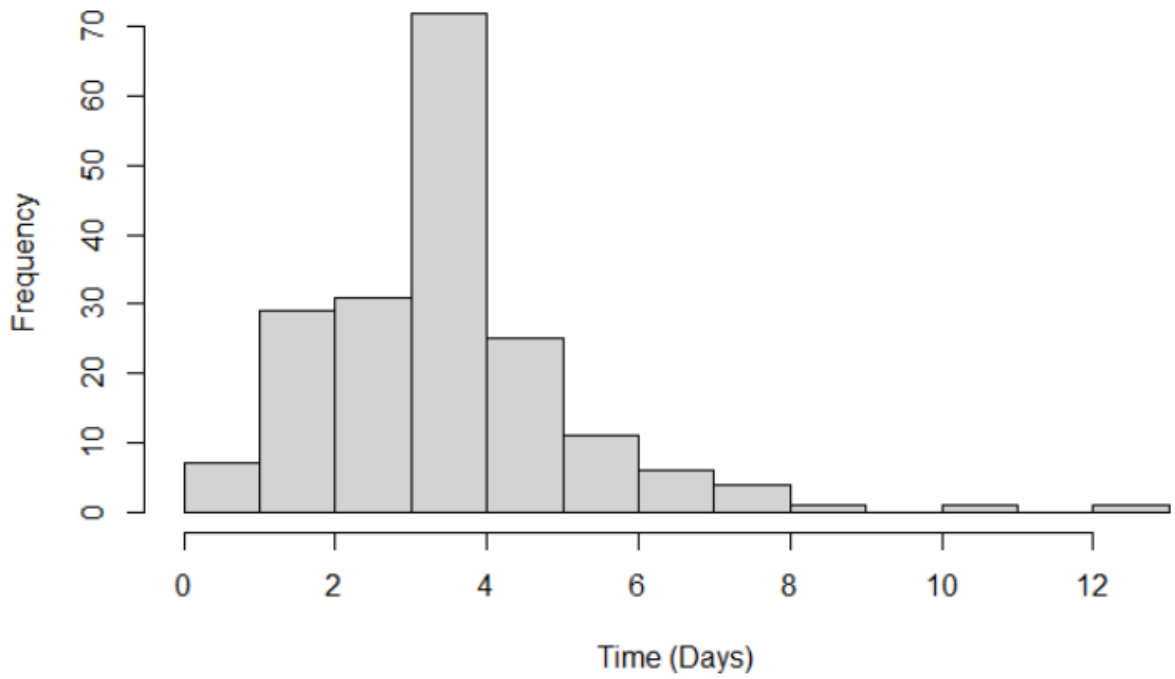


Figure 4: Histogram of the length of the observed prodromal periods

From this plot we can see that the most common observed prodromal period lasts 4 days, with the mean also being approximately equal to 4. However, it can range in length anywhere from 0 to 13 days. To simplify the model, we could make the assumption of a constant prodromal period length, with the most obvious choice being 4. However this

runs the risk of reducing the accuracy of our model as we would not be accurately reflecting the reality that the length of the prodormal period can vary significantly from person to person. Therefore, in order to best match the dynamics of the observed epidemic, rather than assuming a constant length, we will sample with replacement from this set of observed prodromal periods when we come to simulate under our model.

4.1 Classroom Transmission

Recall from Section 2.1 that we have access to data of the classroom that each individual attended. We categorise those children who were too young to attend class, and thus were not present at a physical location, as classroom 0. The remaining children were separated into classroom 1 and classroom 2 based on age. The population in each of these classrooms is not equal; there are 90, 30 and 68 individuals in classroom 0, 1 and 2 respectively. In Section 2.3 we identified that the classroom was likely an area of significant viral transmission and thus it is of interest to investigate this.

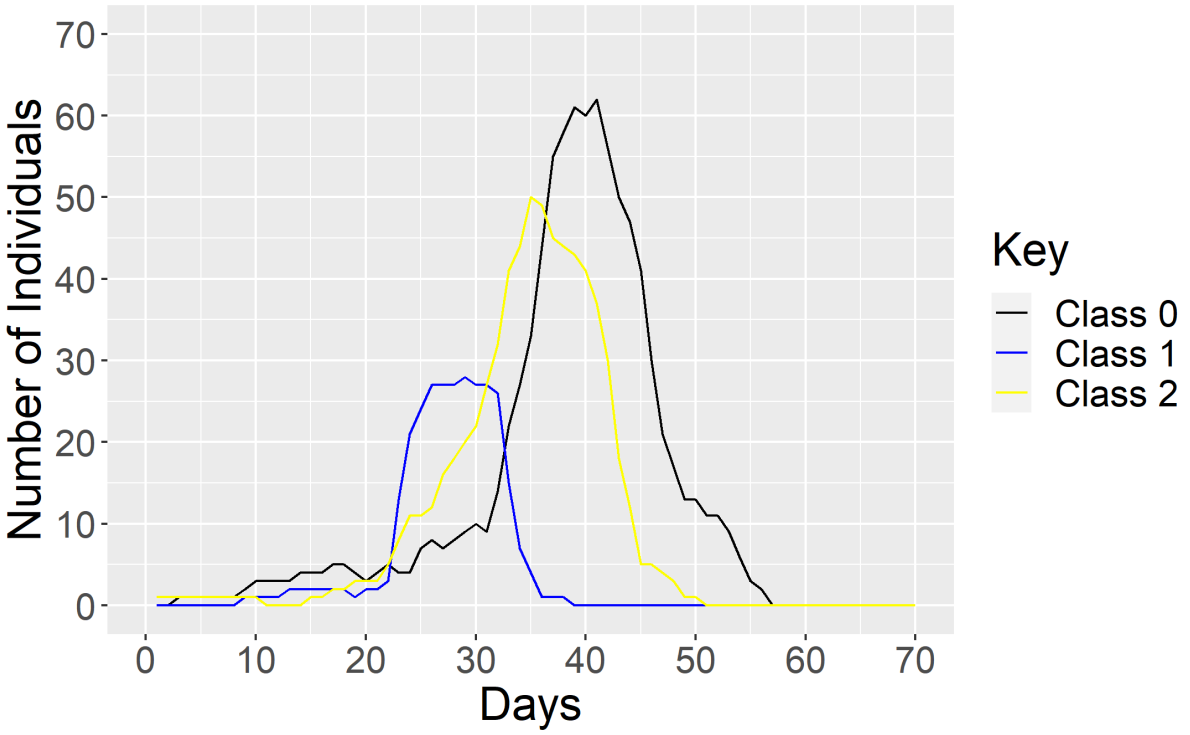


Figure 5: Plot of the number of exposed individuals $E(t)$ separated into classrooms, assuming $E = 10, d = 3$

In Figure 5, we can see that the epidemic clearly spread through the various classrooms in a staggered fashion. The children in classroom 1 were the first to be widely affected with measles. The next widely affected population was those in classroom 2, followed by classroom 0. We also highlight that the individuals in classroom 1 were all infected over a much shorter time frame than the other two classes. These findings suggest that the epidemic spread via the classrooms, as we would expect, and thus it is worth accounting for classroom transmission in our modelling attempts.

4.2 Household Transmission

We also know the household that each individual was a part of. The village is comprised of 56 households which contain an average of 3 individuals and range in size from 1 to 8 individuals. Because of the number of households and the relatively small populations in each, it is infeasible to produce a plot such as Figure 5 that will allow us to assess if household transmission is significant. Instead, recall that due to our model assumptions, we have access to the timeline of each individuals state throughout the epidemic. Then, by focusing in on a single household, we can overlap each individual's timelines and assess whether it was possible, under our model assumptions, that they were infected by a member of their own household. Doing this for each household we can calculate the proportion of infections that could have occurred due to household transmission, discounting the first infection in each household.

Example:

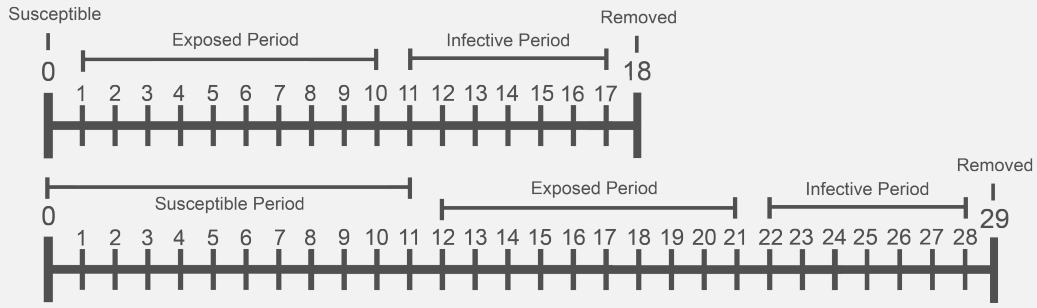


Figure 6: Household of two individuals with overlapped infection timelines

Figure 6 represents a household with two individuals and displays their infection timelines. We can see that the start of the first individual's infective period predates the start of the second individual's exposed period. Therefore it is possible that the second individual was infected by the first. Thus, for this specific household, the proportion of new infections that could have occurred via household transmission is equal to 1.

Calculating this proportion for each household and averaging gives a value of 0.310. Therefore, we can say that approximately 31% of all infections where household transmission was possible, could have occurred as a result of household transmission. This represents a significant proportion of the total number of infections, suggesting that household transmission may have been a significant factor in the spread of the epidemic.

4.3 Spatial Transmission

Recall that we also know the location of each household in the village of Hagelloch. Intuitively, we expect that the epidemic could have spread from household to household, with distance playing a significant factor in influencing how individuals may have interacted in the village. That is, the closer a household is to another household, the bigger the chance their respective individuals will interact with each other, and thus the more likely it is that they will have infected each other. This is difficult to capture precisely, but one way we can attempt to see this is by *clustering* the households together into various groups

based on distance, which we can then analyse for evidence of transmission. If we see significant separation in when each *cluster* began to experience symptoms, then we may have evidence that suggests that measles was spread inside these clusters. A good way to think of what we are modelling here is that a child living in a small village might interact with the children on their road more frequently than the children say, on the other side of town.

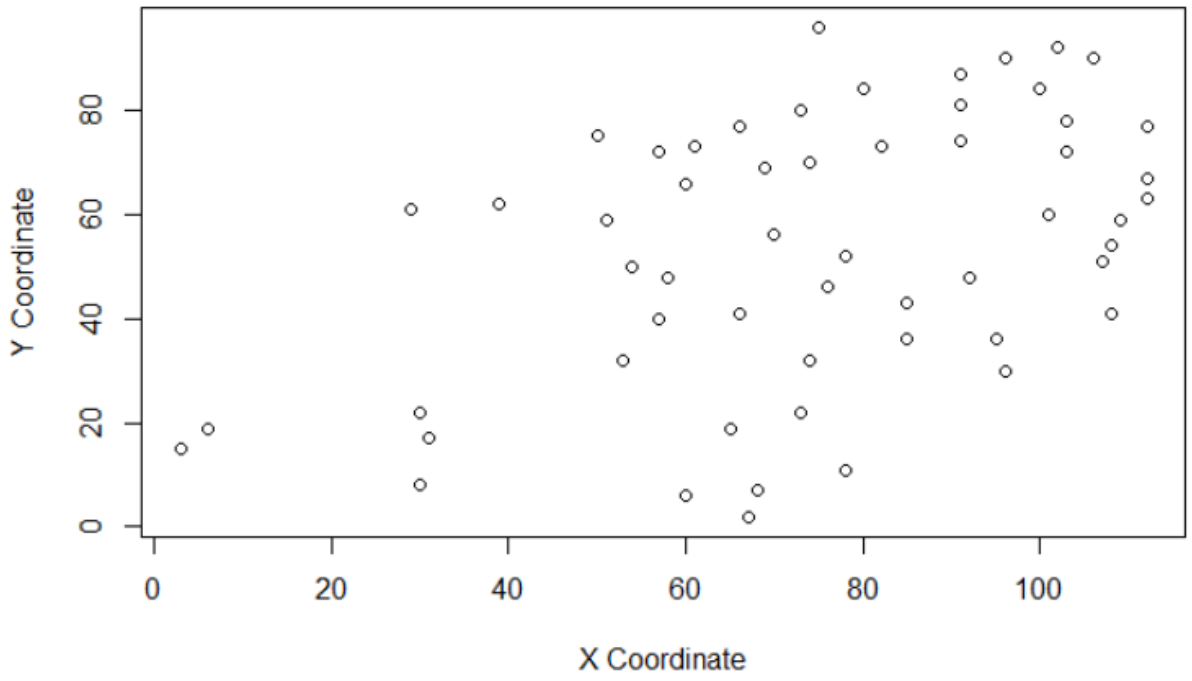


Figure 7: Map of the households of Hagelloch

In order to investigate this, we need an algorithm that clusters the households together; one example is the K-means clustering algorithm. Applying it to the households allows us to partition them into groups which we can then analyse for evidence of transmission. The K-means clustering algorithm works as follows:

Algorithm 1: K-Means Clustering Algorithm [16]

Randomly choose the coordinates of K households, $\mathbf{m}_1, \mathbf{m}_2, ...\mathbf{m}_K$ without replacement ;

Initialise K empty sets $C_1, ..., C_K$;

1. **for** *each household* $\boldsymbol{h}_i = (x_i, y_i)$ **do**

$j = \min_{k'=1,...,K} ||\mathbf{h}_i - \mathbf{m}_{k'}||$;

$\mathbf{h}_i \in C_j$;

end

2. **for** *each* $\boldsymbol{m}_i$ **do**

$\mathbf{m}_i = \frac{1}{|C_i|} \sum_{j \in C_j} \mathbf{h}_k$

end

Repeat 1 and 2 until convergence

We have 56 individual households and we want to separate them into approximately equal groups with enough individuals in each such that we can identify any transmission patterns easily. Choosing $K = 7$, will output 7 groups consisting of an average of 8 households and approximately 26 individuals, depending on which initial households are chosen. These clusters are big enough to take some time for measles to infect each individual but small enough to allow us to conceptualise that a single individual might interact with each person in the cluster on any given day. Applying the algorithm until convergence produces the following clusters:

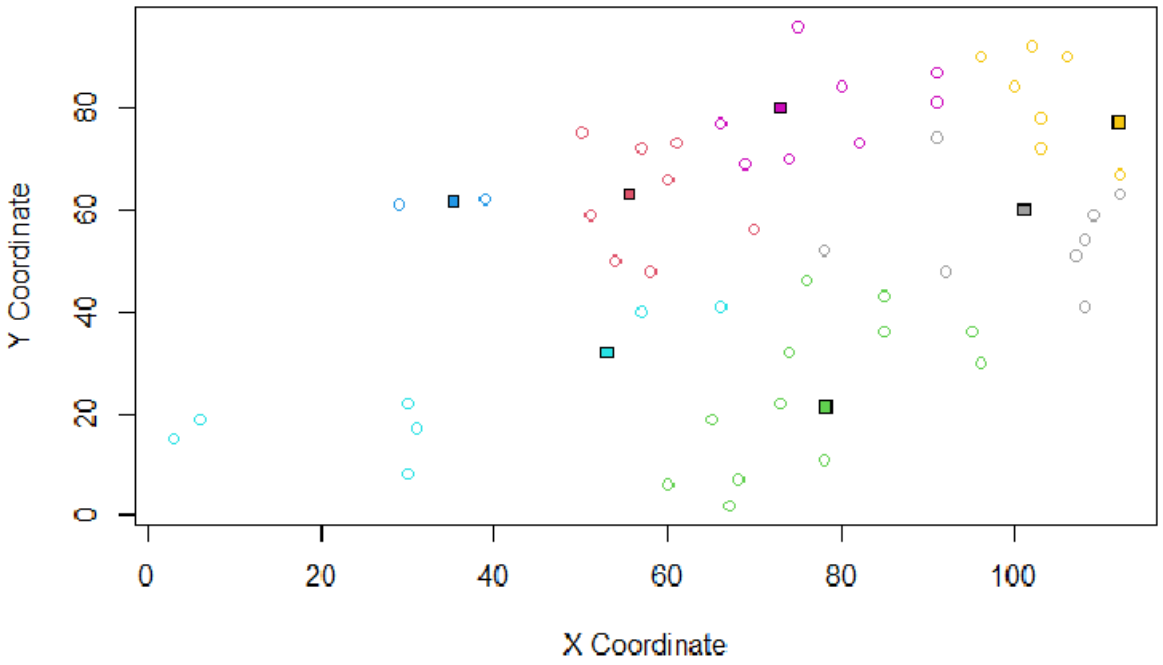


Figure 8: Map of the household clusters of Hagelloch produced by the K-means clustering algorithm with $K = 7$

We can now use these clusters to investigate whether or not the epidemic appeared to spread spatially. To do this we plot when each individual in each cluster started their prodromal period, i.e. began to suffer from symptoms. If the individuals in cluster 1, for example, began experiencing symptoms 20 days before all the individuals in cluster 2, then we have evidence to say that measles may have been spreading inside these clusters separately.

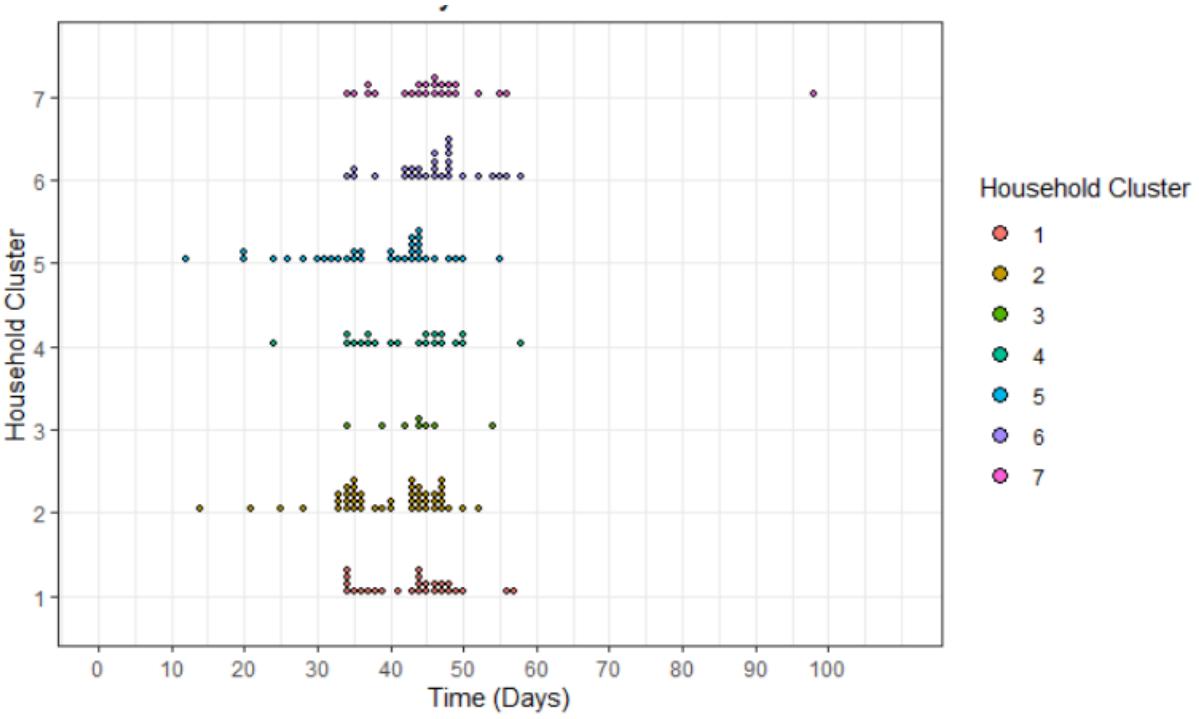


Figure 9: Dotplot showing the prodromal dates of each individual coloured by the household cluster they belong to

From Figure 9, at least using this approach, we can surmise that there does not appear to be a spatial component in the spread of the epidemic; each cluster seems to experience infections over approximately the same time frame. If spatial transmission was a significant factor in the spread of measles we would expect there to be more obvious separation in when each cluster first began to notably suffer from measles, such as the pattern we saw in our investigation of classroom transmission[4].

5 Base Model

We can use our binomial model to produce simulations of the epidemic and assess the strength of the fit to the observed epidemic. Before we proceed, we need to estimate q, we can do this via Maximum Likelihood Estimation (MLE) on the likelihood function,

$$L(q) = \prod_t (q^{I(t)})^{S(t+1)} (1 - q^{I(t)})^{(S(t) - S(t+1))}.$$

[4]Note that we ran the clustering algorithm with a wide range of values of K and found no evidence of spatial transmission in any case

Note that we are only interested in maximising the likelihood, thus we have dropped the constant combinatorial term. Now, taking logs, we find the log-likelihood,

$$\log(L(q)) = \sum_t S(t+1)I(t)\log(q) + [S(t) - S(t+1)]\log(1 - q^{I(t)}).$$

We can then optimise the log-likelihood numerically using the *optim* function in R. For univariate functions, *optim* uses Brent's method which is a hybrid root-finding algorithm combining the bisection method, the secant method and inverse quadratic interpolation [17]. Using this function gives us the estimate $\hat{q} = 0.9929226$. We also know that the MLE is asymptotically normal for large n. Using this fact, and the approximate Hessian produced by *optim*, we can calculate an approximate 95% confidence interval for this estimate.

$$\begin{aligned}
&\hat{q} \pm 1.96\frac{1}{\sqrt{(\sigma_q^2)}} \\
&= 0.9929226 \pm 1.96\frac{1}{\sqrt{(3799613)}} \\
&= 0.9929226 \pm 0.00100551
\end{aligned}$$

With $n = 188$ observations, the asymptotic normality assumption should be fairly valid. This estimate makes intuitive sense; we expect an avoidance parameter q that is very close to 1 such that the avoidance probability $Q(I(t)) = q^{I(t)}$ is not too small. If $Q(I(t))$ was very small, the epidemic would spread to the entire population very quickly.

5.1 Epidemic Simulation

The model is now fully specified. We state a simulation algorithm, recalling that we assume the exposed period E has length 10, the eruption period d has length 3 and for each prodromal period x we sample a value $\hat{x}$ with replacement from the observed prodromal values.

Algorithm 2: Basic Model Simulation Algorithm

Intialise $S(0)$, $I(0)$, $E(0), R(0)$, q, E, d,

Maximum epidemic length T

for t *from 1 to* T **do**

  $S(t) = \text{Bin}(S(t-1), q^{I(t-1)})$

  $\hat{I} = S(t-1) - S(t)$

  for t' *from* t *to* $(t + E - 1)$ **do**

    $E(t') = E(t') + \hat{I}$

  end

  for i *from 1 to* $\hat{I}$ **do**

    $\hat{x}_i$ = sample prodromal period length x

    for $\hat{t}$ *from* $(t + E)$ *to* $(t + E + \hat{x}_i + d)$ **do**

      $I(\hat{t}) = I(\hat{t}) + 1$

    end

    for $\tilde{t}$ *from* $(t + E + \hat{x}_i + d + 1)$ *to* T **do**

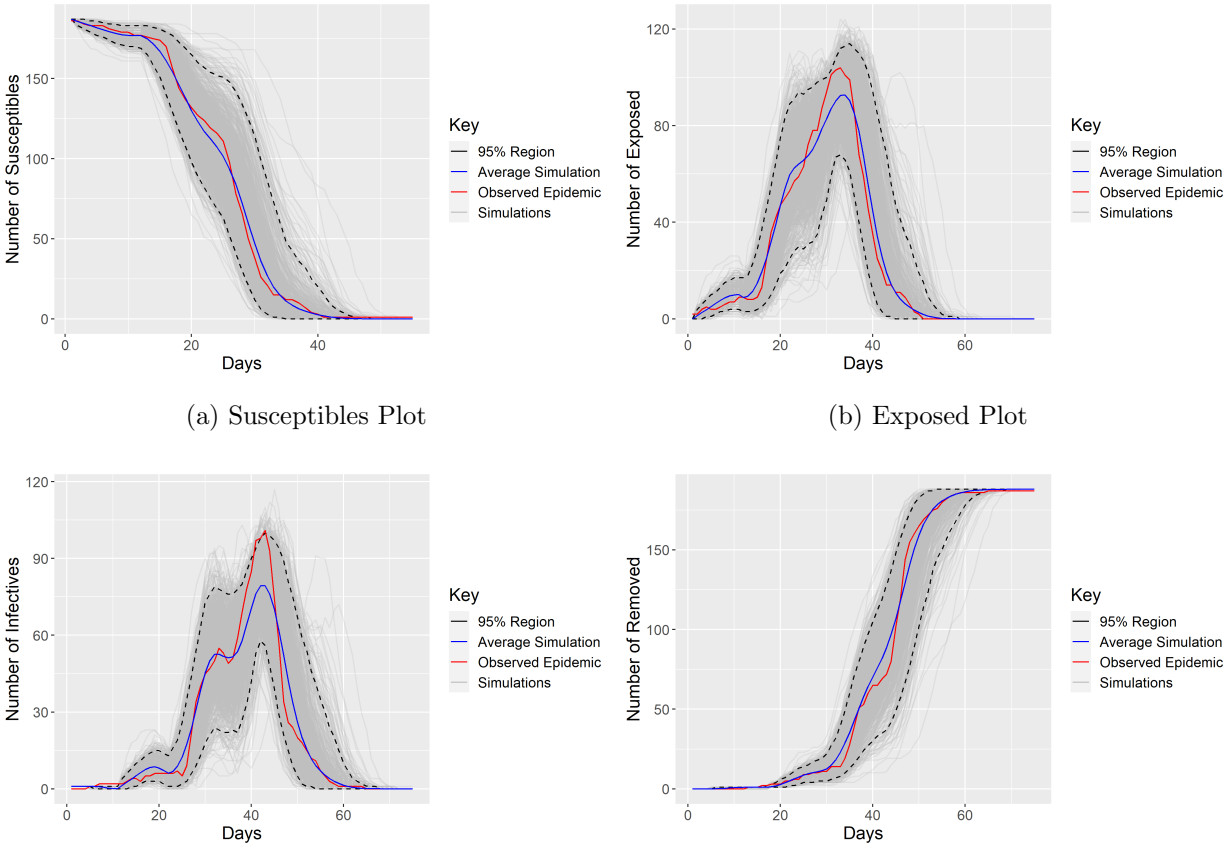
      $R(\tilde{t}) = R(\tilde{t}) + 1$

    end

  end

end

This algorithm simulates an epidemic via *nested for loops* that increment on the various population statistics over the course of the epidemic. We now produce simulations using the algorithm in order to check the performance of our model.



