DETERMINISTIC AND STOCHASTIC SIR MODEL: THEORY, BAYESIAN INFERENCE AND POSTERIOR PREDICTION FOR THE FINAL SIZE OF THE EPIDEMIC

AN ABSTRACT

SUBMITTED ON THE FIFTH DAY OF MAY, 2023

TO THE DEPARTMENT OF MATHEMATICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

OF THE SCHOOL OF SCIENCE AND ENGINEERING

OF TULANE UNIVERSITY

FOR THE DEGREE OF

MASTER OF SCIENCE

BY

LOC XUAN HUYNH

APPROVED:

SCOTT MCKINLEY, Ph.D.

CHAIRMAN

KYLE ZHAO, Ph.D.

TOMMASO BUVOLI, Ph.D.

DETERMINISTIC AND STOCHASTIC SIR MODEL: THEORY, BAYESIAN INFERENCE AND POSTERIOR PREDICTION FOR THE FINAL SIZE OF THE EPIDEMIC

AN ABSTRACT

SUBMITTED ON THE FIFTH DAY OF MAY, 2023

TO THE DEPARTMENT OF MATHEMATICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

OF THE SCHOOL OF SCIENCE AND ENGINEERING

OF TULANE UNIVERSITY

FOR THE DEGREE OF

MASTER OF SCIENCE

BY

LOC XUAN HUYNH

APPROVED:_____

SCOTT MCKINLEY, Ph.D.

CHAIRMAN

KYLE ZHAO, Ph.D.

TOMMASO BUVOLI, Ph.D.

Acknowledgements

I owe the utmost gratitude to the Tulane University Department of Mathematics. I am particularly grateful to Dr. Kyle Zhao, Dr. Victor Moll, and Dr. Ricardo Cortez, who, though not directly involved in my thesis, have been inspiring me to continue my education towards a Ph.D. Specially, I want to thank Dr. Kyle Zhao and Dr. Tommaso Buvoli for always taking the time to discuss many of my problems in their office, not only in Mathematics but also my personal challenges. I must also acknowledge the constant support and encouragement from my friends and family. Their unwavering belief in my abilities and their emotional support have been my backbone throughout this challenging period. I would also like to acknowledge Benjamin Stager for his help with LaTex. Lastly, but most importantly, I extend my deepest gratitude to Dr. Scott McKinley, my advisor. He was also the first one to encourage me to pursue a Ph.D. I truly could not have completed this thesis without his guidance, mentorship, and patience.

Dedication

To the profound world of mathematics.

Contents

List of Figures

Chapter 1

Introduction

Mathematical modeling has long been a valuable tool for studying the dynamics of infectious disease outbreaks. Among the most widely used models is the SIR (Susceptible-Infected-Removed) framework, which classifies a population into these three classes and uses analytical tools to study the dynamics of each classes.

This article provides a comprehensive study of the deterministic SIR model as well as a stochastic SIR model that accounts for randomness. The deterministic model, dating back to the pioneering work of Kermack and McKendrick in 1927, uses a system of differential equations to track the time evolution of the S, I, and R population groups. A key concept is the basic reproduction number $\mathcal{R}_0$, which determines if an outbreak will occur or die out and provides a tool to estimate the final size of the epidemic.

While the deterministic model provides a big-picture view, the stochastic SIR model captures the randomness in real-world problem. Formulated as a continuous-time Markov chain, it includes transition probability at each time step. Interestingly, even when $\mathcal{R}_0 > 1$, the stochastic model shows there is still a chance that only a minor outbreak happen.

The article then dives into statistical methods - namely, Bayesian inference and data augmentation techniques - to estimate the model parameters such as transmission and recovery rates from actual outbreak data.

Our main work explores a framework for sequential daily prediction of new cases and the final size of the outbreak. Both simulated datasets or real world datasets like Ebola and COVID-19 are analyzed using this methodology. We include the model's performance, predictions, and limitations of this simplified stochastic SIR model.

Chapter 2 covers the fundamental knowledge in continuous-time Markov Chain and provides the simple birth and death processes as an example. Chapter 3 introduces the deterministic and stochastic SIR models. Chapter 4 covers Bayesian parameter estimation using data augmentation on a smallpox outbreak dataset. Chapter 5 presents our main work on daily Bayesian prediction of cases and final size, applied to simulated data as well as real Ebola and COVID-19 data and discusses the limitations of the methodology.

We also include our R code - in order to help the reader better understand how we can simulate systems under different scenarios and analyze different datasets. All codes are open to the reader to use for their benefit **here**.

Chapter 2

Overview of continuous-time Markov Chain

In this chapter, we study the continuous-time Markov chain. The continuous-time Markov chain is fundamental to explore the stochastic dynamical system. In particular, we discuss the the properties of the continuous-time Markov chain, focusing on Poisson processes, followed by the simple birth and death chain, which will be used to calculate the probability of having an outbreak in stochastic SIR model in the next chapter. In this chapter, we also assume that the readers have fundamental knowledge in probability theory and discrete-time Markov chain.

A classical reference for stochastic processes in general and continuous-time Markov chain in particular is a book by Durrett [3]. We also consult a book by Allen [1] to introduce an example of a continuous-time Markov chain, namely simple birth and death process.

2.1 Exponential and Poisson distribution

We begin by reviewing the Exponential and Poisson distribution. We also introduce some of their important properties and connections to other types of distribution.

2.1.1 Exponential distribution

Definition 2.1.1. *A continuous random varible T is said to follow an exponential distribution with rate λ, written as $T \sim Exp(\lambda)$, if*

$$F(t) = P(T \leq t) = 1 - e^{-\lambda t}, \quad \textit{for all } t \geq 0,$$

where F is the cummulative distribution function of the random variable T.

Remark 2.1.2. *Let T be a random variable and $T \sim Exp(\lambda)$. Then, the expected value and variance of T is*

$$E(T) = \frac{1}{\lambda}, \quad Var(T) = \frac{1}{\lambda^2}.$$

We discuss the Lack of Memory Property, which is one of the most important results of the Exponential distributed random variable.

Theorem 2.1.3. *Let T be a random variable, $T \sim Exp(\lambda)$, and let $s, t > 0$, then*

$$P(T > t + s | T > t) = P(T > s). \tag{2.1.1}$$