(a) Susceptibles Plot

(b) Exposed Plot

(c) Infectives Plot

(d) Removed Plot

Figure 10: Plot showing 1000 simulated epidemics using the basic model with $q = 0.9929226$, the average simulation and the observed epidemic

Each light grey line indicates a single simulation with the black dashed lines representing the region enclosed by the 2.5% and 97.5% percentiles of these simulations. We can use these figures to to assess how the model is performing by comparing the average simulation in blue to the observed epidemic in red. The overall fit to the observed epidemic is quite good. We can see from Figure 10d, which displays the number of removed individuals throughout the epidemic, that the average simulation is closely following the observed epidemic. However, when inspecting Figure 10c, we can see that around day 40 the number of maximum infectives in the average simulation falls short of the observed epidemic. Up until that point, the average simulation was closely fitted to the up and down motion of the observed data. We also see a similar result in Figure 10b for the number of exposed individuals.

Despite these seemingly good results, the current model does not accurately represent the transmission pathways that we expect were present in the observed epidemic, i.e. the household and classroom transmission. Rather it assumes that the entire population mixes with each other in an equivalent manner. As a result, under this model we do not have the ability to observe the epidemic inside the classrooms and households; we could have a good fit to the overall epidemic but be poorly representing what is going on inside these sub-populations.

6 Classroom Models

We have evidence that suggests that classroom transmission was a significant factor in the spread of measles throughout Hagelloch. This is motivation to include classroom interaction in our binomial epidemic model.

6.1 Initial Classroom Model

We begin by separating the population into the three classrooms. At time t, with $i = 0, 1, 2$, we have,

$S_i(\mathrm{t})$ = number of susceptibles in classroom i,
$E_i(\mathrm{t})$ = number of exposed in classroom i,
$I_i(\mathrm{t})$ = number of infectives in classroom i.
$R_i(\mathrm{t})$ = number of removed in classroom i,

Note that $S_0(t) + S_1(t) + S_2(t) = S(t)$, similarly this holds for the infective, exposed and removed populations statistics. Now, we retain the avoidance parameter q which represents general mixing in the population, however we introduce a further avoidance parameter q_c which accounts for interaction inside a classroom. Then, with this addition, and under the previous general model assumptions, we have a two-parameter binomial

model of the following form,

$$
\begin{aligned}
S_0(t+1) &\sim \text{Binomial}(S_0(t), q^{I(t)}),\\
S_1(t+1) &\sim \text{Binomial}(S_1(t), q_c^{I_1(t)} q^{I_0(t)+I_2(t)}),\\
S_2(t+1) &\sim \text{Binomial}(S_2(t), q_c^{I_2(t)} q^{I_1(t)+I_2(t)}).
\end{aligned}
$$

Note that because class zero individuals are too young to attend a physical classroom, we do not include a parameter for classroom spread in the binomial model for $S_0(t+1)$. A toy example of this model in practice can be seen below.

Example: Let $I_0(0) = I_1(0) = I_2(0) = 1$ and $S_0(0) = S_1(0) = S_2(0) = 2$. We are interested in $S_0(1)$, $S_1(1)$ and $S_2(1)$, i.e. the number of susceptibles in each classroom at time 1. Using the model specified above, we have that

$$
\begin{aligned}
S_0(1) &\sim \text{Binomial}(S_0(0), q^{I(0)}) = \text{Binomial}(2, q^3)\\
S_1(1) &\sim \text{Binomial}(S_1(0), q_c^{I_1(0)} q^{I_0(0)+I_2(0)}) = \text{Binomial}(2, q_c q^2)\\
S_2(1) &\sim \text{Binomial}(S_2(0), q_c^{I_2(0)} q^{I_1(0)+I_2(0)}) = \text{Binomial}(2, q_c q^2)
\end{aligned}
$$

Note that $S_0(1)$ here is equivalent to the example we saw in Section 3. We can also see that, in this example, $S_1(1)$ and $S_2(1)$ use an equivalent model. Focusing on $S_1(1)$, we have that

$$
P(S_1(1) = k) = \binom{S_1(0)}{k} (q_c q^2)^k (1 - q_c q^2)^{S_1(0)-k} = \binom{2}{k} (q_c q^2)^k (1 - q_c q^2)^{2-k}.
$$

Then,

$$
\begin{aligned}
P(S_1(1) = 2) &= \binom{2}{2} (q_c q^2)^2 (1 - q_c q^2)^{2-2} = q_c^2 q^4,\\
P(S_1(1) = 1) &= \binom{2}{1} (q_c q^2)^1 (1 - q_c q^2)^{2-1} = 2 q_c q^2 (1 - q_c q^2),\\
P(S_1(1) = 0) &= \binom{2}{0} (q_c q^2)^0 (1 - q_c q^2)^{2-0} = (1 - q_c q^2)^2.
\end{aligned}
$$

The unknown quantities here are q and q_c, all that remains to do is estimate them via

maximum likelihood estimation on the likelihood. Firstly we have

$$L_0(q, q_c) = \prod_t (q^{I(t)})^{S_0(t+1)}(1 - q^{I(t)})^{(S_0(t) - S_0(t+1))},$$

$$L_1(q, q_c) = \prod_t (q_c^{I_1(t)} q^{I_0(t)+I_2(t)})^{S_1(t+1)}(1 - q_c^{I_1(t)} q^{I_0(t)+I_2(t)})^{(S_1(t) - S_1(t+1))},$$

$$L_2(q, q_c) = \prod_t (q_c^{I_2(t)} q^{I_0(t)+I_1(t)})^{S_2(t+1)}(1 - q_c^{I_2(t)} q^{I_0(t)+I_1(t)})^{(S_2(t) - S_2(t+1))},$$

as the likelihoods for each individual classroom. Then,

$$L(q, q_c) = L_0(q, q_c) L_1(q, q_c) L_2(q, q_c),$$

represents the likelihood for the overall population. Now, taking logs we have,

$$\log(L_0(q, q_c)) = \sum_t S_0(t+1) I_0(t) \log(q) + (S_0(t) - S_0(t+1)) \log(1 - q^{I_0(t)}),$$

$$\begin{aligned} \log(L_1(q, q_c)) = \sum_t & S_1(t+1)[(I_0(t) + I_2(t)) \log(q) + I_1(t) \log(q_c)] \\ & + (S_1(t) - S_1(t+1)) \log(1 - q_c^{I_1(t)} q^{I_0(t)+I_2(t)}), \end{aligned}$$

$$\begin{aligned} \log(L_2(q, q_c)) = \sum_t & S_2(t+1)[(I_0(t) + I_1(t)) \log(q) + I_2(t) \log(q_c)] \\ & + (S_2(t) - S_2(t+1)) \log(1 - q_c^{I_2(t)} q^{I_0(t)+I_1(t)}). \end{aligned}$$

Finally, the overall log-likelihood is as follows,

$$\log(L(q, q_c)) = \log(L_0(q, q_c)) + \log(L_1(q, q_c)) + \log(L_2(q, q_c)).$$

We can once again maximise the log-likelihood numerically using the *optim* function in R. For multivariate functions *optim* uses the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm. This is an iterative optimisation method that works by determining an ascent direction (towards the maximum) by preconditioning the gradient with curvature information. This information is obtained by gradually improving an approximation to the Hessian matrix of the cost function. The Hessian matrix itself is found via a generalised secant method on the gradient [18]. Using this function provides the estimates

$\hat{q} = 0.9952665$ and $\hat{q}_c = 0.9646576$. Once again, using the asymptotic normality of the MLE and the approximate Hessian produced by *optim*, we can calculate an approximate 95% confidence interval for these estimates.

$$\begin{aligned}
&\hat{q} \pm 1.96 \frac{1}{\sqrt{(\sigma_q^2)}} \\
&= 0.9952665 \pm 1.96 \frac{1}{\sqrt{(4822534.20)}} \\
&= 0.9952665 \pm 0.0008925209
\end{aligned}$$

$$\begin{aligned}
&\hat{q}_c \pm 1.96 \frac{1}{\sqrt{(\sigma_{q_c}^2)}} \\
&= 0.9646576 \pm 1.96 \frac{1}{\sqrt{(85499.96)}} \\
&= 0.9646576 \pm 0.00670306
\end{aligned}$$

These estimates make intuitive sense, we have that $q_c < q$, this indicates that an individual is more likely to be infected inside a classroom than outside it, which corroborates the research on viral transmission from Section 2.3. We also note that the confidence interval for q is a factor of 10 smaller than the confidence interval for q_c. This is unsurprising as there are fewer individuals of age to attend school than in the total population.

We are now in a position to state a simulation algorithm and then produce simulations. However, now that we have a model with two transmission pathways, we can estimate the proportion of infections in the observed epidemic that occurred via these pathways. This will give us insight into the significance of classroom transmission in the spread of the measles epidemic.

6.1.1 Transmission Probabilities

We are interested in knowing how significant the various transmission pathways discussed in Section 2.3 are to the evolution of the epidemic. One way to quantify this is to calculate the probability, under our model, that an infection came from a particular source. More specifically, we would like to know the probability that *given* an infection occurred, it was

a sole result of classroom transmission. That is, we want to find

$$P(Classroom\ Infection|Infection).$$

Then, by using the Kolmogorov definition of conditional probability, the above is equivalent to,

$$\frac{P(Classroom\ Infection\ \cap\ Infection)}{P(Infection)}.$$

By noting that an infection as a result of classroom non-avoidance is an event that is a subset of the infection occurring in the first place, we have that the above is equivalent to,

$$\frac{P(Classroom\ Infection)}{P(Infection)}.$$

Now, the probability of a classroom infection is equivalent to the probability that the susceptible avoids infection as a result of general transmission, but fails to avoid infection due to classroom transmission. Thus the above probability is equivalent to,

$$\frac{P(General\ Avoidance\ \cap\ Classroom\ Non\text{-}Avoidance)}{P(Infection)}.$$

Finally, under the Reed-Frost model, we assume each susceptible independently avoids infection from each infective. Therefore we have that,

$$P(Classroom\ Infection|Infection) = \frac{P(General\ Avoidance)P(Classroom\ Non\text{-}Avoidance)}{P(Infection)}.$$

Note that under our binomial epidemic model, we are not explicitly modelling who-infected-whom and in particular, it is possible that two or more individuals can "infect" the same susceptible on any given day. Instead, here we estimate the probability that an infected individual failed to avoid at least one infective inside their classroom while simultaneously successfully avoiding the rest of the infective population outside the classroom. In this way we can get a sense of how many infections occurred as a result of classroom transmission. This is best understood through the use of a toy example.

Example: Consider a population consisting of a single susceptible individual who attends classroom 1 and two infectives; one who attends classroom 2 and the other, classroom 1. We are interested in finding the probability that, given an infection occurred, the susceptible avoided the class 2 infective but failed to avoid the class 1 infective.

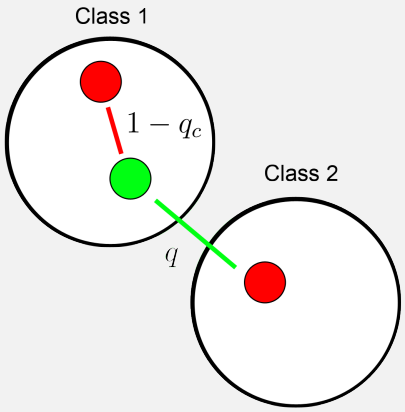


Figure 11: Visualisation of this event

The above figure represents this state; the red line shows an infection and the green line shows an avoidance. Then,

$$\frac{P(General\ Avoidance)P(Classroom\ Non\text{-}Avoidance)}{P(Infection)} = \frac{q(1-q_c)}{1-qq_c},$$

where we have that,

$$\begin{aligned} P(Infection) &= 1 - P(Avoids\ Infection) \\ &= 1 - P(General\ Avoidance \cap Classroom\ Avoidance) \\ &= 1 - qq_c \end{aligned}$$

Now, consider an individual who attends classroom 1. The probability that this individual is infected at time $t+1$ as a direct result of classroom transmission is the following,

$$\frac{P(General\ Avoidance)P(Classroom\ Non\text{-}Avoidance)}{P(Infection)} = \frac{q^{I_0(t)+I_2(t)}(1-q_c^{I_1(t)})}{1-q^{I_0(t)+I_2(t)}q_c^{I_1(t)}}.$$

Similarly, for an individual who attends classroom 2,

$$\frac{P(General\ Avoidance)P(Classroom\ Non\text{-}Avoidance)}{P(Infection)} = \frac{q^{I_0(t)+I_1(t)}(1-q_c^{I_2(t)})}{1-q^{I_0(t)+I_1(t)}q_c^{I_2(t)}}.$$

Note that this is not valid for individuals in classroom 0. Under our model they do not undergo the same classroom interactions and thus the probability they are directly infected due to classroom transmission is zero.

Now, we can calculate this probability for each infection in the observed epidemic and then average to calculate the overall proportion of infections that occurred as a direct result of classroom transmission. We find that, under this model, approximately 36.5% of the total infections in the Hagelloch epidemic occurred as solely due to classroom transmission. Restricting our attention to just those infections where classroom transmission was possible, i.e. discounting those individuals too young to attend a classroom, this proportion rises to 71%.

Using a very similar probability framework discussed above, we can calculate the probability that an infection occurred due to only general transmission. Note that for individuals in classroom 0 this probability is equal to one as they only undergo general interactions. Performing this calculation for each infection and averaging gives us that approximately 61% of total infections were a result of general spread. However, restricting our attention once again to only those infections where classroom transmission was possible, we find that in this case only 26% of infections were a direct result of general transmission.

These figures tell us that classroom transmission was an incredibly significant factor in the spread of the epidemic, with a large percentage of the infections occurring, under

this model, as a direct consequence of classroom spread. This is in line with the research on viral transmission from Section 2.3 and reinforces our decision to include classroom avoidance parameters in the model.

By noting that the proportions of general infections and classroom infections must add to 100%, we can infer that 2.5% of the overall infections occurred due to *both* general and classroom transmission, and restricting our attention once again to just classrooms, this becomes 3%. The idea that an individual can be infected due to both forms of transmission does not make much intuitive sense. Another way to think of this, for example by looking at the whole population, is that on the day of their infection 2.5% of individuals failed to avoid at least one infective both inside and outside their classroom. Therefore they would have been infected on that specific day regardless of the presence of classroom transmission. Conversely, 36.5% of individuals would have avoided infection at their respective time if not for classroom transmission.

6.1.2 Epidemic Simulation

The model is now fully specified. We have coded the following simulation algorithm, recalling that we assume the exposed period E has length 10, the eruption period d has length 3 and for each prodromal period x we sample a value $\hat{x}$ with replacement from the observed values.

Algorithm 3: Classroom Model Simulation Algorithm

Intialise $S_0(0)$, $S_1(0)$, $S_2(0)$,

$I_0(0), I_1(0), I_2(0)$,

$E_0(0), E_1(0), E_2(0)$,

$R_0(0), R_1(0), R_2(0)$,

q, q_c, E, d

Maximum epidemic length T

for *t from 1 to T* **do**

$S_0(t) = \text{Bin}(S_0(t-1), q^{I(t-1)})$, $\hat{I}_0 = S_0(t-1) - S_0(t)$

$S_1(t) = \text{Bin}(S_1(t-1), q^{I_0(t-1)+I_2(t-1)} q_c^{I_1(t-1)})$, $\hat{I}_1 = S_1(t-1) - S_1(t)$

$S_2(t) = \text{Bin}(S_2(t-1), q^{I_0(t-1)+I_1(t-1)} q_c^{I_2(t-1)})$, $\hat{I}_2 = S_2(t-1) - S_2(t)$

for *k from 0 to 2* **do**

for *t' from t to $(t+E-1)$* **do**

$E_k(t') = E_k(t') + \hat{I}_k$

end

for *i from 1 to $\hat{I}_k$* **do**

$\hat{x}_i =$ sample prodromal period length x

for *$\hat{t}$ from $(t+E)$ to $(t+E+\hat{x}_i+d)$* **do**

$I_k(\hat{t}) = I_k(\hat{t}) + 1$

end

for *$\tilde{t}$ from $(t+E+\hat{x}_i+d+1)$ to T* **do**

$R_k(\tilde{t}) = R_k(\tilde{t}) + 1$

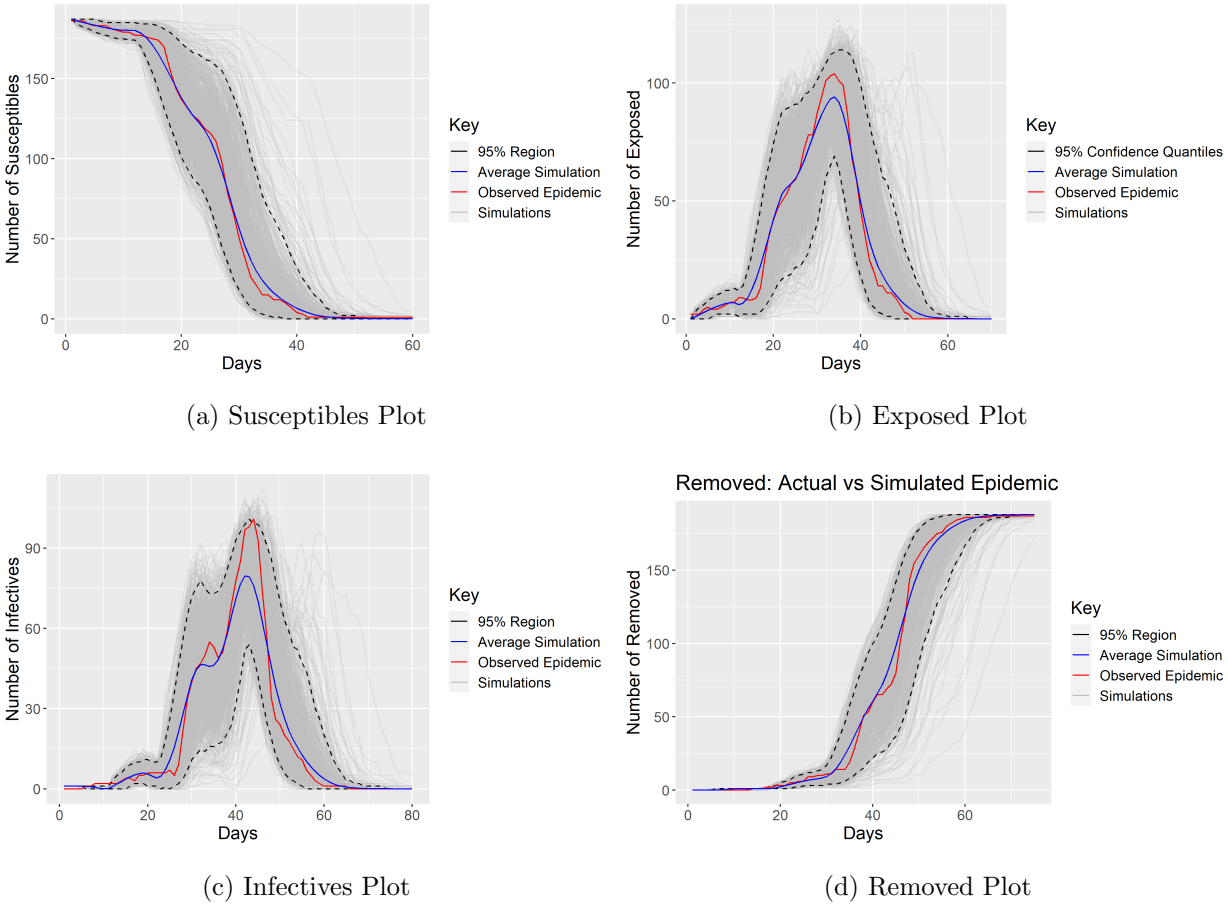
end

end

end

end

Using the above algorithm we can generate epidemics under this classroom model. Note that in the observed epidemic, the first infective on record attended classroom 2 and so we initialise the simulations as such.



(a) Susceptibles Plot

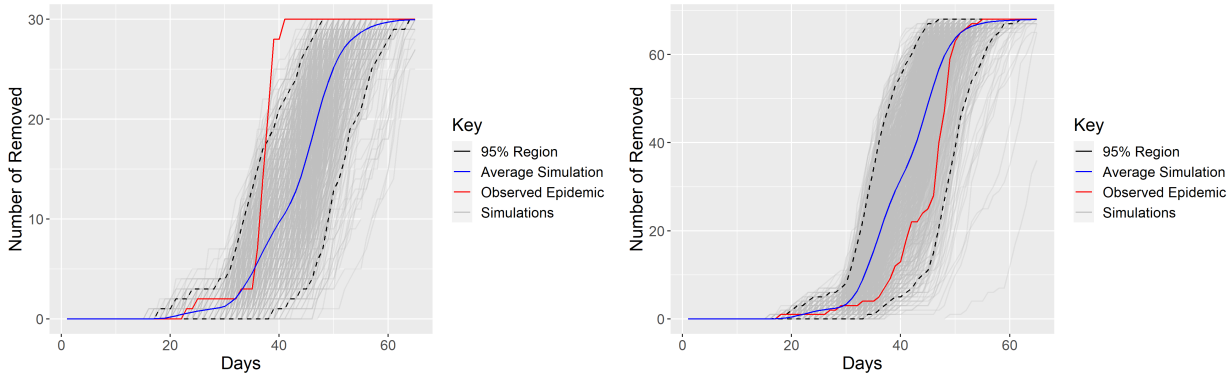
(b) Exposed Plot

(c) Infectives Plot

(d) Removed Plot

Figure 12: Plot showing 1000 simulated epidemics using the initial classroom model with $q = 0.9952665, q_c = 0.9646576$, the average simulation and the observed epidemic

It is difficult to tell if this is an improvement on the simulation results under the base model from Section 5 Figure 10. Despite this, the average simulation does retain a strong fit to the observed epidemic. However, the main area of interest with this model is observing the epidemic inside the physical classrooms. This way, we can see if the model simulations accurately capture the significant classroom transmission we observe.



(a) Class 1 Removed Plot

(b) Class 2 Removed Plot

Figure 13: Plot of the removed statistic $R(t)$ in each classroom showing the 1000 simulated epidemics using the initial classroom model with $q = 0.9952665, q_c = 0.9646576$, the average simulation and the observed epidemic

From these figures which focus on the two classrooms, we can see that this model is performing quite poorly. Figure 13a highlights that the spread of the epidemic in classroom 1 is too slow while conversely, in Figure 13b, the epidemic is progressing too quickly. This highlights a flaw in the model formulation; we assume that the avoidance parameter for each classroom interaction is the same no matter which classroom the interaction is occurring in, namely q_c. In fact, when we focus on the observed epidemic in each figure, we can see that measles spread much quicker in classroom 1 than classroom 2. This suggests that using two parameters, one for each classroom, to account for classroom transmission could produce stronger results.

6.2 Extended Classroom Model

The previous simulation results suggest that using just one parameter q_c to represent interactions in both classrooms is a bad idea due to the varying rate of infection we see in classroom 1 and 2. Thus we introduce two avoidance parameters, q_1 and q_2, in place of q_c, for interactions in classroom 1 and 2 respectively. Then, under the same previous model assumptions, we now have the three-parameter binomial epidemic model,

$$S_0(t+1) \sim \text{Binomial}(S_0(t), q^{I(t)})$$
$$S_1(t+1) \sim \text{Binomial}(S_1(t), q_1^{I_1(t)} q^{I_0(t)+I_2(t)})$$
$$S_2(t+1) \sim \text{Binomial}(S_2(t), q_2^{I_2(t)} q^{I_1(t)+I_2(t)})$$

The unknown quantities here are q, q_1 and q_2. Once again, we can proceed by maximum likelihood estimation . Firstly we have,

$$\begin{aligned} L_0(q, q_1, q_2) &= \prod_t P(S_0(t+1)|S_0(t), I(t)) \\ &= \prod_t (q^{I(t)})^{S_0(t+1)} (1 - q^{I(t)})^{(S_0(t) - S_0(t+1))}, \end{aligned}$$

$$\begin{aligned} L_1(q, q_1, q_2) &= \prod_t P(S_1(t+1)|S_1(t), I_0(t), I_1(t), I_2(t)) \\ &= \prod_t (q_1^{I_1(t)} q^{I_0(t)+I_2(t)})^{S_1(t+1)} (1 - q_1^{I_1(t)} q^{I_0(t)+I_2(t)})^{(S_1(t) - S_1(t+1))}, \end{aligned}$$

$$\begin{aligned} L_2(q, q_1, q_2) &= \prod_t P(S_2(t+1)|S_2(t), I_0(t), I_1(t), I_2(t)) \\ &= \prod_t (q_2^{I_2(t)} q^{I_0(t)+I_1(t)})^{S_2(t+1)} (1 - q_2^{I_2(t)} q^{I_0(t)+I_1(t)})^{(S_2(t) - S_2(t+1))}. \end{aligned}$$

Then, the overall likelihood is as follows,

$$L(q, q_1, q_2) = L_0(q, q_1, q_2) L_1(q, q_1, q_2) L_2(q, q_1, q_2).$$

Now, taking logs, we have

$$\log(L_0(q, q_c)) = \sum_t S_0(t+1) I_0(t) \log(q) + (S_0(t) - S_0(t+1)) \log(1 - q^{I_0(t)}),$$

$$\begin{aligned} \log(L_1(q, q_1, q_2)) = \sum_t & S_1(t+1)[(I_0(t) + I_2(t)) \log(q) + I_1(t) \log(q_1)] \\ & + (S_1(t) - S_1(t+1)) \log(1 - q_1^{I_1(t)} q^{I_0(t)+I_2(t)}), \end{aligned}$$

$$\begin{aligned} \log(L_2(q, q_1, q_2)) = \sum_t & S_2(t+1)[(I_0(t) + I_1(t)) \log(q) + I_2(t) \log(q_2)] \\ & + (S_2(t) - S_2(t+1)) \log(1 - q_2^{I_2(t)} q^{I_0(t)+I_1(t)}). \end{aligned}$$

Finally, the overall log-likelihood is the following,

$$\log(L(q, q_1, q_2)) = \log(L_0(q, q_1, q_2)) + \log(L_1(q, q_1, q_2)) + \log(L_2(q, q_1, q_2)).$$

Maximising with the BFGS algorithm gives the estimates: $\hat{q} = 0.9951997, \hat{q_1} = 0.8285551$ and $\hat{q_2} = 0.9765360$. These results are what we might expect, i.e we have $q_1 < q_2 < q$, which indicates an individual would be more likely to be infected in classroom 1 than classroom 2 than in the general population. This reflects our observations from the previous section. We also note that this finding makes perfect sense when we consider the age groups of each classroom; it is very likely that those individuals in classroom 1, who were between the ages of 7 and 10, would be in more frequent physical contact with each other in comparison to those in classroom 2 who were aged between 11 and 15.

Using the approximate Hessian matrix and the asymptotic normality assumption of the MLE we can produce an approximate 95% confidence interval for each estimate.

$$\begin{aligned}
&\hat{q} \pm 1.96\frac{1}{\sqrt{(\sigma_q^2)}} \\
&= 0.9951997 \pm 1.96\frac{1}{\sqrt{(4737087.254)}} \\
&= 0.9951997 \pm 0.0009005345,
\end{aligned}$$

$$\begin{aligned}
&\hat{q_1} \pm 1.96\frac{1}{\sqrt{(\sigma_{q_1}^2)}} \\
&= 0.8285551 \pm 1.96\frac{1}{\sqrt{(1038.70)}} \\
&= 0.8285551 \pm 0.06081507,
\end{aligned}$$

$$\begin{aligned}
&\hat{q_2} \pm 1.96\frac{1}{\sqrt{(\sigma_{q_2}^2)}} \\
&= 0.976536 \pm 1.96\frac{1}{\sqrt{(57506.48)}} \\
&= 0.976536 \pm 0.008173304.
\end{aligned}$$

The confidence interval for q_1 is a factor of 10 wider than the interval for q_2 which is itself a factor of 10 larger than the interval for q. This is unsurprising as we have an increasing

number of observations for each estimate; for classroom 1 we have just 30 individuals. With such a small amount of data, we are more and more uncertain of our estimates and the normality assumption becomes less and less valid.

6.2.1 Transmission Probabilities

Now that we have introduced separate avoidance parameters for each classroom, it is of interest to recalculate the transmission probabilities in order to get a better sense of the importance of classroom transmission in the spread of the epidemic. To do so, we use the same approach as in Section 6.1.1.

Consider a susceptible individual in classroom 1, the probability that this individual is infected at time $t+1$ as a direct result of classroom transmission is the following,

$$\frac{P(General\ Avoidance)P(Classroom\ Non\text{-}Avoidance)}{P(Infection)} = \frac{q^{I_0(t)+I_2(t)}(1-q_1^{I_1(t)})}{1-q^{I_0(t)+I_2(t)}q_1^{I_1(t)}}.$$

Similarly, for an individual in classroom 2 we have,

$$\frac{P(General\ Avoidance)P(Classroom\ Non\text{-}Avoidance)}{P(Infection)} = \frac{q^{I_0(t)+I_1(t)}(1-q_2^{I_2(t)})}{1-q^{I_0(t)+I_1(t)}q_2^{I_2(t)}}.$$

For a class 0 individual, the probability they are infected due to classroom transmission is zero as the model assumes they do not undergo the same classroom interactions as class 1 and 2 individuals.

Example: Consider the following population, $S_1(0) = 1, I_0(0) = 1, I_1(0) = 2, I_2(0) = 1$. We are interested in finding the probability that, given an infection occurred, it was a sole result of classroom transmission.

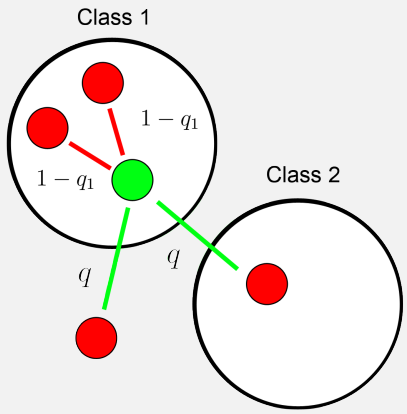


Figure 14: Visualisation of one of the possible events where a susceptible individual is infected solely due to classroom transmission

The above figure represents one possible occurrence that satisfies our requirements; the suscpetible individual fails to avoid *both* infectives inside their classroom but *does* avoid the infective in class 2 and the general population. There are two further possibilities where the susceptible avoids one of these infectives but not the other and vice versa. In total, the three possibilities make up the state space of our probability of interest, which is calculated below.

$$\frac{P(General\ Avoidance)P(Classroom\ Non\text{-}Avoidance)}{P(Infection)} = \frac{q^2(1-q_1^2)}{1-q^2q_1^2}.$$

where we have that,

$$\begin{aligned} P(Classroom\ Non\text{-}Avoidance) &= 1 - P(Classroom\ Avoidance) \\ &= 1 - q_1^2, \end{aligned}$$

and,

$$\begin{aligned} P(Infection) &= 1 - P(Avoids\ Infection) \\ &= 1 - P(General\ Avoidance \cap Classroom\ Avoidance) \\ &= 1 - q^2q_1^2. \end{aligned}$$

Now, as in Section 6.1.1, we can calculate these probabilities for each infection in the observed epidemic and then average to calculate the overall proportion of infections that occurred as a direct result of classroom transmission. We find that, under this model, approximately 37.5% of the total infections in the Hagelloch epidemic happened as a result classroom transmission. This is very similar to the previous model. However, when we restrict our attention to just those infections where classroom transmission was possible, we now find that 91% and 62% of individuals in classroom 1 and 2 respectively were infected solely due to transmission inside their classes[5].

We can see a large difference in the infection source between classroom 1 and classroom 2 individuals. In particular, for classroom 1, we see a significant rise from the results of the previous model. This reflects the observed epidemic much more closely. We have seen that once an individual in classroom 1 was infected, the epidemic then spread through the remainder of the class in a few days. This suggests that classroom transmission was very much responsible for the infection of classroom 1 individuals. In comparison, the virus took much longer to work through the classroom 2 population, this indicates that the virus probably did not spread as much via classroom 2. We see this reflected in the transmission probabilities where a significant percentage of classroom 2 individuals were infected due to general transmission, not classroom transmission.

6.2.2 Epidemic Simulation

The model is now fully specified and we can use a very similar algorithm to what we saw in Section 6.1.2 to produce simulations of the epidemic.

[5]We can also once again calculate the probability that an infection occurred as a singular result of general transmission. We get that approximately 61% of total infections were a result of general spread; comparatively this represented only 7% and 35% of infections within classroom 1 and 2 respectively

Algorithm 4: Extended Classroom Model Simulation Algorithm

Intialise $S_0(0)$, $S_1(0)$, $S_2(0)$,

$I_0(0), I_1(0), I_2(0)$,

$E_0(0), E_1(0), E_2(0)$,

$R_0(0), R_1(0), R_2(0)$,

q, q_1, q_2, E, d,

Maximum epidemic length T

for *t from 1 to T* **do**

$S_0(t) = \text{Bin}(S_0(t-1), q^{I(t-1)})$, $\hat{I}_0 = S_0(t-1) - S_0(t)$

$S_1(t) = \text{Bin}(S_1(t-1), q^{I_0(t-1)+I_2(t-1)} q_1^{I_1(t-1)})$, $\hat{I}_1 = S_1(t-1) - S_1(t)$

$S_2(t) = \text{Bin}(S_2(t-1), q^{I_0(t-1)+I_1(t-1)} q_2^{I_2(t-1)})$, $\hat{I}_2 = S_2(t-1) - S_2(t)$

for *k from 0 to 2* **do**

for *t' from t to $(t+E-1)$* **do**

$E_k(t') = E_k(t') + \hat{I}_k$

end

for *i from 1 to $\hat{I}_k$* **do**

$\hat{x}_i =$ sample prodromal period length x

for *$\hat{t}$ from $(t+E)$ to $(t+E+\hat{x}_i+d)$* **do**

$I_k(\hat{t}) = I_k(\hat{t}) + 1$

end

for *$\tilde{t}$ from $(t+E+\hat{x}_i+d+1)$ to T* **do**

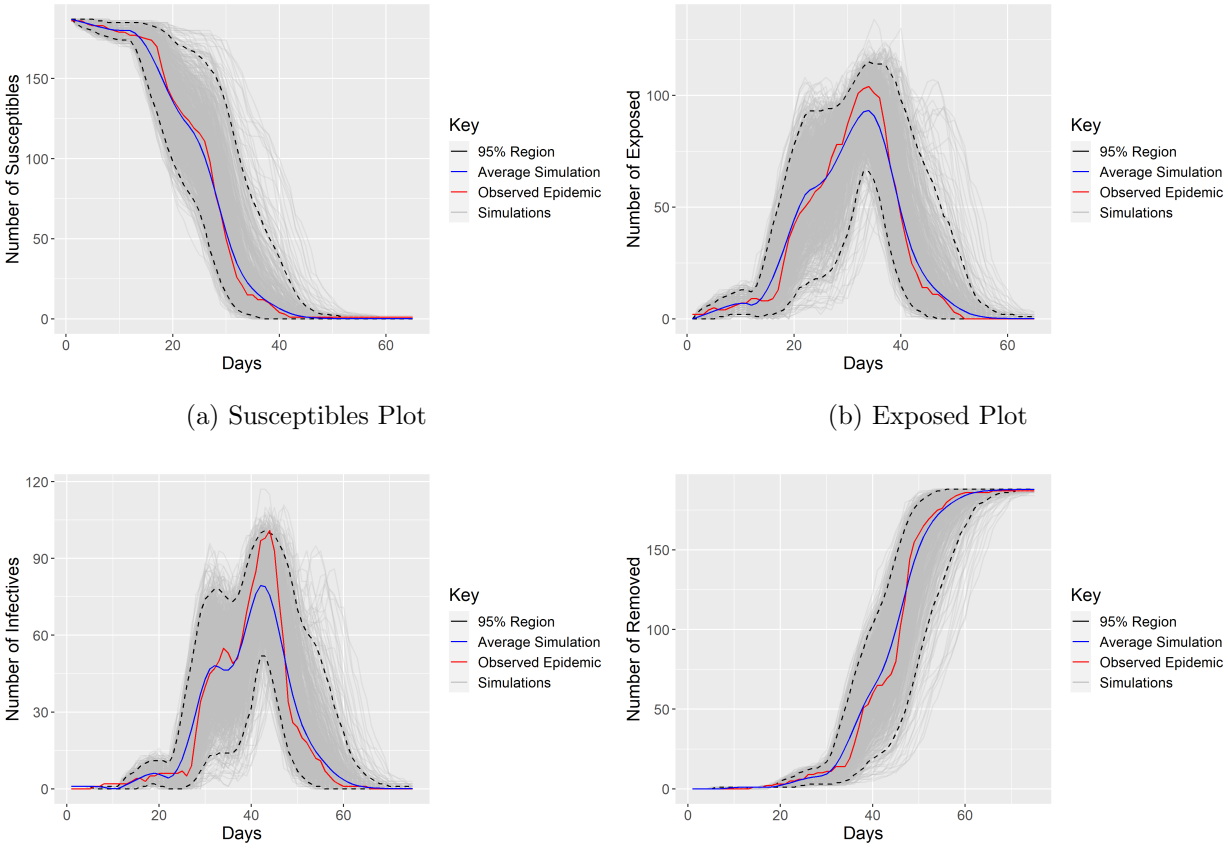
$R_k(\tilde{t}) = R_k(\tilde{t}) + 1$

end

end

end

end



(a) Susceptibles Plot

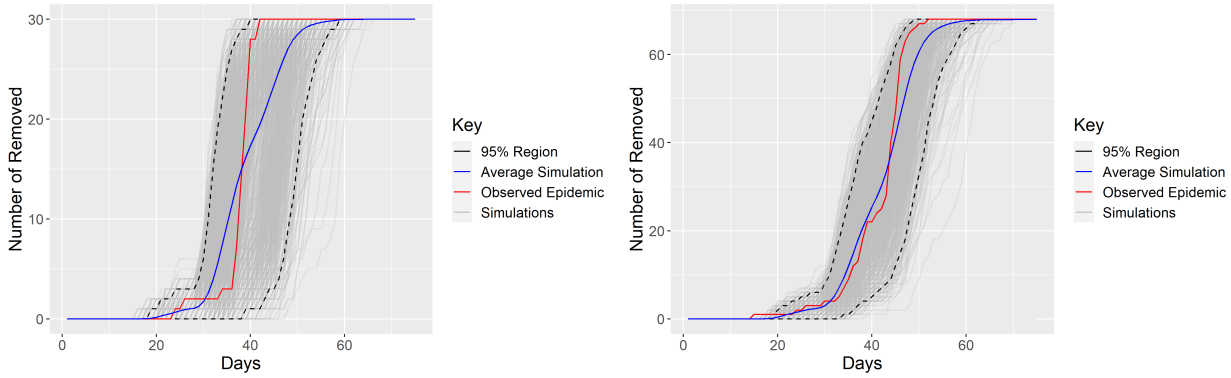
(b) Exposed Plot

(c) Infectives Plot

(d) Removed Plot

Figure 15: Plot showing 1000 simulated epidemics using the extended classroom model with $q = 0.9951997, q_1 = 0.8285551, q_2 = 0.976536$, the average simulation and the observed epidemic

There does not seem to have been a significant improvement in the average fit to the overall observed epidemic in comparison to the previous model. Despite this, the simulations do appear to be more consistent; fewer simulations stray very far from the mean than in the previous simulation. This is likely due to the decreased avoidance probability in classroom 1 ensuring that the epidemic progresses and doesn't spend a lot of time gaining steam. However, once again we are more interested in seeing what is going on inside the classrooms.



(a) Class 1 Removed Plot

(b) Class 2 Removed Plot

Figure 16: Plot of the removed statistic $R(t)$ of each classroom showing the 1000 simulated epidemics using the extended classroom model with $q = 0.9951997, q_1 = 0.8285551, q_2 = 0.976536$, the average simulation and the observed epidemic

In comparison to the results from the first classroom model, this is a huge improvement. The observed epidemic in each case is firmly within the 95% region of the simulations with the average simulation much more closely fitted. We can see however, particularly with classroom 1, that the average simulation is not perfect. In Figure 16a it is clear that the simulated epidemic is not accurately capturing the infection rate within the class 1 population. Similarly, in Figure 16b we can see that the simulated epidemic progresses at the correct pace at first, but does not accelerate in the same way that the observed epidemic did. One possible explanation for this is that we are missing a transmission pathway; in particular we have yet to introduce household transmission to the epidemic model. Introducing this factor could result in simulations that more accurately represent the observed epidemic and in particular, the classroom sub-populations.

7 Household Model

We have seen evidence from the exploratory analysis in Section 4.2 that household transmission was likely a significant factor in the spread of the epidemic. We have also observed from the extended classroom model simulations that the epidemic is not spreading at the pace that we would expect and that introducing another transmission vector could be a solution to this. Before we do this, it is worth exploring a model with just household and

general transmission to isolate the impact of household spread.

Instead of separating the population into the fifty six households and working with a susceptible population statistic for each, we instead work with the individuals separately. This decision was taken for ease of writing and is an equivalent approach. Consider each individual, $i = 1, 2, ..., 188$, then at time t we define the following function,

$$\hat{S}_i(t) = \begin{cases} 1, & \text{if individual } i \text{ is susceptible at time } t \\ 0, & \text{otherwise} \end{cases}.$$

Note that $S(t) = \hat{S}_1(t) + \hat{S}_2(t) + ... + \hat{S}_{188}(t)$. For each individual i, we also define $I_{H_i}(t)$ to be the number of infectives in the household of individual i at time t. Now, let q and q_H represent the avoidance parameter for interaction outside and inside a household, respectively. Then, under the previous model assumptions and the addition of q_H, we have that,

$$\hat{S}_i(t+1) \sim \text{Bernoulli}(q_H^{I_{H_i}(t)} q^{I(t)-I_{H_i}(t)}),$$

that is,

$$\hat{S}_i(t+1) = \begin{cases} 1, & \text{with probability } \; q_H^{I_{H_i}(t)} q^{I(t)-I_{H_i}(t)} \\ 0, & \text{with probability } \; 1 - q_H^{I_{H_i}(t)} q^{I(t)-I_{H_i}(t)}. \end{cases}$$

Example: At time $t = 0$, consider a population with one susceptible ($i = 1$), and three infectives, one inside and two outside the susceptible's household. Then, $\hat{S}_1(0) = 1$, $I(t) = 3$, $I_{H_1}(0) = 1$ and we have that,

$$\hat{S}_1(1) \sim \text{Bernoulli}(q_H^{I_{H_i}(0)} q^{I(0)-I_{H_i}(0)}) = \text{Bernoulli}(q_H q^2),$$

that is,

$$P(\hat{S}_1(1) = 1) = q_H q^2,$$
$$P(\hat{S}_1(1) = 0) = 1 - q_H q^2.$$

The only unknown quantities are q and q_H, we can estimate these once again by maximum likelihood estimation. Before we do this, note that an individual can avoid infection many times but only be infected once. Therefore the model specified above for $\hat{S}(t)$ is only valid for the time up until and including the day the individual is infected. Thus, we let t_i be the time at which individual i is infected, and the likelihood function is then as follows,

$$L(q, q_H) = \prod_i \prod_{t \leq t_i} (q_H^{I_{H_i}(t)} q^{I(t)-I_{H_i}(t)})^{\hat{S}_i(t+1)} (1 - q_H^{I_{H_i}(t)} q^{I(t)-I_{H_i}(t)})^{(1-\hat{S}_i(t+1))}.$$

Then the log-likelihood has the form,

$$\begin{aligned}
\log(L(q, q_H)) &= \sum_i \sum_{t \leq t_i} \hat{S}_i(t+1) \log(q_H^{I_{H_i}(t)} q^{I(t)-I_{H_i}(t)}) + (1 - \hat{S}_i(t+1)) \log(1 - q_H^{I_{H_i}(t)} q^{I(t)-I_{H_i}(t)}) \\
&= \sum_i \sum_{t \leq t_i} \hat{S}_i(t+1)[I_{H_i}(t) \log(q_H) + (I(t) - I_{H_i}(t)) \log(q)] \\
&\qquad + (1 - \hat{S}_i(t+1)) \log(1 - q_H^{I_{H_i}(t)} q^{I(t)-I_{H_i}(t)}).
\end{aligned}$$

We can optimise this using the BFGS algorithm to get the following estimates and approximate 95% confidence intervals,

$$\begin{aligned}
&\hat{q} \pm 1.96 \frac{1}{\sqrt{(\sigma_q^2)}} \\
&= 0.9938498 \pm 1.96 \frac{1}{\sqrt{(4103170.09)}} \\
&= 0.9938498 \pm 0.000967601
\end{aligned}$$

$$\begin{aligned}
&\hat{q}_H \pm 1.96 \frac{1}{\sqrt{(\sigma_{q_H}^2)}} \\
&= 0.8581391 \pm 1.96 \frac{1}{\sqrt{(899.3447)}} \\
&= 0.8581391 \pm 0.06535713
\end{aligned}$$

These estimates reflect our research from Section 2.3. With $q_H < q$, you are more likely to be infected inside a household than outside. We also note that the confidence interval for q_H is a factor of 100 wider than the interval for q. This is due to the comparatively small sample size of infections where household transmission was possible in the observed epidemic. With such small sample sizes, the asymptotic normality assumption of the

MLE is also seriously called into question and the accuracy of the 95% confidence interval is doubtful.

7.1 Transmission Probabilities

We are interested in knowing how significant household transmission was as a pathway to infection in the observed epidemic. We use a very similar argument to that from Section 6.1.1 to derive the general expression for the probability that an indivudal i was infected at time $t+1$ as a direct result of household transmission,

$$\begin{aligned} P(Household\ Infection|Infection) &= \frac{P(General\ Avoidance)P(Household\ Non\text{-}Avoidance)}{P(Infection)} \\ &= \frac{q^{I(t)-I_{H_i}(t)}(1-q_H^{I_{H_i}(t)})}{1-q^{I(t)-I_{H_i}(t)}q_H^{I_{H_i}(t)}}. \end{aligned}$$

Example: Consider a population consisting of a single susceptible ($i = 1$) in a household and five infectives; two inside the household and three outside. That is, $\hat{S}_1(0) = 1, I_{H_1}(0) = 2, I(0) = 5$. We are interested in finding the probability that, given an infection occurred, it was a sole result of household transmission.

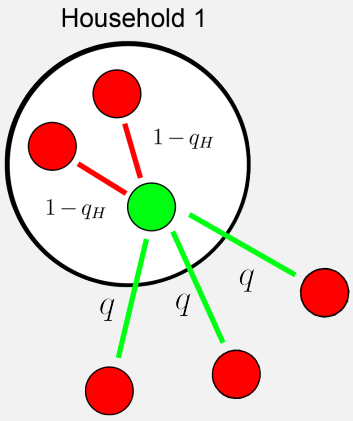


Figure 17: Visualisation of one of the possible events where the susceptible individual is infected solely due to household transmission

The above figure represents one occurrence in the state space of possible events where the individual is infected as a direct result of household transmission. Then,

$$\frac{P(General\ Avoidance)P(Household\ Non\text{-}Avoidance)}{P(Infection)} = \frac{q^3(1-q_H^2)}{1-q^3q_H^2}.$$

where we have that,

$$\begin{aligned} P(Household\ Non\text{-}Avoidance) &= 1 - P(Household\ Avoidance) \\ &= 1 - q_H^2, \end{aligned}$$

and,

$$\begin{aligned} P(Infection) &= 1 - P(Avoids\ Infection) \\ &= 1 - P(General\ Avoidance \cap Household\ Avoidance) \\ &= 1 - q^3q_H^2. \end{aligned}$$

Now, we can calculate this probability for each infection in the observed epidemic and then average to calculate the overall proportion of infections that occurred as a direct result of household transmission. Doing so, we find that approximately 13% of the total infections were a result of just household spread. Conversely, 85% of infections under this model occurred due to general transmission, and therefore we can infer that 2% were a result of both. When compared to the percentages we saw in Section 6.1.1 and 6.2.1, the proportion of direct household infections seems quite low, however when you consider that the average household was inhabited by only 4 people, and the overall population consisted of 188 individuals, the chance that any one of these individuals would have been infected by a member of their own household would be relatively small.

Restricting our attention to just those infections where household transmission was possible, we find that approximately 48% of these occurred due to household transmission. Comparing this number to the results from performing a similar calculation in the exploratory analysis in Section 4.2, we find that this model appears to actually be *over* representing the number of household infections. This could be due to the large uncertainty we have for our value of $\hat{q}_H$; it may be that a larger value of $\hat{q}_H$ would be more appropriate here.

7.2 Epidemic Simulation

We can now state an algorithm that will allow us to simulate epidemics under this fully specified household model.

Algorithm 5: Household Model Simulation Algorithm

Intialise $\hat{S}_1(0), \hat{S}_2(0), ..., \hat{S}_{188}(0), E(0), I(0), I_{H_1}(0), I_{H_2}(0), ..., I_{H_{188}}(0), R(0)$,

q, q_H, E, d,

Maximum epidemic length T

for *t from 1 to T* **do**

  for *i from 1 to 188* **do**

    if $\hat{S}_i(t) = 1$ **then**

      $\hat{S}_i(t) = \text{Bernoulli}(q_H^{I_{H_i}(t-1)} q^{I(t-1)-I_{H_i}(t-1)})$

      if $\hat{S}_i(t) = 0$ **then**

        for *t' from t to $(t + E - 1)$* **do**

          $E(t') = E(t') + 1$

        end

        $\hat{x}_i$ = sample prodromal period length x

        for *$\hat{t}$ from $(t + E)$ to $(t + E + \hat{x}_i + d)$* **do**

          $I(\hat{t}) = I(\hat{t}) + 1$

          $I_{H_i}(\hat{t}) = I(\hat{t}) + 1$

        end

        for *$\tilde{t}$ from $(t + E + \hat{x}_i + d + 1)$ to T* **do**

          $R(\tilde{t}) = R(\tilde{t}) + 1$

        end

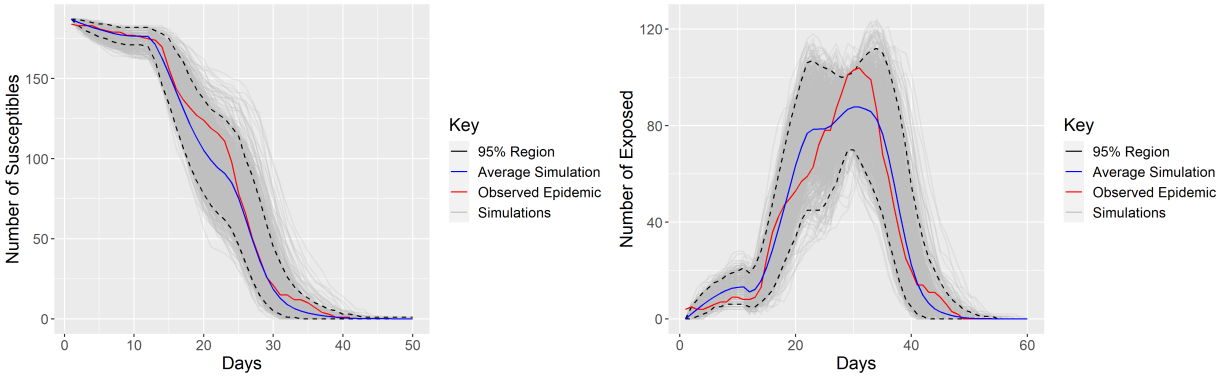
      end

    end

  end

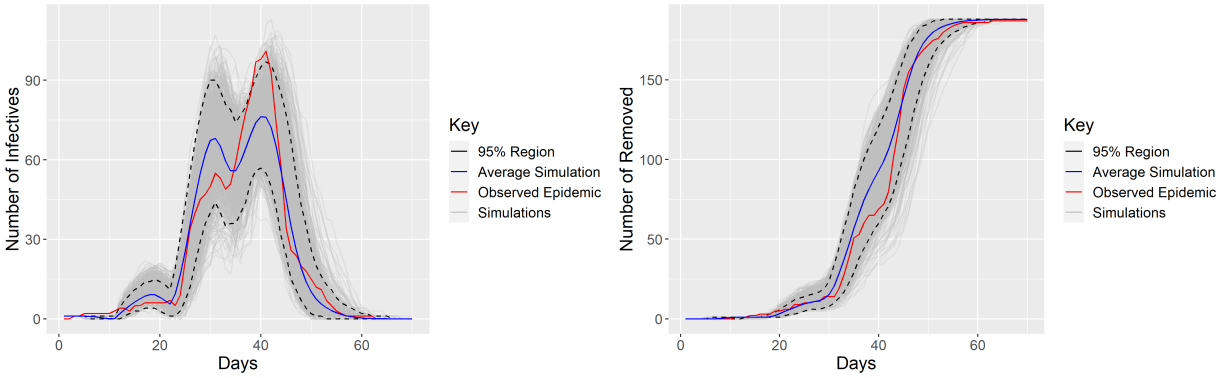
end

In the observed epidemic, the first infective individual was a part of the fifth household which had eight inhabitants. Using these initial values we produced 1000 simulations of the Hagelloch measles epidemic.



(a) Susceptibles Plot

(b) Exposed Plot



(c) Infectives Plot

(d) Removed Plot

Figure 18: Plot showing 1000 simulated epidemics using the household model with $q = 0.9938498, q_H = 0.8581391$, the average simulation and the observed epidemic

This model has produced simulations that have an overall worse fit to the observed epidemic than the classroom and base models. We can see that the average simulation initially follows the observed epidemic closely, similarly near the end of the epidemic we also have a strong fit. However, under this model, the middle phase of the epidemic when most infections are occurring is not being accurately reflected. Specifically, there are too many infections too quickly; in Figure 18c we can see the two waves discussed in the exploratory analysis, however the first wave is much too large and then, due to having a capped number of individuals in the population, the second wave is too small.

These findings could suggest that our MLE for q_H is too small, and as a result the epidemic is spreading too quickly. Our relative uncertainty of the value of q_H is a factor here; due to small sample sizes estimating q_H accurately is difficult. Alternatively, it could be the

case that including household transmission in our binomial model is itself not an accurate representation of the infection dynamics in the observed epidemic. Despite this, we can proceed with introducing household transmission to the extended classroom model to see if our results improve.

8 Classroom-Household Model

Now that we have analysed the effect of household spreading alone, we can reintroduce classroom transmission to the epidemic model to produce simulations that best represent the transmission pathways we observed from the exploratory analysis.

Recall from Section 6.2 that the avoidance parameters q_1 and q_2 represent interaction in classroom 1 and 2 respectively. Now, rather than separating the population into three classes, we work with individuals directly, similar to the approach taken in Section 7. Therefore, for each individual $i = 1, 2, ..., 188$, we define $I_{C_i}(t)$ to be the number of infectives in individual i's classroom. Also, recall that from the household model we have the avoidance parameter q_H, and we let $I_{Hi}(t)$ be the number of infectives in individual i's household.

Now, note that a single infective can be in both an individual's household *and* classroom. Thus, to avoid the issue of double counting, we define the population statistic $I_{Hi} \cap I_{C_i}(t)$ to be the number of infectives in individual i's household *and* classroom.

Then, under the same model assumptions from previously, for each individual $i = 1, 2, ..., 188$, we have that,

$$\hat{S}_i(t+1) \sim \begin{cases} \text{Bernoulli}(q_H^{I_{H_i}(t)} q^{I(t)-I_{H_i}(t)}), & \text{if individual } i \text{ is in classroom 0} \\ \text{Bernoulli}(q_H^{I_{H_i}(t)} q_1^{I_{C_i}(t)} q^{I(t)-I_{H_i}(t)-I_{C_i}(t)+I_{H_i}\cap I_{C_i}(t)}), & \text{if individual } i \text{ is in classroom 1} \\ \text{Bernoulli}(q_H^{I_{H_i}(t)} q_2^{I_{C_i}(t)} q^{I(t)-I_{H_i}(t)-I_{C_i}(t)+I_{H_i}\cap I_{C_i}(t)}), & \text{if individual } i \text{ is in classroom 2} \end{cases}.$$

Example: At time $t = 0$, consider a population with one susceptible individual ($i = 1$), who is in classroom 1 and household 1. Now, let there be two infectives in the individual's household and two in their classroom. Then, we let one of these infectives be in both their household *and* classroom. Finally, let there be one remaining infective in the general population. That is, $I_{C_i}(0) = 2$, $I_{H_i}(0) = 2$, $I_{H_i} \cap I_{C_i}(0) = 1$, and $I(0) = 4$.

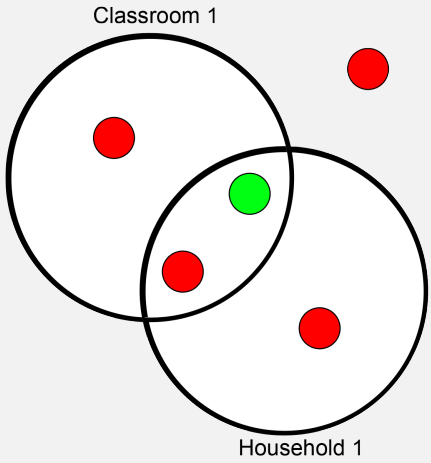


Figure 19: Visualisation of this population

Then, the model for this susceptible individual is as follows

$$\begin{aligned}\hat{S}_1(1) &\sim \text{Bernoulli}(q_H^{I_{H_i}(0)} q_1^{I_{C_i}(0)} q^{I(0)-I_{H_i}(0)-I_{C_i}(0)+I_{H_i}\cap I_{C_i}(0)}) \\ &= \text{Bernoulli}(q_H^2 q_1^2 q^{4-2-2+1}) \\ &= \text{Bernoulli}(q_H^2 q_1^2 q)\end{aligned}$$

that is,

$$\begin{aligned}P(\hat{S}_1(1) = 1) &= q_H^2 q_1^2 q, \\ P(\hat{S}_1(1) = 0) &= 1 - q_H^2 q_1^2 q.\end{aligned}$$

Note that one way to think of the above example of this model is as follows,

$$\hat{S}_1(1) = \text{Bernoulli}(q_H^1 q_1^1 (q_H q_1)^1 q^1).$$

Here we have one susceptible-infective interaction governed by q_H, one by q_1, one by q and one by *both* q_H and q_1 i.e. $q_H q_1$, for a total of four interactions; one for each infective. In

this interpretation, when an infective is in a susceptible's household *and* classroom, $q_H q_1$ acts as an avoidance parameter that represents a singular susceptible-infective interaction. We could introduce avoidance parameters to govern these interactions directly. To do so, we would need two new parameters, one to replace $q_H q_1$ and another for $q_H q_2$. However, this would introduce further variability to the simulations for little potential gain as these interactions are relatively rare and only occur on a small scale throughout the epidemic.

Alternatively, we can think of the case when an infective is in a susceptible's household and classroom as generating two unrelated susceptible-infective interactions. One governed by q_H and then another by q_1 or q_2, depending on the classroom. This makes more intuitive sense as we can imagine that the susceptible individual will interact with this infective in two separate ways throughout a typical day; at school and then again at home. For the example above, this would result in a total of five susceptible-infective interactions from four infectives. This, and the above interpretation are mathematically equivalent when we calculate $\hat{S}_1(1)$ due to our Reed-Frost assumption of independence in the susceptible-infective interactions. However, these interpretations do differ when we come to calculate transmission probabilities in Section 8.1. The second is more consistent with previous sections. Therefore we choose to think of these susceptible-infective interactions in the second way, i.e. as two different interactions.

Now, we need to estimate q, q_H, q_1 and q_2. Like with the previous models, we proceed by maximum likelihood estimation. Once again, we let t_i be the time at which individual i is infected. Then, for individuals in classroom 1, we have the likelihood,

$$L_0(q, q_H, q_1, q_2) = \prod_i \prod_{t \leq t_i} (q_H^{I_{H_i}(t)} q^{I(t) - I_{H_i}(t)})^{\hat{S}_i(t+1)} (1 - q_H^{I_{H_i}(t)} q^{I(t) - I_{H_i}(t)})^{(1 - \hat{S}_i(t+1))}.$$

For classroom 1 individuals,

$$L_1(q, q_H, q_1, q_2) = \prod_i \prod_{t \leq t_i} (q_H^{I_{H_i}(t)} q_1^{I_{C_i}(t)} q^{I(t) - I_{H_i}(t) - I_{C_i}(t) + I_{H_i} \cap I_{C_i}(t)})^{\hat{S}_i(t+1)}$$
$$(1 - q_H^{I_{H_i}(t)} q_1^{I_{C_i}(t)} q^{I(t) - I_{H_i}(t) - I_{C_i}(t) + I_{H_i} \cap I_{C_i}(t)})^{(1 - \hat{S}_i(t+1))},$$

and classroom two individuals,

$$L_2(q, q_H, q_1, q_2) = \prod_i \prod_{t \le t_i} (q_H^{I_{H_i}(t)} q_2^{I_{C_i}(t)} q^{I(t) - I_{H_i}(t) - I_{C_i}(t) + I_{H_i} \cap I_{C_i}(t)})^{\hat{S}_i(t+1)}$$
$$(1 - q_H^{I_{H_i}(t)} q_2^{I_{C_i}(t)} q^{I(t) - I_{H_i}(t) - I_{C_i}(t) + I_{H_i} \cap I_{C_i}(t)})^{(1 - \hat{S}_i(t+1))}.$$

We define the overall likelihood function to be,

$$L(q, q_H, q_1, q_2) = L_0(q, q_H, q_1, q_2) L_1(q, q_H, q_1, q_2) L_2(q, q_H, q_1, q_2).$$

Taking logs we have,

$$\log(L_0(q, q_H, q_1, q_2)) = \sum_i \sum_{t \le t_i} \hat{S}_i(t+1) \log(q_H^{I_{H_i}(t)} q^{I(t) - I_{H_i}(t)})$$
$$+ (1 - \hat{S}_i(t+1)) \log(1 - q_H^{I_{H_i}(t)} q^{I(t) - I_{H_i}(t)})$$

$$= \sum_i \sum_{t \le t_i} \hat{S}_i(t+1) [I_{H_i}(t) \log(q_H) + (I(t) - I_{H_i}(t)) \log(q)]$$
$$+ (1 - \hat{S}_i(t+1)) \log(1 - q_H^{I_{H_i}(t)} q^{I(t) - I_{H_i}(t)}),$$

$$\log(L_1(q, q_H, q_1, q_2)) = \sum_i \sum_{t \le t_i} \hat{S}_i(t+1) \log(q_H^{I_{H_i}(t)} q_1^{I_{C_i}(t)} q^{I(t) - I_{H_i}(t) - I_{C_i}(t) + I_{H_i} \cap I_{C_i}(t)})$$
$$+ (1 - \hat{S}_i(t+1)) \log(1 - q_H^{I_{H_i}(t)} q_1^{I_{C_i}(t)} q^{I(t) - I_{H_i}(t) - I_{C_i}(t) + I_{H_i} \cap I_{C_i}(t)})$$

$$= \sum_i \sum_{t \le t_i} \hat{S}_i(t+1) [I_{H_i(t)} \log(q_H) + I_{C_i}(t) \log(q_1)$$
$$+ (I(t) - I_{H_i}(t) - I_{C_i}(t) + I_{H_i} \cap I_{C_i}(t)) \log(q)]$$
$$+ (1 - \hat{S}_i(t+1)) \log(1 - q_H^{I_{H_i}(t)} q_1^{I_{C_i}(t)} q^{I(t) - I_{H_i}(t) - I_{C_i}(t) + I_{H_i} \cap I_{C_i}(t)}),$$

$$\log(L_2(q, q_H, q_1, q_2)) = \sum_i \sum_{t \le t_i} \hat{S}_i(t+1) \log(q_H^{I_{H_i}(t)} q_2^{I_{C_i}(t)} q^{I(t) - I_{H_i}(t) - I_{C_i}(t) + I_{H_i} \cap I_{C_i}(t)})$$
$$+ (1 - \hat{S}_i(t+1)) \log(1 - q_H^{I_{H_i}(t)} q_2^{I_{C_i}(t)} q^{I(t) - I_{H_i}(t) - I_{C_i}(t) + I_{H_i} \cap I_{C_i}(t)})$$

$$= \sum_i \sum_{t \le t_i} \hat{S}_i(t+1) [I_{H_i(t)} \log(q_H) + I_{C_i}(t) \log(q_2)$$
$$+ (I(t) - I_{H_i}(t) - I_{C_i}(t) + I_{H_i} \cap I_{C_i}(t)) \log(q)]$$
$$+ (1 - \hat{S}_i(t+1)) \log(1 - q_H^{I_{H_i}(t)} q_2^{I_{C_i}(t)} q^{I(t) - I_{H_i}(t) - I_{C_i}(t) + I_{H_i} \cap I_{C_i}(t)}).$$

Finally, the overall log-likelihood is the following,

$$\log(L(q, q_H, q_1, q_2)) = \log(L_0(q, q_H, q_1, q_2)) + \log(L_1(q, q_H, q_1, q_2)) + \log(L_2(q, q_H, q_1, q_2)).$$

We can then optimise using the BFGS algorithm to get the following estimate and approximate 95% confidence intervals,

$$\begin{aligned}
&\hat{q} \pm 1.96\frac{1}{\sqrt{(\sigma_q^2)}} \\
&= 0.9961663 \pm 1.96\frac{1}{\sqrt{(5477248.584)}} \\
&= 0.9961663 \pm 0.0008374808
\end{aligned}$$

$$\begin{aligned}
&\hat{q}_H \pm 1.96\frac{1}{\sqrt{(\sigma_{q_H}^2)}} \\
&= 0.8475385 \pm 1.96\frac{1}{\sqrt{(894.836747)}} \\
&= 0.8475385 \pm 0.06552155
\end{aligned}$$

$$\begin{aligned}
&\hat{q_1} \pm 1.96\frac{1}{\sqrt{(\sigma_{q_1}^2)}} \\
&= 0.8319394 \pm 1.96\frac{1}{\sqrt{(1058.568269)}} \\
&= 0.8319394 \pm 0.06024162
\end{aligned}$$

$$\begin{aligned}
&\hat{q_2} \pm 1.96\frac{1}{\sqrt{(\sigma_{q_2}^2)}} \\
&= 0.9788626 \pm 1.96\frac{1}{\sqrt{(62269.666)}} \\
&= 0.9788626 \pm 0.007854487
\end{aligned}$$

We have that $q_1 < q_H < q_2 < q$; this indicates that an individual attending classroom 1 is more likely to be infected in their classroom than their household, while for individuals attending classroom 2 the opposite is the case. However, we do note that the confidence intervals for q_1 and q_H are quite wide. Therefore we are too uncertain to make this observation with high confidence.

8.1 Transmission Probabilities

We have a model that allows for measles to spread through households *and* classrooms, thus we are interested in knowing how significant these transmission pathways are in the observed epidemic now that they are both possibilities. Say we wanted to know the probability that a susceptible individual was infected solely due to classroom transmission, that is,

$$P(Classroom\ Infection|Infection).$$

Then, by using the same arguments from Section 6.1.1, this time extended to three possible sources of infection, we find that our probability of interest is equivalent to,

$$\frac{P(General\ Avoidance)P(Household\ Avoidance)P(Classroom\ Non\text{-}Avoidance)}{P(Infection)}.$$

If we now consider an individual i who attends classroom 1, the probability that this individual is infected at time $t+1$ as direct result of classroom transmission is the following,

$$\frac{P(General\ Avoidance)P(Household\ Avoidance)P(Classroom\ Non\text{-}Avoidance)}{P(Infection)}$$

$$= \frac{q^{I(t)-I_{H_i}(t)-I_{C_i}(t)+I_{H_i}\cap I_{C_i}(t)} q_H^{I_{H_i}(t)} (1-q_1^{I_{C_i}(t)})}{1-q^{I(t)-I_{H_i}(t)-I_{C_i}(t)+I_{H_i}\cap I_{C_i}(t)} q_H^{I_{H_i}(t)} q_1^{I_{C_i}(t)}}.$$

Similarly, if we wanted to find the probability that this individual is infected at time $t+1$ solely due to household transmission, we have that

$$\frac{P(General\ Avoidance)P(Household\ Non\text{-}Avoidance)P(Classroom\ Avoidance)}{P(Infection)}$$

$$= \frac{q^{I(t)-I_{H_i}(t)-I_{C_i}(t)+I_{H_i}\cap I_{C_i}(t)} (1-q_H^{I_{H_i}(t)}) q_1^{I_{C_i}(t)}}{1-q^{I(t)-I_{H_i}(t)-I_{C_i}(t)+I_{H_i}\cap I_{C_i}(t)} q_H^{I_{H_i}(t)} q_1^{I_{C_i}(t)}}.$$

To see this more clearly, we can consider a toy example.

Example: Consider the population from the previous example. We are interested in finding the probability that, given an infection occurred, it was solely due to classroom transmission.

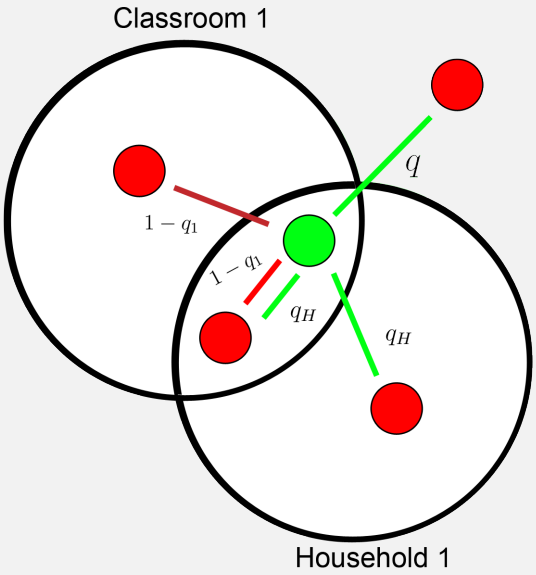


Figure 20: Visualisation of one possible occurrence where the individual was infected as a direct result of classroom transmission

Note that the infective who is in both the susceptible's household and classroom generates two interactions. In this particular event, the susceptible avoids infection from the household interaction but fails to avoid infection from the classroom interaction. The overall probability is then,

$$\frac{P(General\ Avoidance)P(Household\ Avoidance)P(Classroom\ Non\text{-}Avoidance)}{P(Infection)}$$

$$= \frac{qq_H^2(1-q_1^2)}{1-qq_H^2q_1^2}$$

where we have that,

$$P(Infection) = 1 - P(Avoids\ Infection)$$

$$= 1 - P(General\ Avoidance \cap Household\ Avoidance \cap Classroom\ Avoidance)$$

$$= 1 - qq_H^2q_1^2$$

Calculating these transmission probabilities for each infection in the observed epidemic and averaging allows us to estimate the overall proportion of infections that occurred due to classroom, household or general transmission. We find that approximately 35%, 15% and 47% of all infections occurred due to classroom, household and general transmission respectively. Restricting our attention to just those infections where classroom transmission was possible, we find that 90% and 55% of individuals in classroom 1 and 2 were infected solely due to transmission inside their classrooms. This represents a small decrease from the equivalent figures we saw under the extended classroom model which is likely due to the presence of a new source of infection in the form of household transmission.

Now, if we only look at those infections where there was a non-zero chance of household transmission, we find that approximately 55% of these infections occurred solely due to household spread. Once again, we see that the model appears to be over representing the number of household infections when we compare to the 31% figure from the exploratory analysis in Section 4.2. This is possibly due to our relatively low confidence in the estimate of q_H; a different value may be more appropriate here. Overall, from these findings we can see that classroom transmission appears to be a much more significant source of infection than household transmission, even when we consider the proportions of the total number of infections. This is likely due to the relatively small number of individuals in each household who could possibly be infected. In comparison, there are large subpopulations in each classroom and a single infective can quickly affect a significant number of individuals.

8.2 Epidemic Simulation

We can now state our final epidemic simulation algorithm that will allow us to simulate epidemics under the classroom-household model.

Algorithm 6: Classroom-Household Model Simulation Algorithm

Intialise $\hat{S}_1(0)$, $\hat{S}_2(0), ..., \hat{S}_{188}(0)$,

$E(0)$, $I_{H_1}(0), I_{H_2}(0), ..., I_{H_{188}}(0)$, $I_{C_1}(0), I_{C_2}(0), ..., I_{C_{188}}(0), I(0), R(0)$,

q, q_H, q_1, q_2, E, d,

Maximum epidemic length T,

for *t from 1 to T* **do**

  for *i from 1 to 188* **do**

    Let k = classroom number of individual i

    if $\hat{S}_i(t) = 1$ **then**

      if $k = 0$ **then**

        $\hat{S}_i(t) = \text{Bernoulli}(q_H^{I_{H_i}(t-1)} q^{I(t-1)-I_{H_i}(t-1)})$

      end

      if $k = 1$ *or* $k = 2$ **then**

        $\hat{S}_i(t) =$

        $\text{Bernoulli}(q_H^{I_{H_i}(t-1)} q_k^{I_{C_i}(t-1)} q^{I(t-1)-I_{H_i}(t-1)-I_{C_i}(t-1)+I_{H_i} \cap I_{C_i}(t-1)})$.

      end

      if $\hat{S}_i(t) = 0$ **then**

        for t' *from* t *to* $(t + E - 1)$ **do**

          $E(t') = E(t') + 1$

        end

        $\hat{x}_i$ = sample prodromal period length x

        for $\hat{t}$ *from* $(t + E)$ *to* $(t + E + \hat{x}_i + d)$ **do**

          $I(\hat{t}) = I(\hat{t}) + 1$, $I_{C_i}(\hat{t}) = I_{C_i}(\hat{t}) + 1$, $I_{H_i}(\hat{t}) = I_{H_i}(\hat{t}) + 1$

        end

        for $\tilde{t}$ *from* $(t + E + \hat{x}_i + d + 1)$ *to* T **do**

          $R(\tilde{t}) = R(\tilde{t}) + 1$

        end

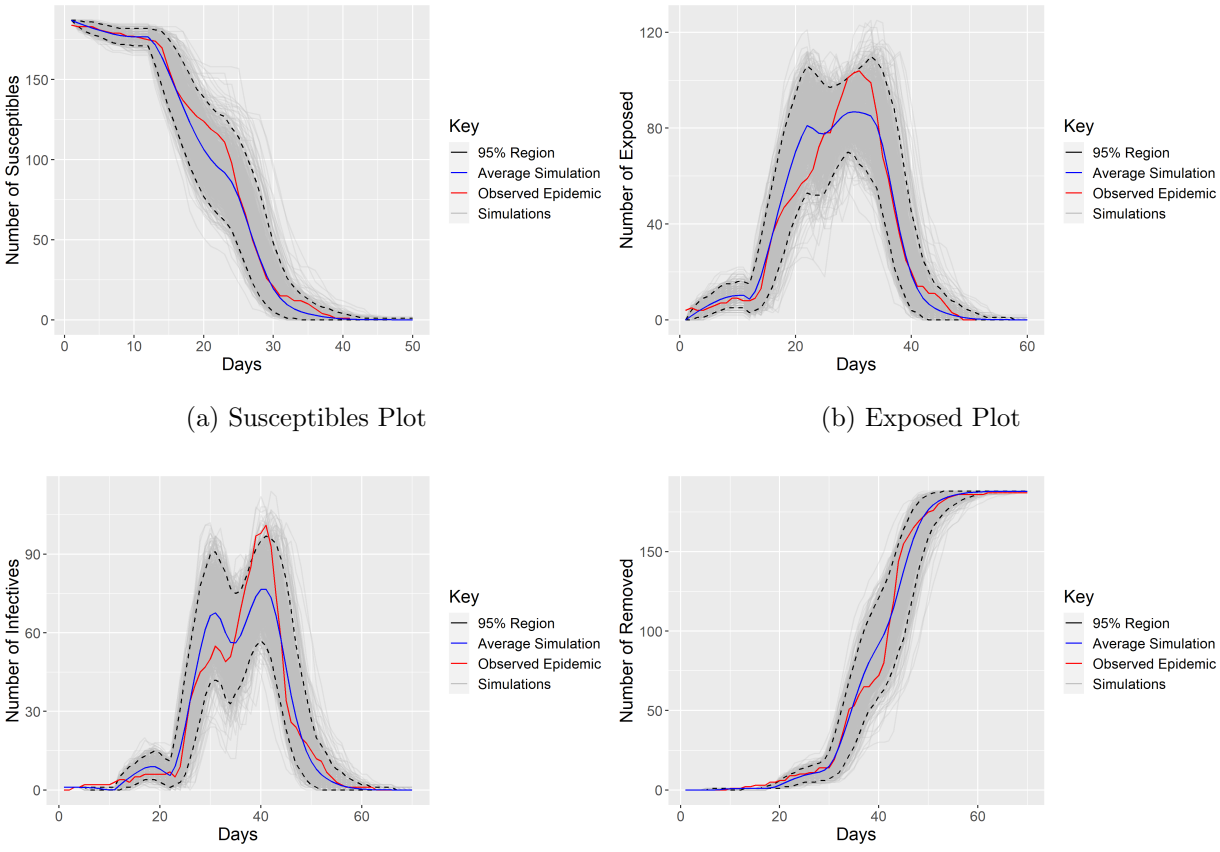
      end

    end

  end

end

We note that in the observed epidemic, the first infective individual attended classroom 2 and was a part of the fifth household. Therefore, we use these initial values in the algorithm to produce 1000 simulations of the epidemic under this model.



(a) Susceptibles Plot

(b) Exposed Plot

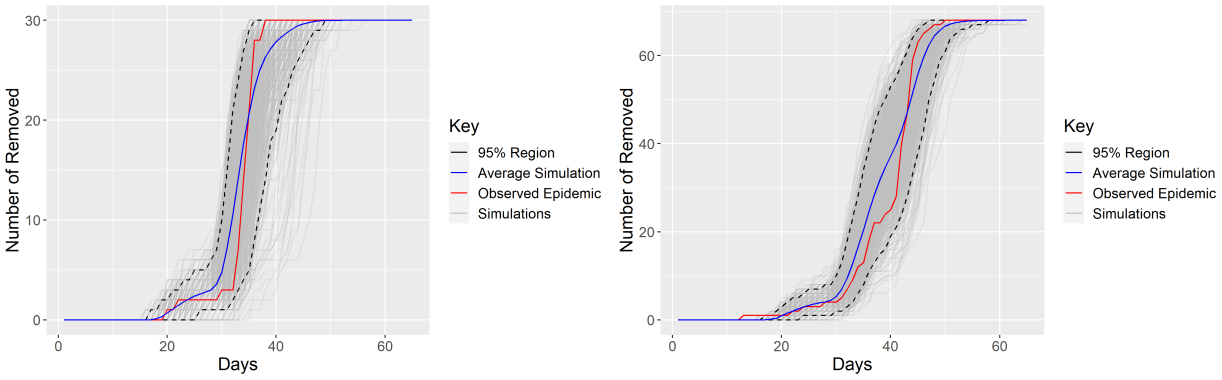
(c) Infectives Plot

(d) Removed Plot

Figure 21: Plot showing 1000 simulated epidemics under the classroom-household model with $q = 0.9961663, q_1 = 0.8319394, q_2 = 0.9788626, q_H = 0.8475385$, the average simulation and the observed epidemic

We would hope that reflecting what we saw in the exploratory analysis more closely by adding further routes of transmission would improve our simulations. In fact, we are quite disappointed in these simulation results; the fit to the observed epidemic is relatively poor in the middle phase of the epidemic when most infections are occurring, especially when compared to the results from the extended classroom model. It seems that the addition of household transmission to our extended classroom model has worsened the overall fit. Despite this, we can still inspect the classroom sub-populations to see how the model has

performed.



(a) Class 1 Removed Plot

(b) Class 2 Removed Plot

Figure 22: Plot of the removed statistic $R(t)$ of each classroom under the classroom-household with $q = 0.9961663, q_1 = 0.8319394, q_2 = 0.9788626, q_H = 0.8475385$, showing the 1000 simulated epidemics, the average simulation and the observed epidemic

Surprisingly, these results are quite strong. In particular, the model has accurately reflected the spread of measles through classroom 1, much more so than any other model thus far. It seems that adding households as an extra transmission pathway has allowed this model to infect individuals at the rate we see in the observed epidemic. Conversely, when we consider classroom 2, we can see that the epidemic progresses too quickly in the middle phase of the epidemic producing a slightly worse fit than what we saw in the extended classroom model. It seems as though the addition of household transmission has improved the fit to the classroom 1 sub-population at the expense of the overall epidemic.

These findings seem to suggest that the inclusion of household transmission, at least under this formulation, does not accurately reflect the infection dynamics in the observed epidemic. The extended classroom model, where we just consider classroom and general transmission, produces significantly better simulation results. However, as stated previously, this could also be due to poor parameter estimates which may be distorting the simulation results. Therefore, it is of interest to use *Bayesian* inference to produce alternate estimates and intervals which we may have more confidence in.

9 Bayesian Inference

Throughout the dissertation we have used classical statistical techniques to estimate model parameters. That is, we treat the unknown model parameters as having some fixed unknown value, and all probabilistic statements relate to random variables which we observe. In our case these random variables are the susceptible population statistics. Estimation is then performed by maximising the likelihood function to produce MLE's such that under our assumed statistical model, the observed data is most probable. These MLE's are *point estimates*, i.e. a single value which serves as our best guess of the unknown parameter. In order to produce approximate confidence intervals for these point estimates, we rely on an asymptotic assumption of the normality of the MLE. The validity of this assumption has frequently been called into question due to a lack of sample data. Thus we are interested in exploring another method of gaining information about our parameters.

The *Bayesian* approach treats the parameters themselves as random variables about which we can make probabilistic statements. Inference then depends on a mixture of our initial knowledge of these parameters and how the introduction of data updates this knowledge. Using this method, we gain access to entire probability distributions for the parameters and "confidence intervals" (here they are *credible intervals*) which are not based on any asymptotic assumptions. Discounting our concerns about the classical approach, it is also of interest to apply Bayesian methods in their own right.

We adopt the following framework. Consider the random variables $\boldsymbol{X} = (X_1, ..., X_n)$ with observed values $\boldsymbol{x} = (x_1, ..., x_n)$. Then, given the parameter vector $\boldsymbol{\theta}$, we denote the probability model for this data by $p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta})$. Note that is equivalent to the likelihood function. Then, we summarise our initial knowledge about $\boldsymbol{\theta}$ by the pdf/pmf $\pi(\boldsymbol{\theta})$, which we call the *prior* distribution on $\boldsymbol{\theta}$.

We can then use Bayes' Theorem,

$$P(A|B) = P(B|A)\frac{P(A)}{P(B)},$$

to get the *posterior* pdf/pmf of $\boldsymbol{\theta}$ given the data x,

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) = p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta})\frac{\pi(\boldsymbol{\theta})}{m(\boldsymbol{x})},$$

where $m(\boldsymbol{x})$ is the marginal pdf/pmf of $\boldsymbol{x}$. The posterior is an expression of the updated knowledge of $\boldsymbol{\theta}$ after we observe the data. Note that we will use the formula

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) \propto p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}),$$

as for our purposes it is unnecessary to find the actual value of the normalising constant.

We need to choose the prior distribution ourselves. This choice falls into one of two categories; *informative* and *non-informative*. We have little prior knowledge about our model parameters, so we use a non-informative prior. There are multiple ways to do this, but by noting that our model parameters are avoidance probabilities (i.e q, q_c, q_H etc) and therefore must fall in the interval [0,1], we can choose a uniform distribution on [0,1]. Then, we have that each parameter θ is independently drawn from a $U(0,1)$ distribution and thus,

$$\pi(\boldsymbol{\theta}) = \begin{cases} 1 & \text{if} \quad \boldsymbol{\theta} \in [0,1] \\ 0, & \text{otherwise} \end{cases}$$

Therefore, the posterior distribution has the following simplified form,

$$\begin{aligned} \pi(\boldsymbol{\theta}|\boldsymbol{x}) &\propto p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ &= \begin{cases} p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}) & \text{if} \quad \boldsymbol{\theta} \in [0,1] \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

Now that we know the entire posterior distribution of a parameter, point estimates and intervals are useful as summaries of the posterior. In particular, we will be considering the posterior mean $E[\boldsymbol{\theta}|\boldsymbol{x}]$[6] and *Bayesian Credible Regions* (BCR).

[6]Other summaries of loaction include the posterior median and posterior mode. Note that with our choice of prior, the posterior mode is equivalent to the MLE

Definition [19]: The random region C_α is a $100(1-\alpha)\%$ Bayesian Credible Region for θ if

$$P(\boldsymbol{\theta} \in C_\alpha | \boldsymbol{x}) = 1 - \alpha.$$

Note that these regions are not unique. Therefore, as we will be considering the posterior mean for our parameter estimates, we choose the BCR that is centred on the mean. We interpret Bayesian credible intervals C_{B_α} differently than classical confidence intervals C_{C_α}. In the Bayesian setting, we think of θ as a random variable and the bounds of C_{B_α} as fixed. Conversely, in the classical approach, we say that C_{C_α} contains the true fixed value of θ, with the bounds as random variables.

Now that we have an understanding of the basic ideas behind Bayesian inference, we can explore a method of sampling from the posterior distribution such that we can calculate the posterior means and Bayesian credible regions for our parameters.

9.1 Markov Chain Monte Carlo Methods

We want to sample, up to proportionality, from the posterior density $\pi(\boldsymbol{\theta}|\boldsymbol{x})$. We use Markov Chain Monte Carlo (MCMC) methods which consist of a class of algorithms used to sample from probability distributions. These methods were born out of the Los Alomos National Laboratory during World War II as part of the research done by physicists during their work on mathematical physics and the atomic bomb [20].

The main idea is to construct a Markov chain whose stationary distribution is our probability distribution of interest. Then, by simulating the chain, in the long run it will "forget" its starting point and its values will approximately be from the target distribution, $\pi(\boldsymbol{\theta}|\boldsymbol{x})$. Before we discuss the details of our choice of MCMC algorithm, we introduce some key facts about Markov chains which are vital for fully understanding MCMC.

9.1.1 Key Facts About Markov Chains

Definition [21]: A sequence $X_2, X_2, X_3, ...$ of random variables taking values in a state space S is a *Markov chain* if

$$P(X_{t+1} \in A | X_1 = x_1, ..., X_t = x_t) = P(X_{t+1} \in A | X_t = x_t)$$

for all $A \subset S$ for which $P(A)$ is defined.

The main takeaway here is that inside a Markov chain, the value of a random variable is only influenced by the value of the random variable immediately preceding it.

Definition [21]: The *transition kernel* $Q(.,.)$ defines the dynamics of the chain

$$Q(x, y) = f_{X_{t+1}|X_t}(y|x).$$

That is, $Q(x, y)$ is the conditional pdf of a point y if the chain is currently at x, or equivalently, the likelihood of moving to y given that the state is currently at x. We assume *time-homogeneity*, so Q does not depend on t.

Definition [21]: A distribution π is a *stationary distribution* of a chain with transition kernel Q if

$$\pi = \pi Q.$$

That is, $\pi(y) = \int \pi(x)Q(x, y)dx$ for all y. It can be shown that π is a stationary distribution of Q if the *detailed-balance* equations hold:

$$\pi(x)Q(x, y) = \pi(y)Q(y, x)$$

for all x, y. The intuition behind the detailed-balance equations is that the flow of probability is the same going from x to y as it going from y to x. Thus the chain is stationary, or in *equilibrium*.

A chain is said to be *irreducible* if it it is possible to move from any given state to any other state in a finite number of steps. We also have that a chain is *aperiodic* if there is no periodic relationship between when it is possible for a Markov chain to return areas

within its state space [21].

Convergence Theorem [21]: An aperiodic, irreducible Markov chain $X_1, X_2, ...$ with transition kernel Q and stationary distribution π will converge to its stationary distribution. That is,

$$P(X_t \in A) \to \pi(A) \; \forall \; A \;\; \text{as} \;\; t \to \infty.$$

Ergodic Theorem [21]: An aperiodic, irreducible Markov chain $X_1, X_2, ...$ with transition kernel Q and stationary distribution π is ergodic. That is,

$$\bar{g}_n = \frac{1}{n} \sum_{t=1}^{n} g(X_t) \to E_\pi[g(x)] \;\; \text{as} \;\; n \to \infty.$$

In combination, the ergodic and convergence theorem provide the toolset with which MCMC methods work. That is, we estimate $E_\pi[g(x)]$ by evaluating $\bar{g}_n$ at the simulated values of a Markov chain with dynamics such that its stationary distribution is our target distribution $\pi(\boldsymbol{\theta}|\boldsymbol{x})$. Therefore, we can estimate the posterior mean by,

$$E[\boldsymbol{\theta}|\boldsymbol{x}] \approx \frac{1}{n} \sum_{t=1}^{n} X_t.$$

9.1.2 Metropolis-Hastings Algorithm

The first MCMC algorithm for symmetric proposal distributions was developed in 1953 by Greek-American physicist Nicholas Metropolis at the Los Alomos National Laboratory [22]. In 1970 Canadian statistician W. K. Hastings extended it to the more general Metropolis-Hastings algorithm which tells us how to build transition kernels $Q(x, y)$ such that the Markov chain converges to the stationary distribution $\pi(x)$ [23].

Algorithm 7: Metropolis-Hastings Algorithm [21]

1. Choose a starting location $\boldsymbol{X}_0 = \boldsymbol{x}_0$
2. Suppose at time t, we have $\boldsymbol{X}_t = \boldsymbol{x}$. Sample a candidate value $\boldsymbol{y}$ from a proposal distribution $q(\boldsymbol{x}, \boldsymbol{y}) = q(\boldsymbol{y}|\boldsymbol{x})$
3. Calculate the acceptance ratio $\alpha(\boldsymbol{x}, \boldsymbol{y})$ given by

$$\alpha(\boldsymbol{x}, \boldsymbol{y}) = \frac{\pi(\boldsymbol{y})q(\boldsymbol{y}, \boldsymbol{x})}{\pi(\boldsymbol{x})q(\boldsymbol{x}, \boldsymbol{y})}$$

4. Generate a value u from a U(0,1) distribution
5. Accept the move if $u < \alpha(\boldsymbol{x}, \boldsymbol{y})$. Otherwise, we reject the move and stay at $\boldsymbol{x}$. That is, we set

$$\boldsymbol{X}_{t+1} = \begin{cases} \boldsymbol{y}, & \text{if } u < \alpha(\boldsymbol{x}, \boldsymbol{y}) \\ \boldsymbol{x}, & \text{otherwise} \end{cases}$$

In the limit, it can be shown that the chain produced by the Metropolis-Hastings algorithm will converge to the stationary distribution $\pi(\boldsymbol{x})$. We also note that the algorithm works for any proposal distribution $q(\boldsymbol{x}, \boldsymbol{y})$ where the chain ends up being ergodic and therefore this choice is essentially arbitrary. Although this does give us lots of freedom, we do still need to ensure that the choice of proposal distribution is a good one.

The question is then, how do we assess the performance of an MCMC algorithm? We want the resulting Markov chain to converge quickly to the posterior density and to "mix" well throughout the support of the density. That is, to explore the support efficiently. This involves balancing a chain that has a high acceptance rate for moves, but only explores the support slowly, and a chain that proposes moves far from the current value but which are only rarely accepted. This can be a very difficult task if the target distribution is particularly complicated. We will see that our target distribution, the posterior density, or equivalently here, the likelihood function, is relatively simple and thus the choice of proposal distributions is not too difficult.

Markov Chain Monte Carlo methods do have certain drawbacks. For our needs, we need to

be aware of the fact that we are not independently sampling from the posterior density. We have to take this into account when doing estimation and can be countered by "thinning" the chain, i.e. take every kth observation and discard the rest. A sufficiently thinned chain can leave us with a sample that is close to being independent and identically distributed from the posterior density. This is a particular issue when calculating the variance of our estimators and in our case is something we consider when designing good proposal distributions for our multivariate models. There are also computational factors at play. For example, the classroom-household model took many hours to produce a chain with sufficient length to be fully confident in our estimations. If we considered models that were any more complicated it would require moving to a coding language that is more computationally efficient than R, such as $C++$.

9.2 MCMC Parameter Estimation

Now that we have introduced the MCMC method, we can apply the Metropolis-Hastings algorithm to each of our models. We only detail the full process for the base model, extended classroom model and the classroom-household model.

9.2.1 Base Model MCMC

We begin with the base model with the parameter q. The first step is to identify a suitable proposal distribution, to do this we need to plot the posterior density, or equivalently, the likelihood function $L(q)$.

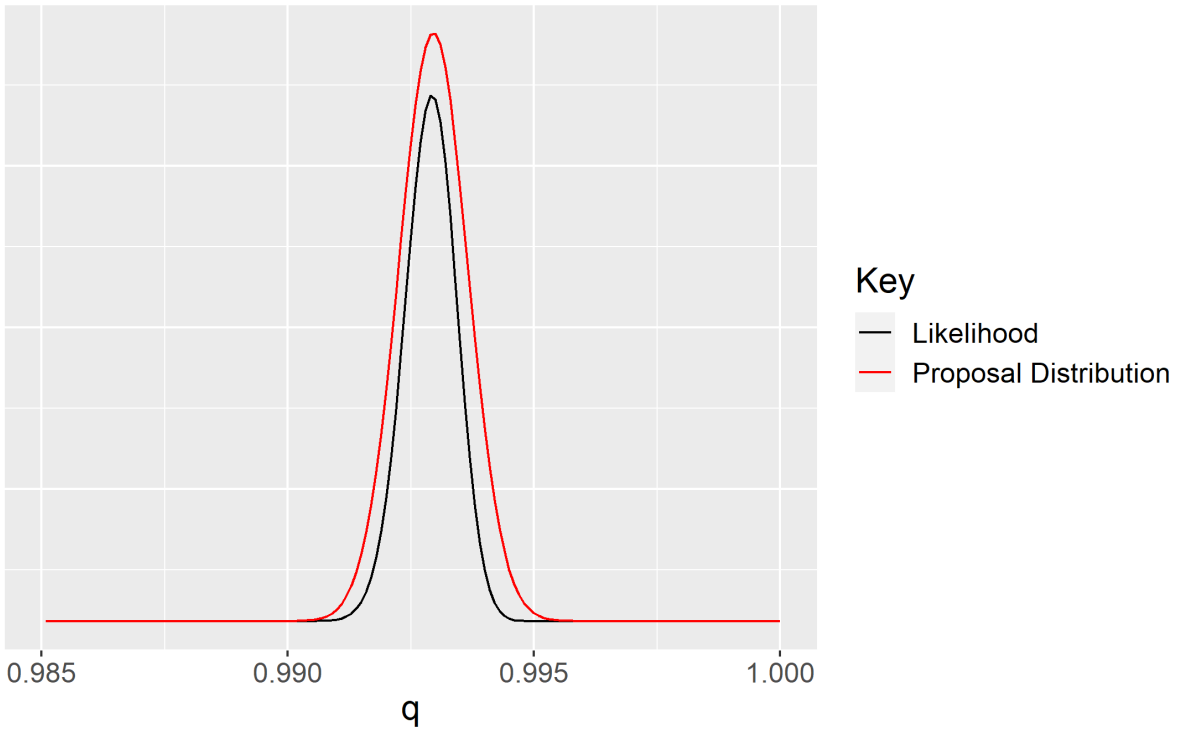


Figure 23: Plot of the base model likelihood function with a Normal proposal distribution

From this figure we can see that the likelihood function can be approximated well by a normal distribution $N(\mu, \sigma^2)$. Specifically, we choose $\sigma^2 = 0.00125$ and μ to be the most recently sampled value of q. Then we let $f(x)$ represent the pdf of this normal distribution. Now that we have a candidate proposal distribution f, we can apply the algorithm to generate samples from the posterior. Note that in the below algorithm we use the log-likelihood $\log(L(q))$ in place of the likelihood, this is an equivalent approach as we can simply take the log of the acceptance ratio and proceed as normal.

Algorithm 8: Base Model Metropolis-Hastings Algorithm

1. Choose a starting location $q = q_0$
2. Suppose at time t, we have $Q_t = q$. Sample a candidate value y from a $N(q, 0.00125)$ distribution
3. Calculate the log-ratio

$$\alpha(q, y) = \log(L(y)) + \log(f(q)) - \log(L(q)) - \log(f(y))$$

4. Generate a value u from a U(0,1) distribution
5. Accept the move if $\log(u) < \alpha(q, y)$. Otherwise we reject the move and stay at q. That is, we set

$$\boldsymbol{Q}_{t+1} = \begin{cases} y, & \text{if } \log(u) < \alpha(q, y) \\ q, & \text{otherwise} \end{cases}$$

We use this algorithm to generate one hundred thousand samples from the posterior distribution.

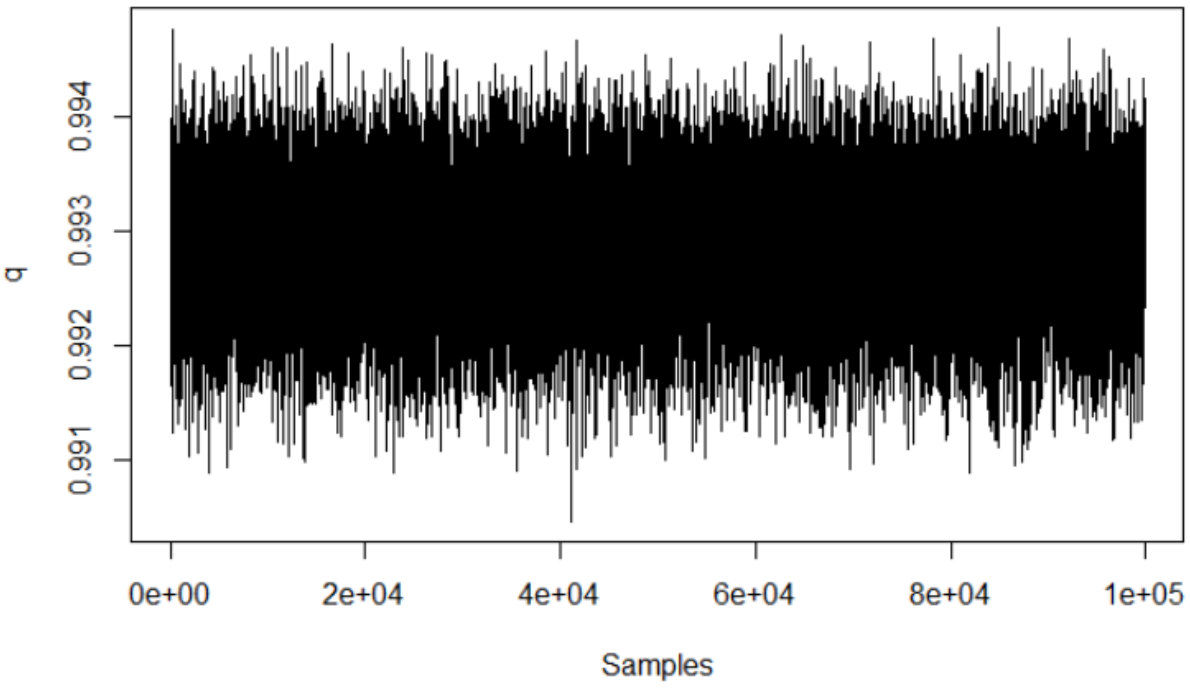


Figure 24: Trace plot showing the 100000 samples under the base model versus the iteration number

It appears that the chain is quickly moving around the state space achieving a good balance between a high acceptance rate and fully exploring the support of the posterior. Indeed, with this choice of proposal distribution, we get an acceptance rate of 44% which is generally thought to be optimal for low-dimensional problems such as this [24]. It is quite clear that the chain is mixing well and has converged to the posterior distribution, despite this we can formally check by plotting the *autocorrelation* of the samples. The lag-k autocorrelation is the correlation between each sample and the sample k steps previously. This should become smaller as k increases, i.e. samples can be considered as effectively independent as they become further and further apart from one another.

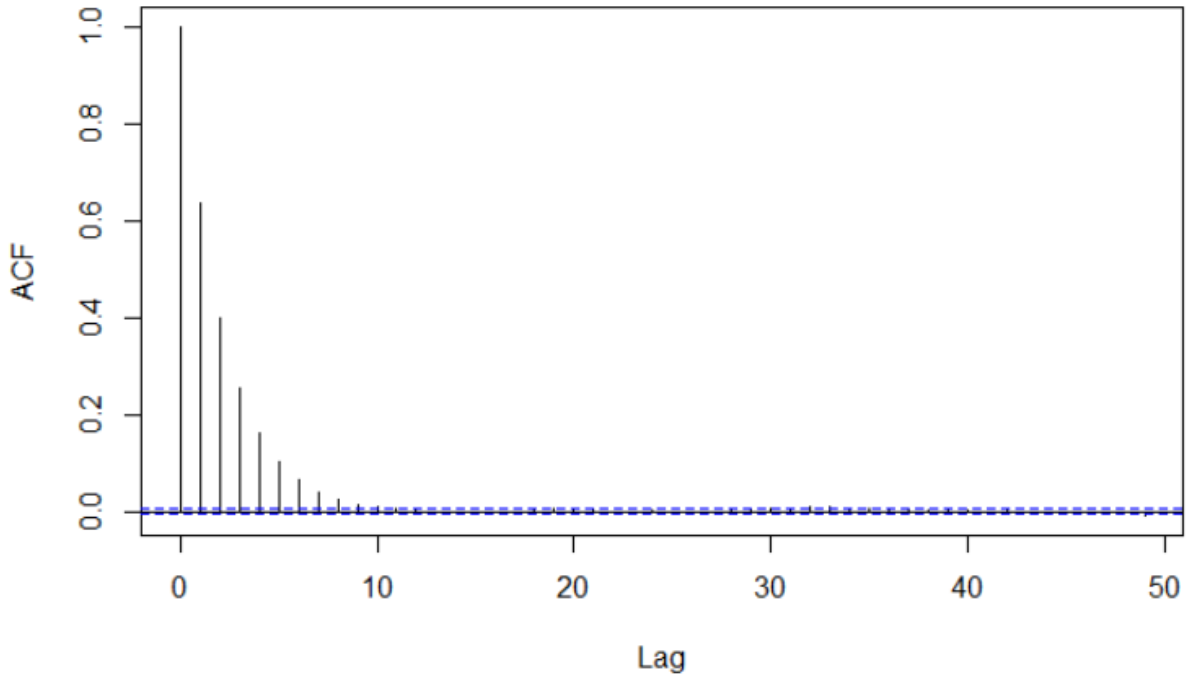


Figure 25: ACF plot of the sample under the base model

We can see that the chain is mixing well, by the 10th lag the correlation is no longer statistically significant (indicated by the horizontal lines falling between the blue dashed lines). If we saw the autocorrelation persisting for high lag we could try thinning the chain. In this case it is unnecessary, but for illustrative purposes by choosing every 10th iteration and discarding the rest, we get the following ACF.

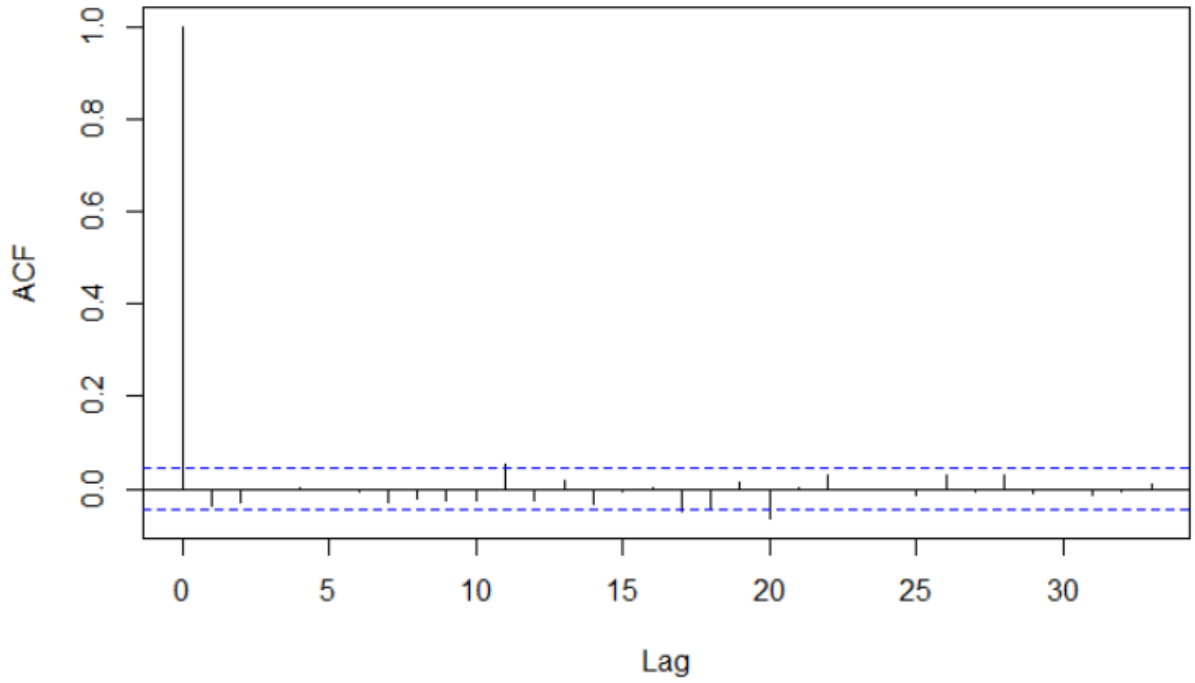


Figure 26: ACF plot of the thinned sample under the base model

Here we note that the autocorrelation at lag 0 is always equal to 1. Thus we we can say for the thinned chain we have effectively zero autocorrelation and the sample can be treated as independently and identically distributed. Now, we have one hundred thousand samples from the posterior density and averaging them allows us to estimate the posterior mean. We can also calculate the 95% credible region centred on this mean to get that $q = 0.9928855 \pm 0.001026197$.

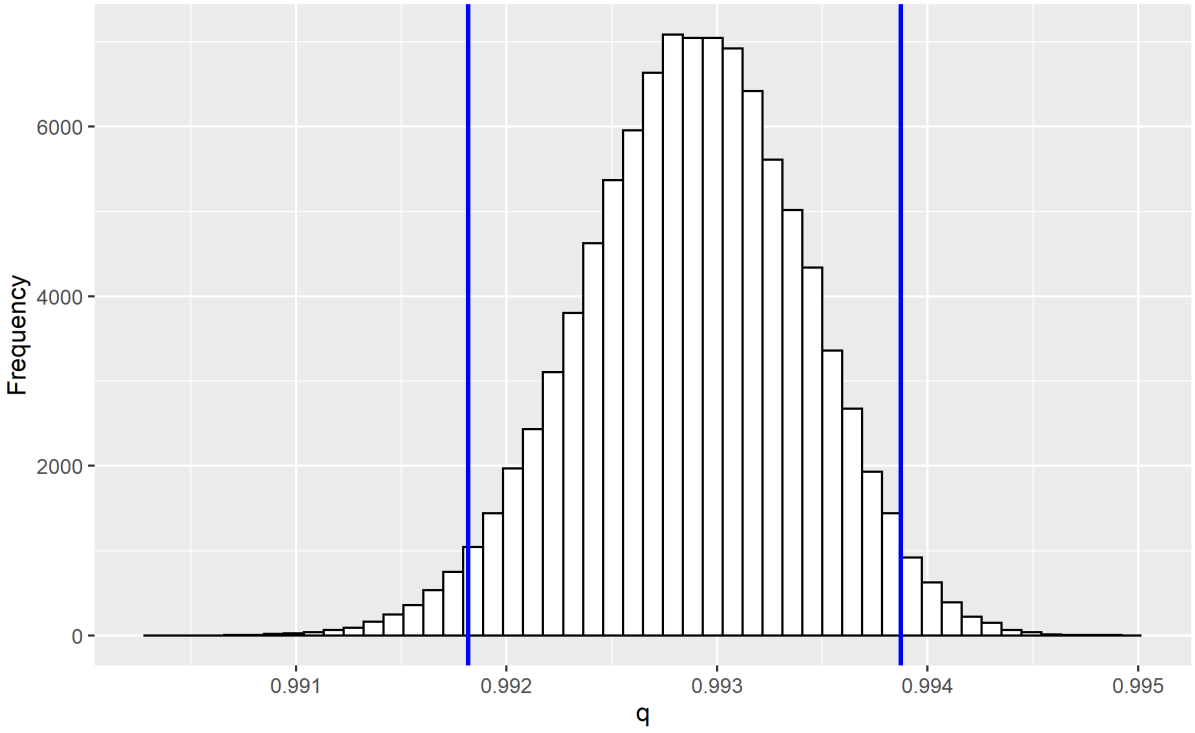


Figure 27: Histograms of the q samples under the base model with associated 95% BCR

The vertical blue lines indicate the 95% credible region centred on the posterior mean. This parameter estimate and BCR is very similar to the estimate and 95% confidence interval we found in Section 5 using classical statistical techniques.

9.2.2 Extended Classroom Model MCMC

We have shown earlier that the extended classroom model produces the best simulation results and seems to accurately represent the infection dynamics of the observed epidemic. Thus we are interested in applying MCMC methods to the extended classroom model in order to estimate the parameters q_1, q_2 and q.

Like previously, we can use the likelihood function to identify suitable proposal distributions. We have two choices here; we can use three separate independent normal distributions to propose values for each parameter, or we can use a multivariate normal distribution. The latter is more computationally efficient and more likely to produce stronger results as we will be taking into account the covariance structure of the parameters. However, we have no way of knowing the covariance matrix exactly. Instead, we

first run the algorithm with three manually identified separate proposal distributions[7], then we can estimate the covariance matrix by using the resulting samples of q_1, q_2 and q. Once this was done, we then proceeded with the following algorithm.

Algorithm 9: Extended Classroom Model Metropolis-Hastings Algorithm

---

1. Choose starting locations $\boldsymbol{q} = (q_0,\ q_{1_0}, q_{2_0})$
2. Suppose at time t, we have $\mathbf{Q}_t = \boldsymbol{q}$. Sample candidate values $\boldsymbol{y}$ from a $\mathcal{N}(\boldsymbol{q}, \Sigma)$ distribution where Σ is our estimated covariance matrix
3. Calculate the log-ratio

$$\alpha(\boldsymbol{q}, \boldsymbol{y}) = \log(L(\boldsymbol{y})) + \log(f(\boldsymbol{q})) - \log(L(\boldsymbol{q})) - \log(f(\boldsymbol{y}))$$

4. Generate a value u from a U(0,1) distribution
5. Accept the move if $\log(u) < \alpha(\boldsymbol{q}, \boldsymbol{y})$. Otherwise we reject the move and stay at $\boldsymbol{q}$. That is, we set

$$\boldsymbol{Q}_{t+1} = \begin{cases} \boldsymbol{y}, & \text{if } \log(u) < \alpha(\boldsymbol{q}, \boldsymbol{y}) \\ \boldsymbol{q}, & \text{otherwise} \end{cases}$$

Once again we use this algorithm to generate one hundred thousand samples from the posterior distribution.

[7] To do this we plot the likelihood varying one parameter and keeping the others constant

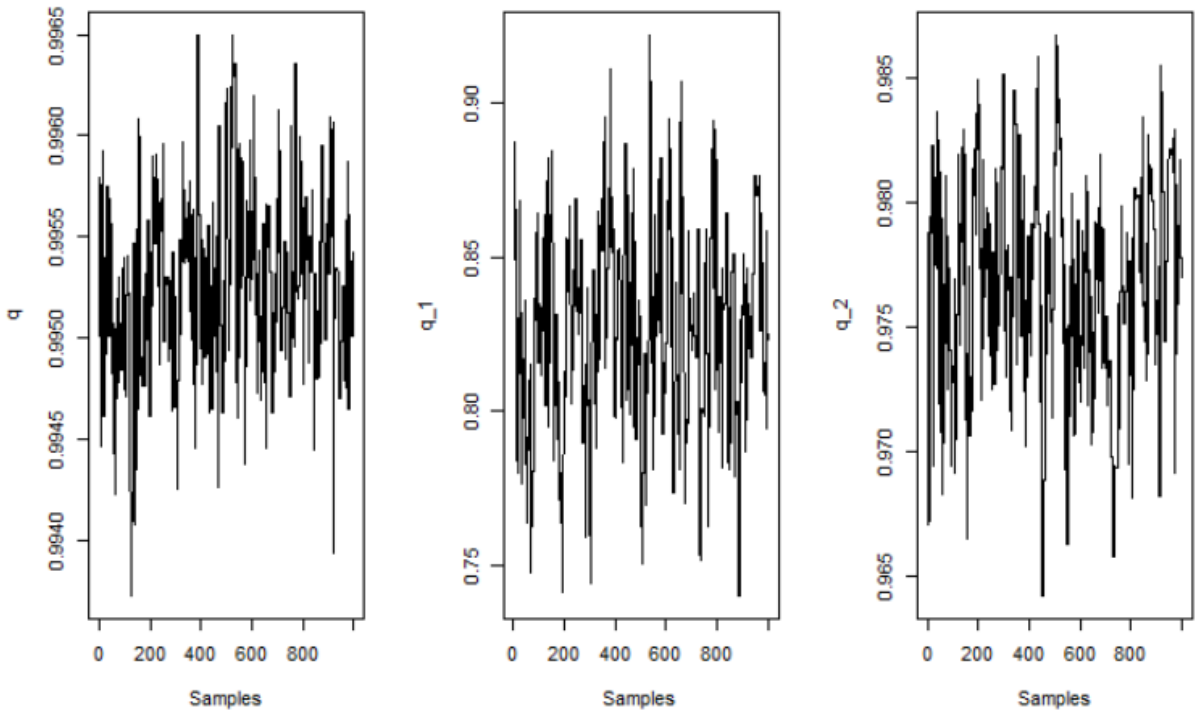


Figure 28: Trace plot of 1000 samples of each parameter under the extended classroom model. From left to right: q, q_1 and q_2

Figure 28 shows a zoomed trace plots highlighting 1000 of the 100000 samples of each parameter. This clearly shows that the chain is exploring the support effectively without getting "stuck" too frequently or for too long. Too see this more formally we can check the autocorrelation plots.

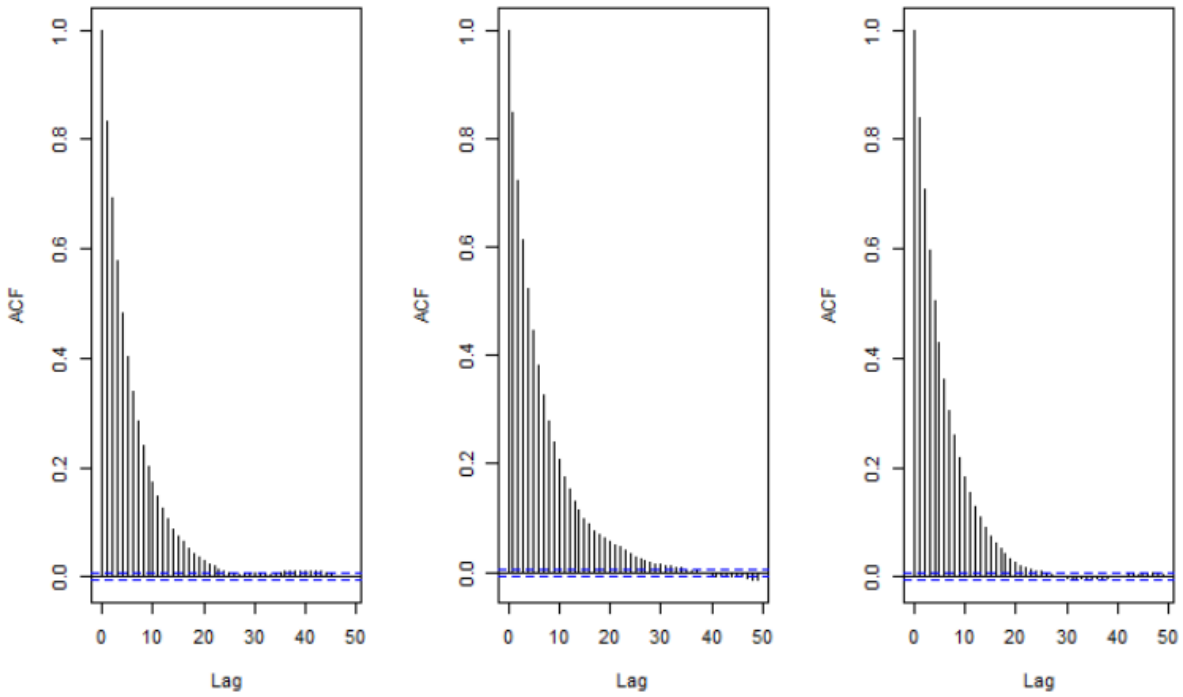


Figure 29: ACF plot of the samples of each parameter under the extended classroom model. From left to right: q, q_1 and q_2

Once again we see that the autocorrelation dies down relatively quickly, this time it takes a few more lags than in the base model, but this is still within the bounds of acceptability and therefore thinning is unnecessary.

Another way we can assess the performance of our MCMC algorithm is to check the posterior correlations between parameters. This is different to autocorrelation which is concerned with the correlation between successive samples of the same parameter. We do not have reason to suspect that the posterior distribution of our parameters are highly correlated. For example, if the avoidance parameter for general transmission q increases, there is nothing to suggest that the avoidance parameter for classroom 1 transmission q_1 will increase or decrease. If our parameter samples appear to be highly correlated, then that indicates that our samples are only allowing us to estimate combinations of the parameters and not each parameter separately. Let our parameter vector be $\boldsymbol{\theta} = (q, q_1, q_2)$, then

$$\boldsymbol{\rho_\theta} = \begin{bmatrix} 1 & -0.0229987051 & -0.2055232000 \\ -0.02299871 & 1 & 0.0001738542 \\ -0.2055232000 & 0.0001738542 & 1 \end{bmatrix}$$

These posterior correlations are easily visualised through the use of a pairs plot.

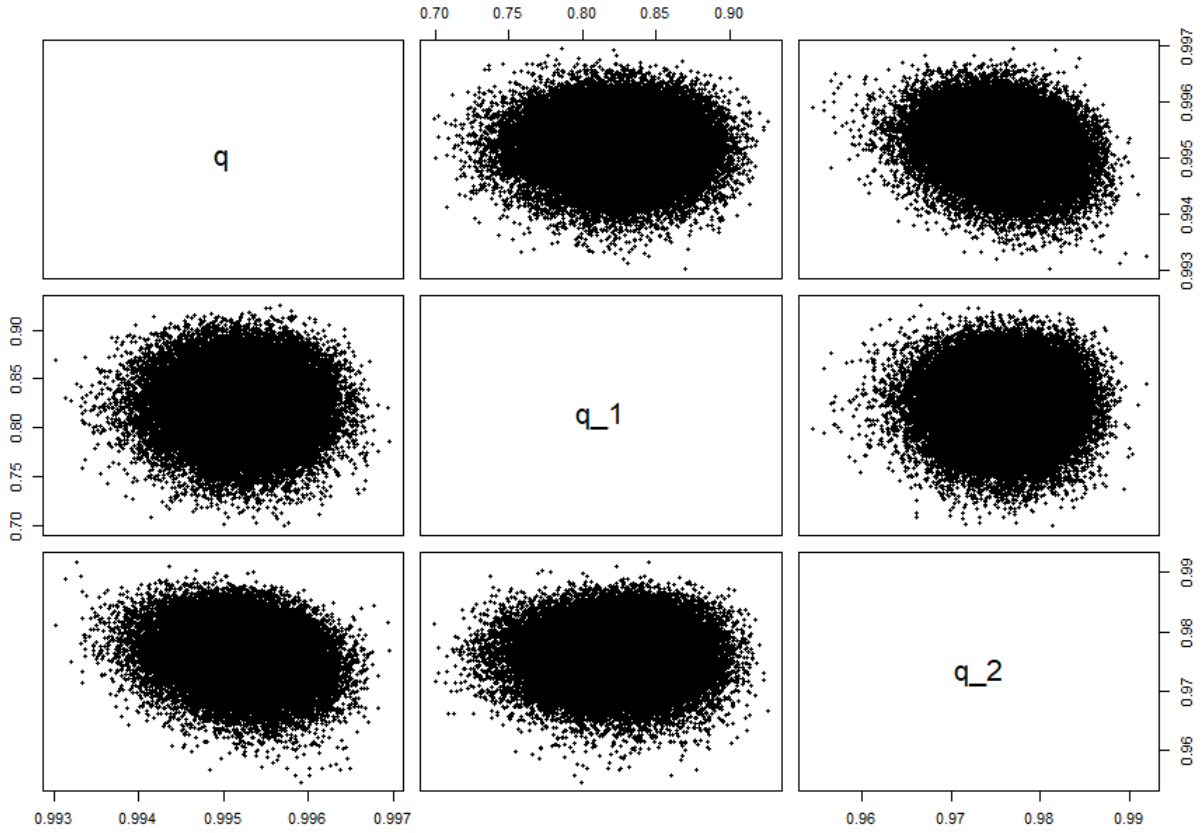


Figure 30: Pairs plot of the posterior samples under the extended classroom model

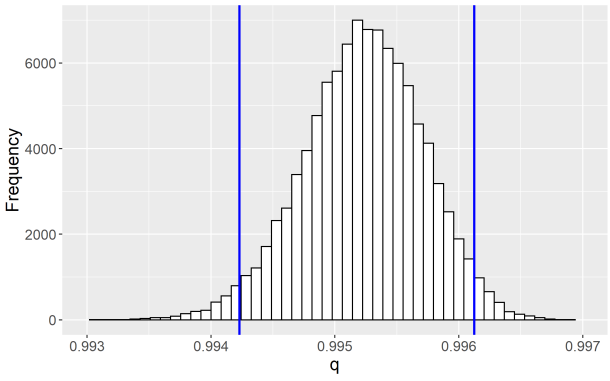
We can see the slight negative correlation between q and q_2 which reflects the value of -0.2055232000 from the correlation matrix. The rest of the correlations are very minimal. A good rule of thumb for when posterior correlations are a cause for concern is if they are greater in modulus than 0.5, at this point we would begin to worry that we are not accurately estimating each parameter separately. We can see from the above correlation matrix that the rule of thumb is not breached by the posterior correlations in this case. Therefore, we can be confident that will not run into issues relating to this with our estimates of the posterior means.

Now that we are completely satisfied with our samples, we can average each set to estimate the posterior means. We also find the 95% BCR for each parameter. Doing so, we get that,

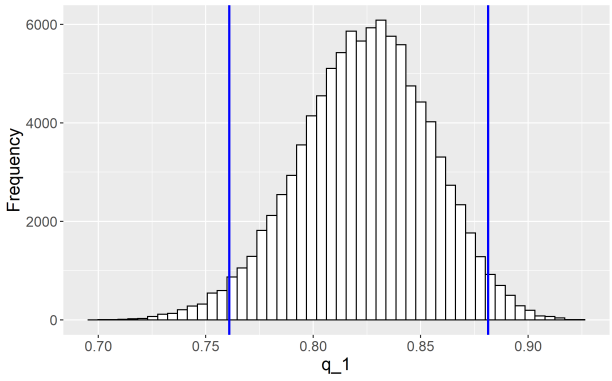
$$\hat{q} = 0.99523 \pm 0.0009473393$$

$$\hat{q}_1 = 0.8241714 \pm 0.06021979$$

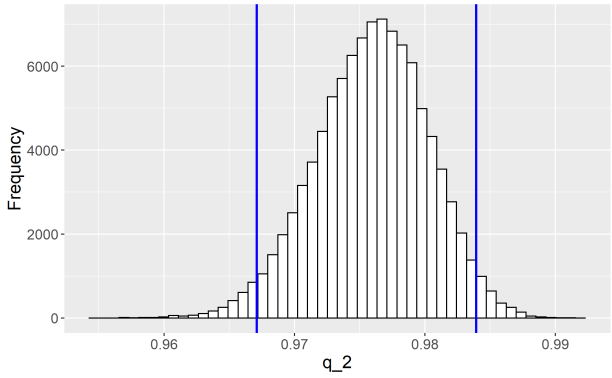
$$\hat{q}_2 = 0.9759522 \pm 0.008415046$$



(a) Samples for q



(b) Samples for q_1



(c) Samples for q_2

Figure 31: Histograms for each parameter samples under the extended classroom model with associated 95% BCRs

These parameter estimates are quite close to those we found through a classical statistical approach and produce simulations that perform very similarly. This is good news as it gives us further confidence in the extended classroom model itself, which produced strong results.

9.2.3 Classroom-Household Model MCMC

In Section 8 we saw that we were very uncertain of our classical parameter estimates for the classroom-household model, with lots of doubt cast on the validity of the asymptotic normal approximation of the MLE. This could have contributed to the relatively poor simulation results produced by the model, thus it is of interest to approach this model in a Bayesian manner to see if our conclusion that household transmission does not accurately reflect the dynamics of the observed epidemic is valid.

We use a very similar algorithm to the one we saw in previous section to produce a Markov chain that converges to the posterior density of the classroom-household model. Computational limitations meant that generating more than fifty thousand samples from the posterior was time prohibitive, however this is more than enough to achieve convergence.

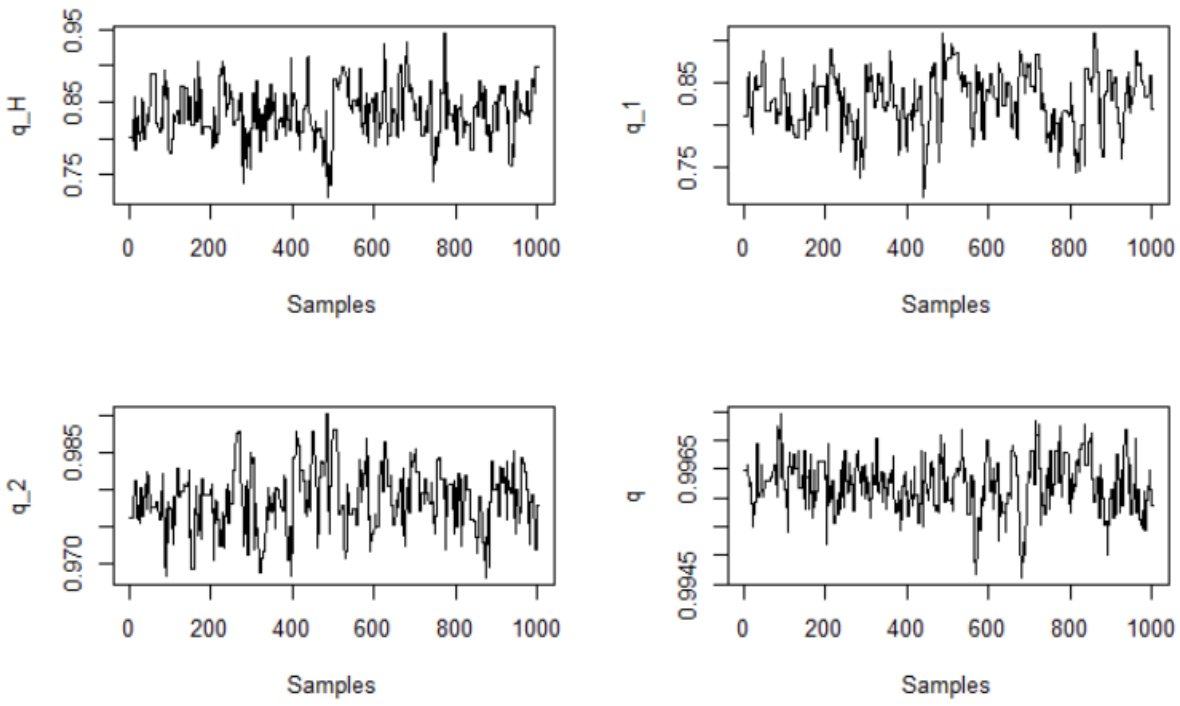


Figure 32: Trace plots of 1000 samples of each parameter q_H, q_1, q_2 and q under the classroom-household model

The above figure shows the first 1000 samples of each overall parameter sample. We can see that the chain does seem to get stuck relatively frequently, sometimes for a significant

number of samples. To see this informally, note that these trace plots are more "blocky" than the previous ones we have seen. This indicates that the chain is not mixing as well as we would like. We can inspect the ACF plots to see what they tell us.

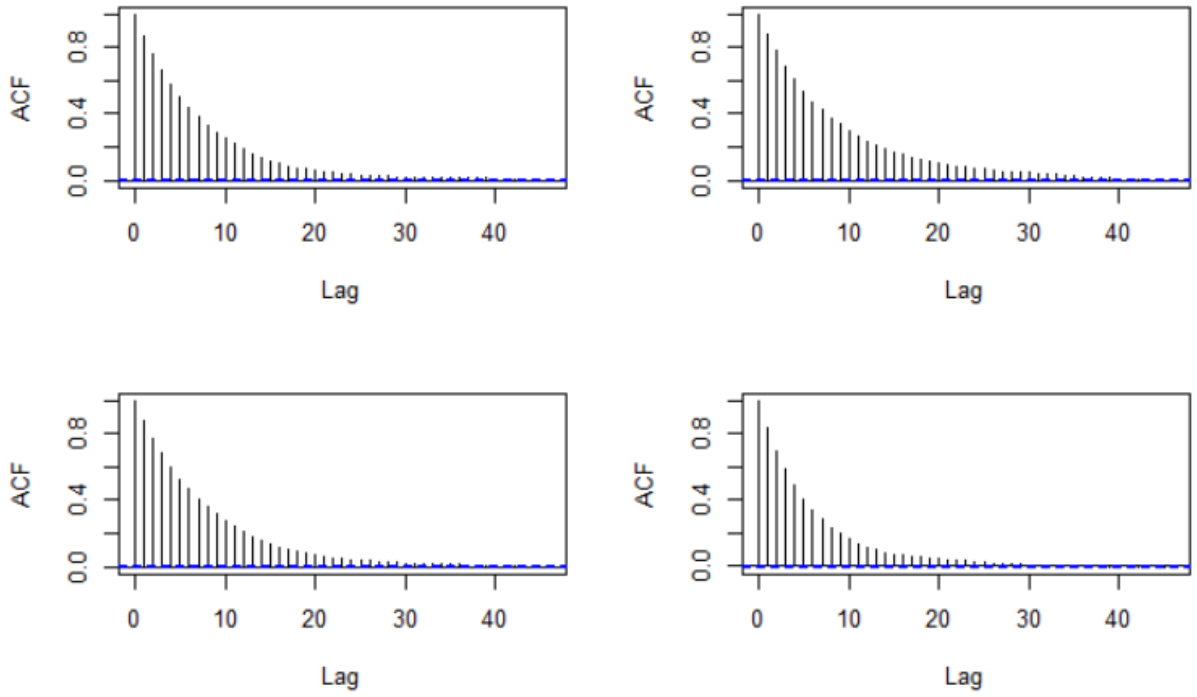


Figure 33: ACF plots of the whole sample for each parameter under the classroom-household model. Top left q_H, top right q_1, bottom left q_2 and bottom right q

The autocorrelation we see here is actually sufficiently good, with the ACF for q_H, q_2 and q dying down before lag 30. We do see that the autocorrelation for q_1 remains significant up to lag 40. This on its own would not be enough to suggest thinning, but in conjunction with the "boxiness" of the trace plots, we can try it to see if we get improvement.

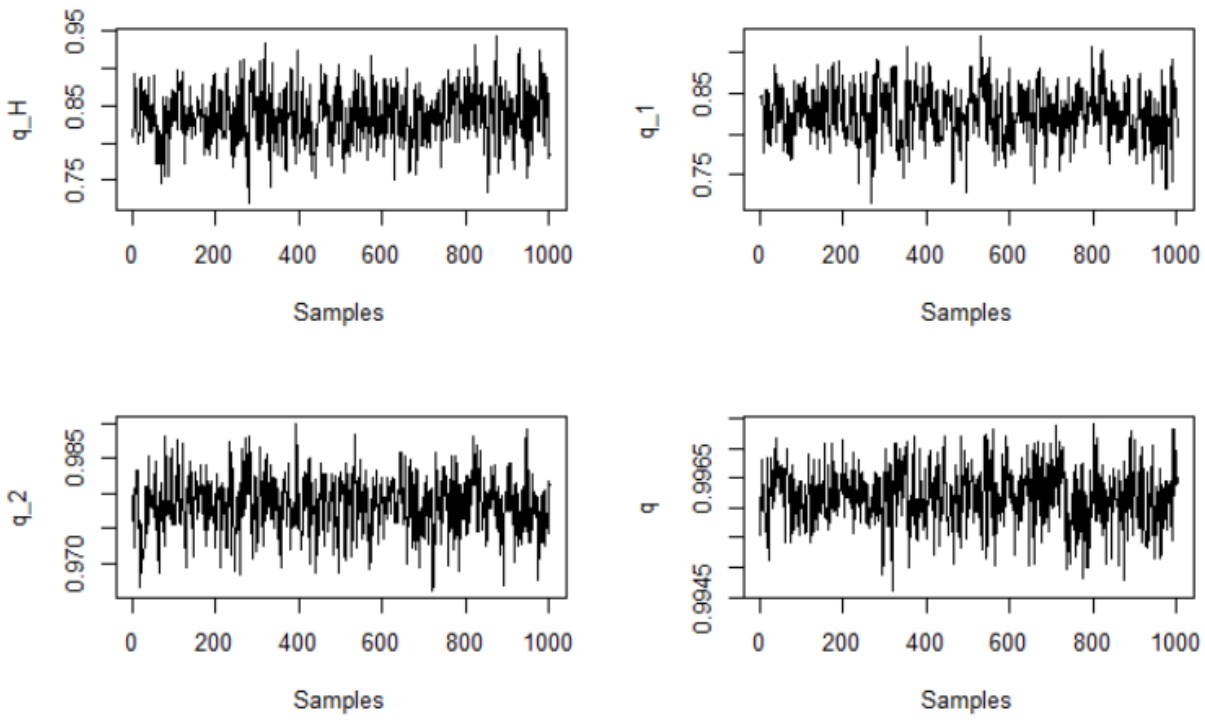


Figure 34: Thinned trace plots of 1000 samples of each parameter q_H, q_1, q_2 and q under the classroom-household model

Here we have taken every fifth observation and discarded the rest. This leaves us with 10000 samples which should be more than enough to produce strong estimates given that the chain clearly converged within the first 1000 initial samples. From these new trace plots, zoomed in on 1000 samples, we see a significant improvement in the mixing of the chain. The "boxiness" we observed previously is now gone and the chain looks to be exploring the whole support of the posterior effectively. We can see the effects of this thinning on the ACF below.

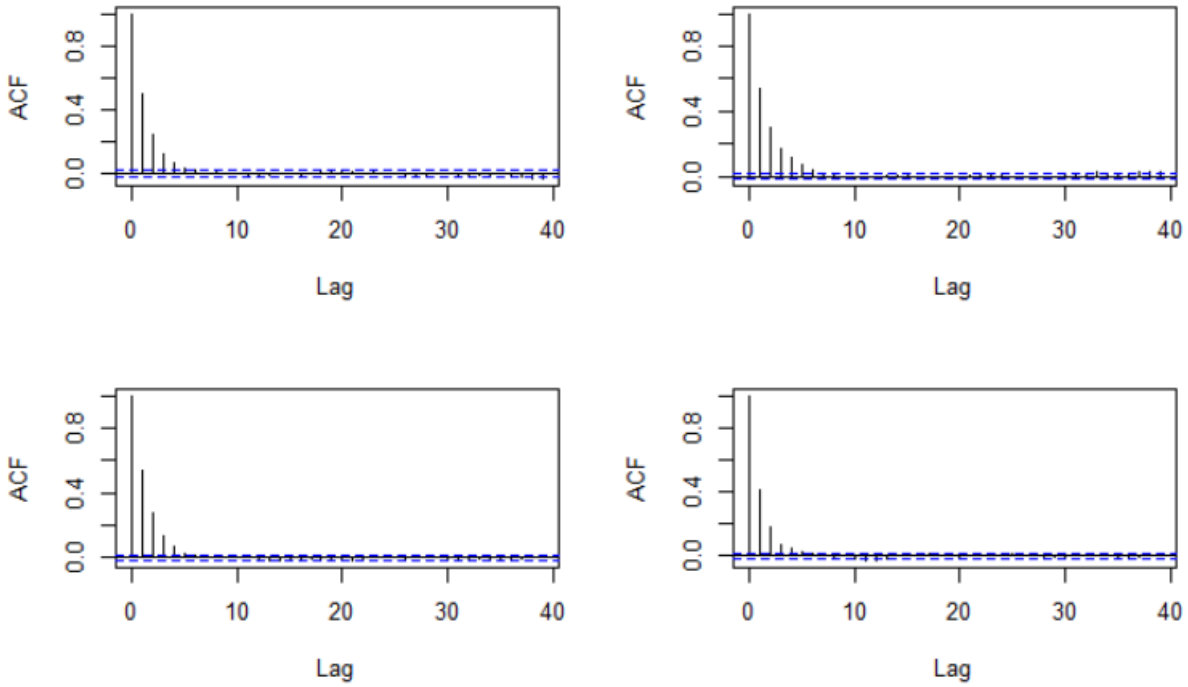


Figure 35: ACF plots of the thinned sample for each parameter under the classroom-household model. Top left q_H, top right q_1, bottom left q_2 and bottom right q

We can clearly see the positive effects of thinning with the ACF dying down for each parameter before the 10th lag. Now that we are happy with the autocorrelation in each parameter sample, we should once again calculate the posterior correlations between the samples. Let our parameter vector be $\boldsymbol{\theta} = (q_H, q_1, q_2, q)$, then

$$\boldsymbol{\rho_\theta} = \begin{bmatrix} 1 & -0.006998378 & -0.06475087 & -0.21006791 \\ -0.006998378 & 1 & -0.01017729 & -0.010050647 \\ -0.06475087 & -0.01017729 & 1 & -0.19891870 \\ -0.21006791 & -0.010050647 & -0.19891870 & 1 \end{bmatrix}$$

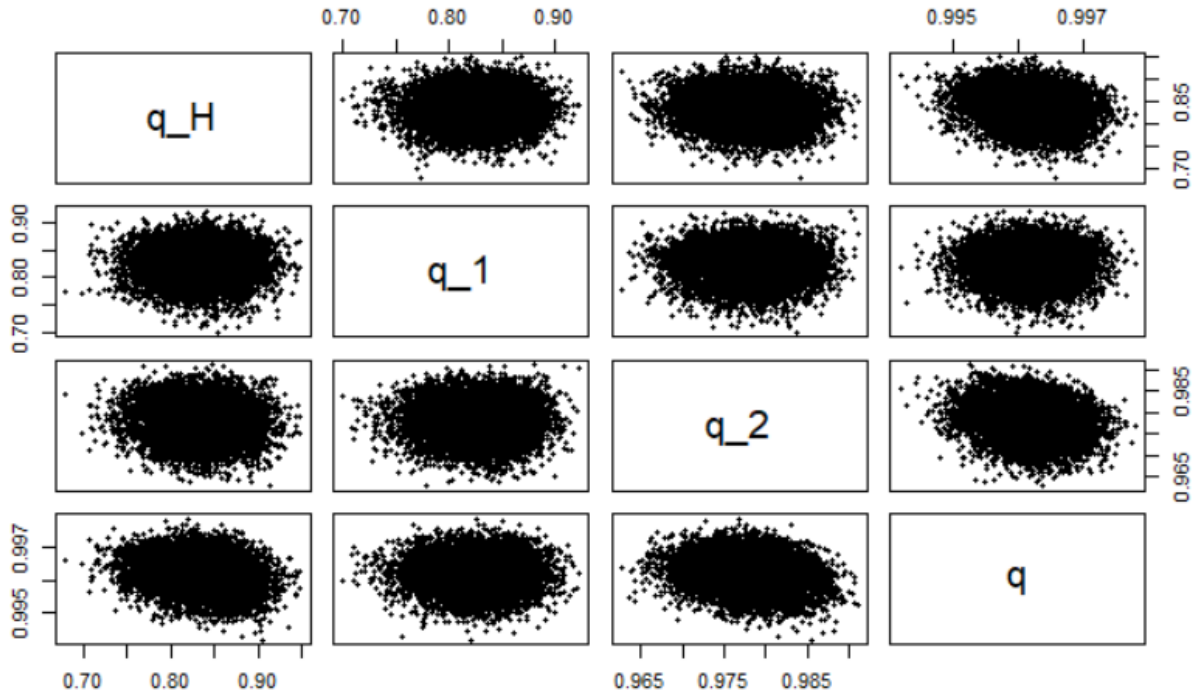


Figure 36: Pairs plot of the posterior samples under the classroom-household model

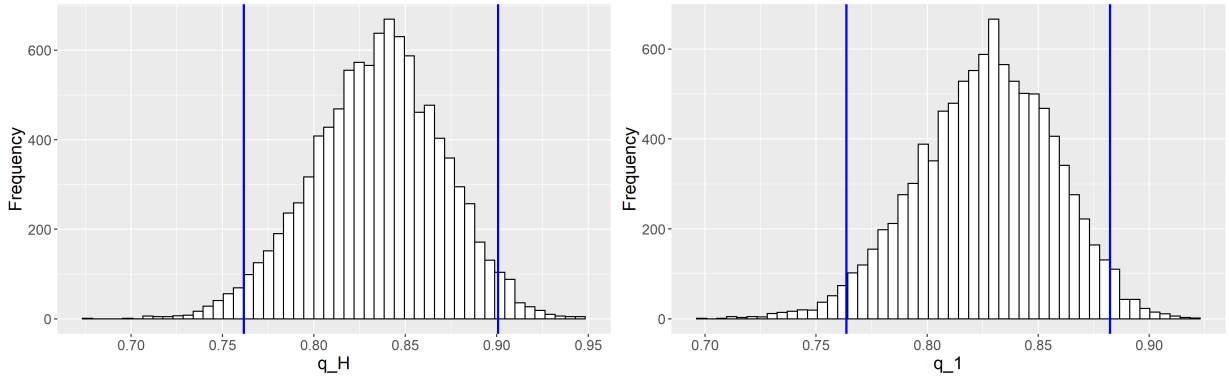
We can see that the majority of the posterior correlations are very small, with a few reaching approximately -0.2, in particular these are the correlations between q_H and q, and q and q_2. Neither of these values violate the rule of thumb outlined previously, and therefore we are confident that our samples will allow us to accurately and separately estimate each parameter. Now, we can average the thinned samples to estimate the posterior means of our parameters. We also calculate the 95% BCRs centred on these posterior means. Doing so, we find that,

$$\hat{q_H} = 0.8347213 \pm 0.06945016$$

$$\hat{q}_1 = 0.8263905 \pm 0.05937389$$

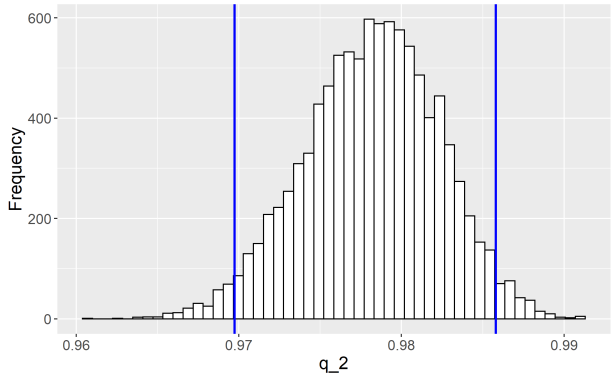
$$\hat{q}_2 = 0.9781636 \pm 0.008036094$$

$$\hat{q} = 0.9962155 \pm 0.0009354049$$

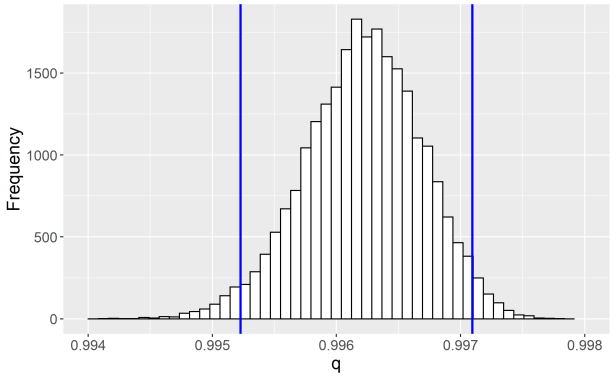


(a) Samples for q_H

(b) Samples for q_1



(c) Samples for q_2



(d) Samples for q

Figure 37: Histograms for each parameter samples under the classroom-household model with associated 95% BCRs

Our Bayesian parameter estimates for this model are once again similar to those we calculated through classical means, and running simulations with them produces results that show no significant improvement. Despite this, we are much more confident in these parameter estimates, and the fact that simulation results remain poor gives us more confidence in our previous conclusion. That is, that the addition of household transmission to the extended classroom model does not produce an accurate reflection, at least under this formulation, of the observed epidemic. Therefore, using these results, we can once again say that classroom transmission seemed to be much more important to the spread of measles and the severity of the epidemic than household transmission.

9.2.4 Summary of Parameter Estimates

We have detailed the process of producing Bayesian parameter estimates for the base, extended classroom and classroom-household models. We also underwent the same process for the initial classroom and the household model. Below is a summary table of the classical and Bayesian parameter estimates and confidence/credible intervals for each model.

Model	**Classical Estimates**	**Bayesian Estimates**	**% Difference**
Base	$q = 0.9929226 \pm 0.00100551$	$q = 0.9928855 \pm 0.001026197$	0.0037%
Initial Classroom	$q = 0.9952665 \pm 0.0008925209$, $q_c = 0.9646576 \pm 0.00670306$	$q = 0.9952084 \pm 0.0009904821$, $q_c = 0.9645587 \pm 0.009225594$	0.0058%, 0.01%
Extended Classroom	$q = 0.9951997 \pm 0.0009005345$, $q_1 = 0.8285551 \pm 0.06081507$, $q_2 = 0.976536 \pm 0.008173304$	$q = 0.99523 \pm 0.0009473393$, $q_1 = 0.8241714 \pm 0.06021979$, $q_2 = 0.9759522 \pm 0.008415046$	-0.003%, 0.53%, 0.59%
Household	$q = 0.9938498 \pm 0.000967601$, $q_H = 0.8581391 \pm 0.06535713$	$q = 0.9938429 \pm 0.0009795776$, $q_H = 0.8520551 \pm 0.001104022$	0.00069%, 0.71%
Classroom-Household	$q = 0.9961663 \pm 0.0008374808$, $q_1 = 0.8319394 \pm 0.06024162$, $q_2 = 0.9788626 \pm 0.007854487$, $q_H = 0.8475385 \pm 0.06552155$	$q = 0.9962155 \pm 0.0009354049$, $q_1 = 0.8263905 \pm 0.05937389$, $q_2 = 0.9781636 \pm 0.008036094$, $q_H = 0.8347213 \pm 0.06945016$	-0.0049%, 0.67%, 0.071%, 1.54%

9.3 MCMC Sensitivity Analysis

In Section 2.4 we discussed how a typical measles infection progresses and provided approximate ranges for the lengths of the various phases of such an infection. We then used this understanding to make model assumptions in Section 3 that allowed us to calculate the various population statistics that form the basis of our models. In particular, we assumed that the length of the exposed period E was 10 and the length of the eruption period d was 3. While this approach is perfectly valid, it is possible that these assumed values are not optimal for the observed epidemic. One way to address this potential problem is to manually run simulations under each model with different values of E and d,

and then pick the ones that produced the best results. This is referred to as *sensitivity analysis*. It is always a good idea to do a sensitivity analysis on any assumptions, if possible. However, we are given further motivation by the fact that the classroom-household model, which our exploratory analysis suggests should produce the best results, is performing poorly.

In the context of our problem, we have to be slightly careful; the length of the exposed and eruption periods vary from person to person, but only over a small range. Therefore, we may run into the problem of having to weigh up the optimal result with what makes sense given the extensive research on how a measles infection typically proceeds. In a similar vein, we are assessing the performance of the models based on simulations, and more specifically, a human interpretation of how well they have done. Therefore, it is difficult to decide what counts as the "best" result unless there are very stark differences in the outcomes.

Another, more objective approach to sensitivity analysis, is to use our MCMC framework where we can directly estimate E and d by updating their values inside the Metropolis-Hastings algorithm. That is, we treat them as additional parameters in our model, rather than assumed constants. However, we may run into a problem where different models produce different optimal values. This is nonsensical when we consider that these values are properties of the clinical features of measles, which shouldn't change based on the model we use. This is also a potential issue for the manual approach, but here it is far simpler to see. There would be no obvious way to choose between differing estimates, as long as they were in a sensible range, and so we make the decision to use this approach on our base model only. This will be computationally the least expensive and the most stable algorithm. If the resulting estimates are poor, i.e. very far out of the clinical range, then we can revisit this.

Now, focusing on the base model with original avoidance parameter q, we treat E and d as additional parameters, letting $\boldsymbol{\theta} = (q, E, d)$. Before we state the MCMC algorithm, we

need to modify our target distribution, i.e the posterior density,

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) \propto p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

Recall that in the base model, for our avoidance parameter q we used a non-informative uniform prior on $[0,1]$. We now need to introduce prior densities for both E and d. This time, unlike the avoidance parameters which are probabilities, E and d are positive integers which are otherwise unrestricted. Therefore, for both parameters, we propose the use of independent non-informative uniform priors on {1,2,...,M}, where M is a large integer. Because we have some knowledge about the approximate ranges of E and d we could attempt to use informative priors, such as more restrictive uniform distributions. However, this is largely unnecessary here as the resulting effect on the MCMC algorithm would be very minimal. Introducing these new priors, we get that

$$\begin{aligned}
\pi(\boldsymbol{\theta}|\boldsymbol{x}) &\propto p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\
&= \begin{cases} \frac{p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta})}{(M-1)^2} & \text{if} \quad q \in [0,1] \quad \text{and} \quad E, d \in \{1,2,...,M\} \\ 0, & \text{otherwise} \end{cases} \\
&\propto \begin{cases} p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}) & \text{if} \quad q \in [0,1] \quad \text{and} \quad E, d \in \{1,2,...,M\} \\ 0, & \text{otherwise} \end{cases}
\end{aligned}$$

We are not quite done yet. Recall that $p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta})$ is the probability model for the data, i.e. the likelihood function. The original base model has the following likelihood,

$$L(q) = \prod_t (q^{I(t)})^{S(t+1)}(1-q^{I(t)})^{(S(t)-S(t+1))}.$$

In order to calculate $S(t)$ and $I(t)$ for each time t, we need to know the values of E and d. Therefore, these population statistics and the likelihood, are now also functions of E and d,

$$L(q,E,d) = \prod_t (q^{I(t,E,d)})^{S(t+1,E,d)}(1-q^{I(t,E,d)})^{(S(t,E,d)-S(t+1,E,d))}.$$

Like in Section 9.2.1 we use a Normal distribution f to propose new values for q. All that is left to do is to choose how we propose new values for E and d. We do this by, for

example, proposing $d \to d+1$ or $d \to d-1$ with equal probability, being careful to reject moves when proposals occur that would be smaller than 1 or larger than M. Now, we can state the modified MCMC algorithm[8],

---

Algorithm 10: Base Model Sensitivity Metropolis-Hastings Algorithm

---

1. Choose a starting location $q = q_0, E = E_0, d = d_0$
2. Suppose at time t, we have $\boldsymbol{Q}_t = (q, E, d)$. For q, sample a candidate value $\hat{q}$ from a $N(q, 0.00125)$ distribution
3. Let $\hat{E} = E + 1$ or $\hat{E} = E - 1$ with equal probability. If $\hat{E} < 1$ or $\hat{E} > M$ reject the move and skip to the next iteration
4. Let $\hat{d} = d + 1$ or $\hat{d} = d - 1$ with equal probability. If $\hat{d} < 1$ or $\hat{d} > M$ reject the move and skip to the next iteration
5. Calculate the log-ratio

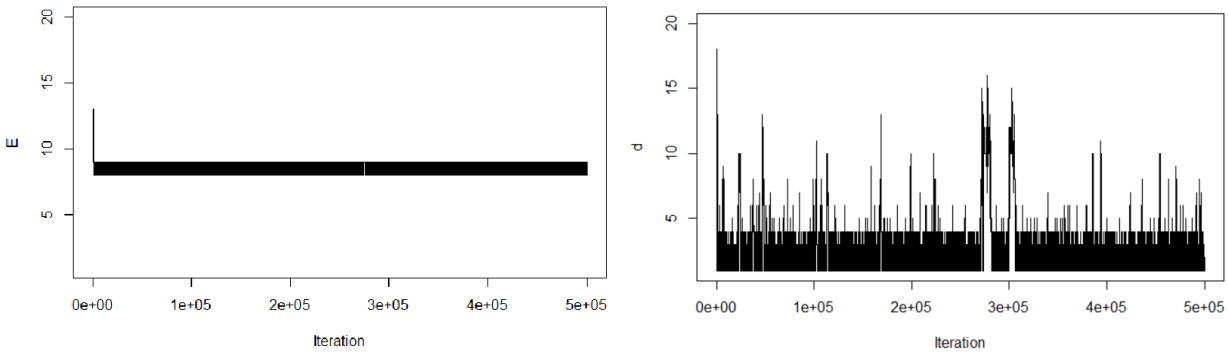
$$\alpha((q, E, d), (\hat{q}, \hat{E}, \hat{d})) = \log(L(\hat{q}, \hat{E}, \hat{d})) + \log(f(q)) - \log(L(q, E, d)) - \log(f(\hat{q}))$$

6. Generate a value u from a U(0,1) distribution
7. Accept the move if $\log(u) < \alpha((q, E, d), (\hat{q}, \hat{E}, \hat{d}))$. Otherwise we reject the move and stay at (q, E, d). That is, we set

$$\boldsymbol{Q}_{t+1} = \begin{cases} (\hat{q}, \hat{E}, \hat{d}), & \text{if } \log(u) < \alpha((q, E, d), (\hat{q}, \hat{E}, \hat{d})) \\ (q, E, d), & \text{otherwise} \end{cases}$$

Running this algorithm for 500000 iterations with initial values of $E = 10, d = 3$ and $q = 0.99$, gives the following results. Note that we are only interested in the samples of E and d here.

[8] When calculating the acceptance ratio in step 5 of the algorithm, the effect of the proposal distributions for E and d cancel out

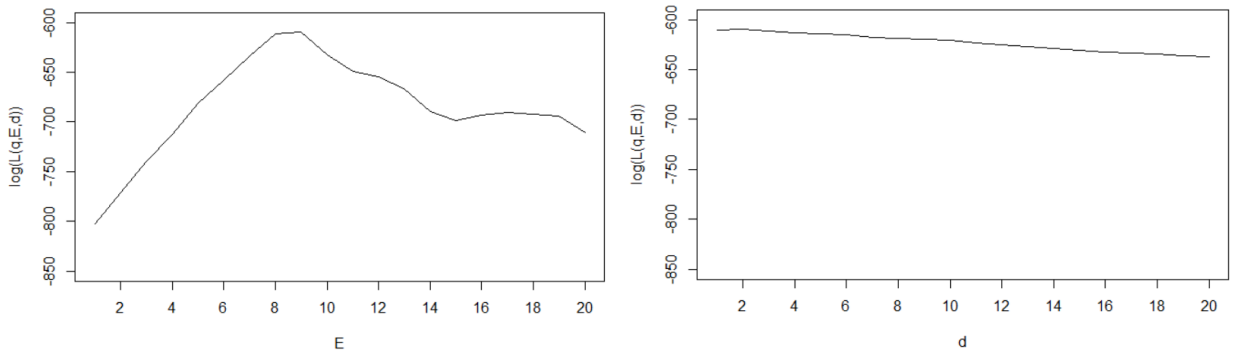


(a) Trace plot of E sample

(b) Trace plot of d sample

Figure 38: Trace plots of the full samples of each parameter of interest E and d under the base model

The behaviour of the sample for E is quite strange, the chain appears to be "stuck" moving between $E = 8$ and $E = 9$. Conversely, the samples for d look much more natural, with a wider range of values in the support being explored. Before we can be fully confident in these samples, we need to check if the behaviour we see is a problem or if it is reflecting the support of the log-likelihood accurately. To do this, we can plot the log-likelihood with one of d and E varying and with the rest of the parameters fixed at their estimate posterior mean. Finding the posterior means of our samples gives the parameter estimates $q = 0.99379$,$E = 8.615782$ and $d = 2.31661$, however noting that the new parameters must be integers, we use the posterior modes instead to get that $E = 9$ and $d = 2$. The value for d is slightly outside the approximate clinical ranges given in Section 2.4 but is still sensible for a measles infection.



(a) $d = 2, q = 0.99379$ fixed and E varying

(b) $E = 9, q = 0.99379$ fixed and d varying

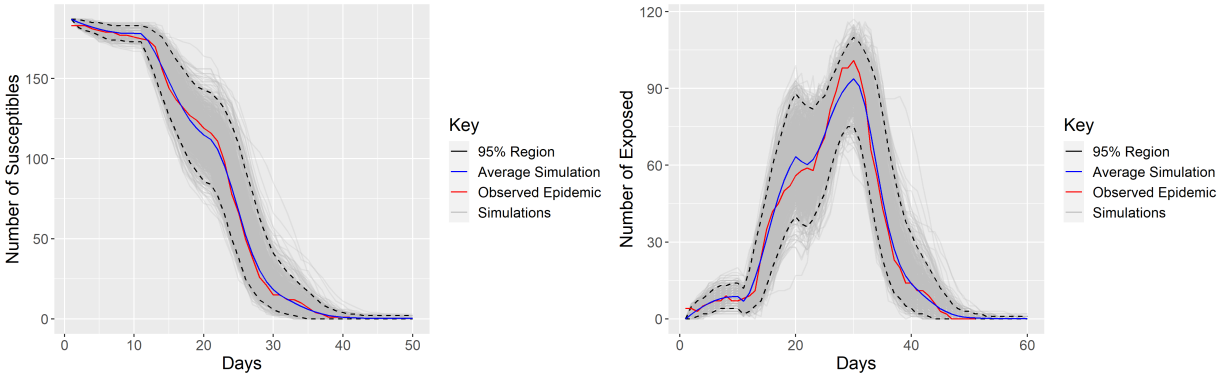
Figure 39: Plot of the base model log-likelihood with $q = 0.99379$ fixed and E or d varying

From Figure 39a we can see that the log-likelihood sharply drops either side of $E = 8$ or 9. This explains why the chain is only exploring these values; when proposals are made outside of this area, they are rejected due to the large decrease in the log-likelihood value. Conversely, in Figure 39b we can see that the largest values of the function occur when $d = 2$, however this time the function decreases much more slowly either side of this maximum and thus it is more likely that a wider range of proposals are accepted. This explains the behaviour of our chain and gives us confidence in the resulting parameter estimates[9].

9.3.1 Effect on Simulations

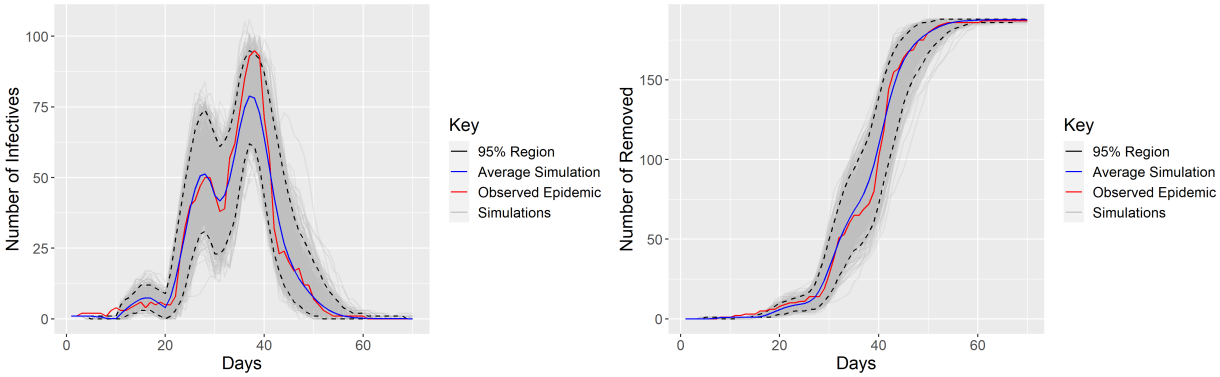
It is now of interest to move forward with the new assumption that $E = 9$ and $d = 2$, apply this to our different models, produce new parameter estimates and then run simulations. Using the Bayesian approach to parameter estimation in each case, we did not find any significant improvement or deterioration in the quality of the results from the classroom models or the base model. However, the household model and the classroom-household model saw a significant increase in the strength of the simulations produced, with the latter showing the largest improvement. We will focus on the results of the classroom-household model.

[9]We also looked at the ACF of of our samples; the autocorrelation was high and stayed for a large number of lags. This is due to the small range of values that our samples explore and does not indicate that the chain is mixing poorly here



(a) Susceptibles Plot

(b) Exposed Plot

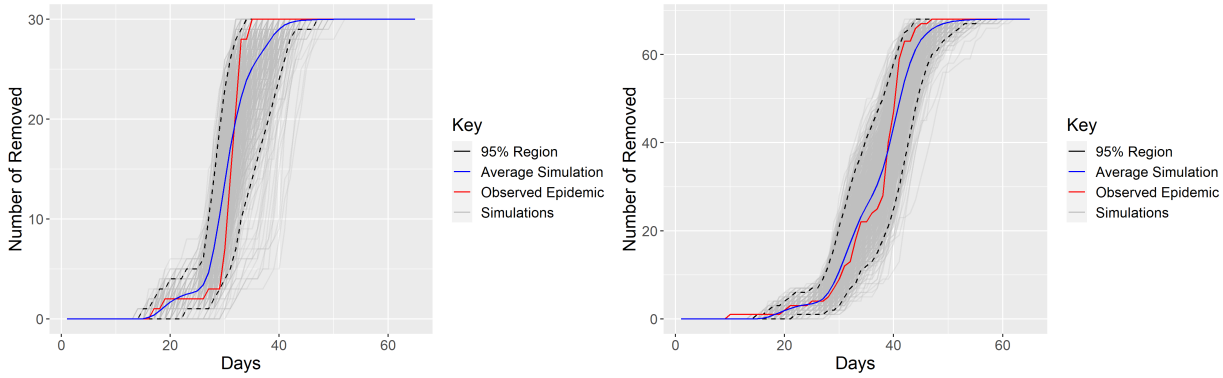


(c) Infectives Plot

(d) Removed Plot

Figure 40: Plot showing 1000 simulated epidemics under the classroom-household model with $q = 0.997019, q_1 = 0.8685684, q_2 = 0.9819253, q_H = 0.8117354$, $E = 9$ $d = 2$, the average simulation and the observed epidemic

We can see that the average simulation here is a vast improvement to the results we saw from the previous use of the classroom-household model in Section 8.2. The fit is extremely close, particularly when we look at the susceptible and removed population statistics in Figure 40a and 40d respectively. In Figure 40c we can see that the simulation falls short of the maximum number of infectives in the observed epidemic but the first wave is represented very accurately here; the most accurate of all the simulations thus far. Similarly, the results in Figure 40b for the exposed population statistic are also very strong.



(a) Class 1 Removed Plot

(b) Class 2 Removed Plot

Figure 41: Plot of the removed statistic $R(t)$ of each classroom under the classroom-household with $q = 0.997019, q_1 = 0.8685684, q_2 = 0.9819253, q_H = 0.8117354$, $E = 9$, $d = 2$, showing the 1000 simulated epidemics, the average simulation and the observed epidemic

The strength of the results continue here with very close fits to the observed epidemic inside classroom 1 and classroom 2. In particular, this is a large improvement for the results for classroom 2. The classroom-household model, using the new exposed and eruption period lengths, has produced results that are stronger than those from the extended classroom model from Section 6.2.2[10]. This is what we initially expected to find given that the model introduces another avenue of infection that more accurately reflects reality. In fact, the new parameter estimates $q = 0.997019, q_1 = 0.8685684, q_2 = 0.9819253, q_H = 0.8117354$ suggest that households are an environment that is more prone than classrooms to spreading measles. This is the opposite of what we found earlier where q_1 was smaller than q_H. Therefore, we should also inspect the transmission probabilities of this model to see if our previous conclusion that classroom transmission was more important than household transmission in the spread of the epidemic still holds.

Calculating the transmission probabilities under this model, with the new assumptions, we find that 39.5%, 32.5% and 24% of infections occurred directly due to general, classroom and household transmission respectively. Therefore, between household and classroom

[10]The results of the extended classroom model under the new assumptions were very similar to those in Section 6.2.2

transmission, we can still say that the latter was a bigger contributor to the spread of measles, however it is a much closer contest than what we saw earlier in Section 8.1. Restricting attention to just those individuals where classroom transmission was possible, i.e those old enough to attend school, we found that in classroom 1, 87%, and in classroom 2, 51%, of infections occurred solely due to classroom spread. This is a small decrease from previous results, however this still represents a very large proportion, particularly for the youngest children attending classroom 1.

10 Conclusions and Areas of Further Work

The goal of this dissertation was to analyse and compare the importance of different transmission pathways in the 1861 Hagelloch measles epidemic. Specifically, exploratory analysis showed that transmission via classroom attendance and household spread both appeared to be significant. By formulating different models which isolated or combined the effects of these transmission pathways, we initially found that the most accurate simulations were produced when we discarded the possibility of household transmission and focused just on classrooms. Of the two classroom models produced, the one that allowed the epidemic to spread inside the two classrooms at different rates was the stronger of the two, with the classroom attended by the younger children having a higher rate of infection.

Throughout the initial model fitting process we used classical statistical inference to estimate the model parameters. Small sample sizes meant that the asymptotic normal assumption of the MLE was called into question. Thus, particularly for the classroom-household model, we were unsure if poor results were due to lack of confidence in our parameter estimates. By applying a Bayesian approach with Markov Chain Monte Carlo methods, we found parameter estimates and credible regions that were very similar to those we found through classical means. This gave us confidence in our initial conclusion that the classroom-household model itself was somehow not an accurate representation of the infection dynamics of the observed epidemic. However, an MCMC sensitivity analysis on our model assumptions showed us that we were using non-optimal values for the length of the exposed and eruption periods. Using the new optimal values, we found that in actuality, the classroom-household model produced the strongest simulations. This makes intuitive sense and is what we expected as this model more closely reflects the available transmission pathways in the observed epidemic.

For each model we also calculated the proportion of infections that were a result of transmission via the different pathways. Following the sensitivity analysis, we found that when both household and classroom transmission were included, classroom transmission was a more significant factor in the spread of the epidemic, with infected individuals more likely

to have caught measles from an individual in their classroom than a member of their household. However, household transmission still represented a significant, but smaller, proportion of the infections. Restricting attention to just those individuals where classroom transmission was possible, i.e those old enough to attend school, we found that in classroom 1, 87%, and in classroom 2, 51%, of infections occurred solely due to classroom spread. This is a very large proportion, particularly for those children in classroom 1. These results lead us to conjecture that if classrooms were closed, the epidemic may not have been as severe as it turned out to be in reality. This keeps with the intuition that leaving areas open where large numbers of individuals interact results in a high rate of infection.

Epidemic modelling is a rich field of research and there are many areas where further work could be done. Keeping closely with the focus of the dissertation on analysing transmission pathways, we could extend the idea to modelling exactly who infected who. In the observed epidemic this information is available and was recorded to the best guess of Dr Pfeilsticker. This would be accomplished by calculating transmission probabilities and randomly assigning to each infection, based on these probabilities, an individual within the relevant sub-population. This would give us another measure by which we could compare the importance of transmission pathways. It is not clear how much new information this would give us as we have already calculated the exact theoretical transmission probabilities under each model for the observed epidemic.

Another area of significant interest would be to relax the assumption of constant infectivity. Of all the model assumptions made, this is the most tenuous, with many studies showing that an individuals infectivity varies significantly during the contagious period [14]. This would be done by choosing a function to model the change in infectivity over time, say $f(\tilde{t})$ on the range $(0, 1]$. Then, when an individual becomes infective, rather than adding 1 to the number of infectives $I(t)$, we could could add $f(\tilde{t})$. This would likely add significant computational cost, especially when simulating, as we would need to keep track of each individuals infectivity as time goes on. Studies have also shown that

an individual can be more or less infective depending on the viral load they are exposed to. Further, there is a link between severity of symptoms and viral load [15]. Dr Pfeilsticker recorded the maximum temperature suffered by each individual; we could choose a function $\lambda(T)$ to estimate this viral load based on the temperature T. Then, rather than adding $f(\tilde{t})$, we could add $\lambda(T)f(\tilde{t})$ to $I(t)$. It is difficult to know how successful this would be as there would likely be significant difficulty in choosing functions $\lambda(T)$ and $f(t)$ that accurately reflect the reality of a measles infection.

Finally, we note that our models do not allow for individuals to die as a result of the infection. Recall that in the observed epidemic, 12 of the 188 individuals ended up dying from measles following the eruption of their rash. It would be relatively simple to incorporate a death chance to each infection. We could then sample from the eruption periods of those individuals who died to replace the assumed length under our model. Similar to the inclusion of sampled prodromal periods, this may result in a better match to the dynamics of the observed epidemic. This time however, with such a small amount of data to sample from, it is possible that it could have the opposite effect as any outliers in the data could end up being over-represented in the simulations.

11 Appendix: R Code

Over 5000 lines of code was written in the programming language *R* to produce the results discussed throughout this report. What is shown here is a snippet of this code, chosen to best represent the techniques and algorithms used throughout the dissertation. In particular, the following code includes the simplest applications of our algorithms; the base model log-likelihood function, the base model epidemic simulation function (and accessory functions) and the Metropolis-Hastings algorithm applied to the base model log-likelihood.

```
base_log_likelihood = function(q){
  S = SUS_count
  I = INF_count
  value = 0
  for (t in 2:103){
    term = S[t]*I[t-1]*log(q)
    inf_term = (S[t-1] - S[t])*log(1-q^I[t-1])
    if(is.nan(inf_term) == FALSE && is.infinite(inf_term) == FALSE){
      term = term + inf_term
    }
    value = value + term
  }
  return(value)
}
```

Listing 1: Function which takes an input q and outputs the value of the base model log-likelihood

```
sample_x = function(I){
  x_samples = data$ERU - data$PRO
  x_values = sample(x_samples, size=I, replace=TRUE)
  return(x_values)
```

```
}
```

Listing 2: Accessory function which samples with replacement I prodromal periods from the observed prodromal periods

```
fill = function(start, I, T, d, epidemic){
  if(I == 0){
    return(epidemic)
    }
  x_values = sample_x(I)
  end = start + x_values + d
  for (j in 1:I){
    if (start> T | end[j] > T){break}
    epidemic[start:end[j],'I'] = epidemic[start:end[j],'I'] + 1
    if ((end[j] + 1)> T){break}
    epidemic[(end[j] + 1):T,'R'] = epidemic[(end[j] + 1):T,'R'] + 1
  }
  return(epidemic)
}
```

Listing 3: Accessory function which takes the following inputs: the current state of the epidemic, represented by a matrix of the population statistics $S(t), E(t), I(t)$ and $R(t)$ and denoted as "epidemic", the number of new infections I, the current time t= "start", the maximum epidemic length T and the assumed value of the eruption period d. It then "fills" in and outputs the epidemic matrix with each new infection accounted for

```
initialise = function(I, S, d, T){
  epidemic = matrix(0,T,4, dimnames = list(c(), c("S","E","I","R")))
  epidemic[1,] = c(S,0,0,0)

  epidemic = fill(1, I, T, d, epidemic)

  return(epidemic)
```

```
}
```

Listing 4: Accessory function which initialises and outputs a matrix to hold the population statistics $S(t), E(t), I(t)$ and $R(t)$ for a single simulation. Here, S is the number of initial susceptibles, d is the assumed eruption period length, T is the maximum epidemic length and I is the number of initial infectives

```
simulate = function(S, I, q, E, d, T){
  epidemic = initialise(I, S, d, T, epidemic)
  day = 1
  for (i in 2:T){
    day = day + 1
    epidemic[day,'S'] = rbinom(1, epidemic[day - 1,'S'],
                               q^epidemic[day-1,'I'])
    new_infections = epidemic[day - 1,'S'] - epidemic[day,'S']
    if ((day + E - 1) > T){break}
    epidemic[day:(day + E - 1),'E'] = epidemic[day:(day + E - 1),'E']
                                      + new_infections
    epidemic = fill((day+E), new_infections, T, d, epidemic)
  }
  return(epidemic)
}
```

Listing 5: Function which performs the epidemic simulation uner the base model. Requires the number of initial susceptibles and infectives, S and I respectively, the estimated value of q, the assumed exposed and eruption period lengths, E and d respectively, and the maximum epidemic length T. Outputs the epidemic after simulating T days

```
MH = function(q0,N){
  qS = numeric(N)
  q = q0
  nacc = 0
  ntry = 0
```

```
  for(i in 1:N){
    ntry = ntry + 1
    y = rnorm(1,q,0.00125)
    logratio = base_log_likelihood(y) + log(dnorm(q,y,0.00125)) -
               base_log_likelihood(q) - log(dnorm(y,q,0.00125))
    if(log(runif(1)) < logratio){
      q = y
      qS[i] = y
      nacc = nacc + 1
    }
    else{
      qS[i] = q
    }
  }
  return(list(sample=qS,accRate=nacc/ntry))
}
```

Listing 6: Base model Metropolis-Hastings algorithm. Requires an initial starting value for q, q_0 and the number of required iterations N

References

[1] A. Pfeilsticker, 1863. *Beiträge zur Pathologie der Masern mit besonderer Berücksichtigung der statistischen Verhältnisse*, Eberhard-Karls-Universität Tübingen

[2] P.J. Neal, G.O. Roberts, 2004. *Statistical inference and model selection for the 1861 Hagelloch measles epidemic*, Lancaster University

[3] 2018, *How long do babies carry their mother's immunity?*. [https://www.nhs.uk/common-health-questions/childrens-health/how-long-do-babies-carry-their-mothers-immunity/], [Accessed 6th April 2021]

[4] P. Ghosh, 2010. *Rinderpest virus has been wiped out, scientists say* . [https://www.bbc.co.uk/news/science-environment-11542653], [Accessed 6th April 2021]

[5] S.G. Cohen, 2008. *Measles and immunomodulation*, The Allergy Archives, Pioneers and Milestones

[6] Y. Furuse, H. Oshitani, A. Suzuki, 2010. *Origin of measles virus: divergence from rinderpest virus between the 11th and 12th centuries* , Tohoku University Graduate School of Medicine

[7] J.P. Byrne, 2008. *Encyclopedia of Pestilence, Pandemics, and Plagues*

[8] W.H. McNeill, 2010. *Most and probably all of the distinctive infectious diseases of civilization have been transferred to human populations from animal herds*. [https://web.archive.org/web/20091003164005/http://www.birdflubook.com/a.php?id=40], [Accessed 6th April 2021]

[9] M. Ludlow, S. McQuaid, D. Milner, 2014. *Pathological consequences of systemic measles virus infection*, The Journal of Pathology

[10] World Health Organisation, 2019. *Measles*. [https://www.who.int/news-room/fact-sheets/detail/measles], [Accessed 6th April 2021]

[11] NEWS DESK, 2020. *DRC: More Ebola and plague cases reported, End of measles epidemic declared*. [http://outbreaknewstoday.com/drc-more-ebola-and-plague-cases-reported-end-of-measles-epidemic-declared-74655/] [Accessed 6th April 2021]

[12] Centre for Disease Control and Prevention, 2020. *Measles*. [https://www.cdc.gov/vaccines/pubs/pinkbook/meas.html], [Accessed 6th April 2021]

[13] PHE Transmission Group , 2020. *Factors contributing to risk of SARS-CoV2 transmission associated with various settings*. [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/945978/S0921_Factors_contributing_to_risk_of_SARS_18122020.pdf], [Accessed 6th April 2021]

[14] M. Çevik, A. Ho, 2020. *COVID-19: when are you most infectious?*. [https://theconversation.com/covid-19-when-are-you-most-infectious-150760], [Accessed 7th April 2021]

[15] L. Geddes, 2020. *Does a high viral load or infectious dose make covid-19 worse?*. [https://www.newscientist.com/article/2238819-does-a-high-viral-load-or-infectious-dose-make-covid-19-worse/], [Accessed 7th April 2021]

[16] J. MacQueen, 1967. *Some methods for classification and analysis of multivariate observations*, The University of California

[17] R.P Brent 1973. *Chapter 4: An Algorithm with Guaranteed Convergence for Finding a Zero of a Function*, Algorithms for Minimization without Derivatives

[18] R.B Schnabel, J.E. Dennis 1996. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations* , Society for Industrial and Applied Mathematics

[19] T. Kypraios, 2020. *Statistical Inference*, University of Nottingham

[20] C. Robert, G. Casella 2011. *A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data*, Institute of Mathematical Statistics

[21] C. Fallaize, 2021. *Computational Statistics - Chapter III: Monte Carlo Methods*, University of Nottingham

[22] N. Metropolis, A.W. Rosenbluth, M.N Rosenbluth, A.H. Teller, 1953. *Equation of State Calculations by Fast Computing Machines* , Los Alamos Scientific Laboratory

[23] W.K Hastings, 1970. *Monte Carlo sampling methods using Markov chains and their applications*, University of Toronto

[24] J. Ellis, 2018. *A Practical Guide to MCMC Part 1: MCMC Basics* . [https://jellis18.github.io/post/2018-01-02-mcmc-part1/], [Accessed 20th April 2021]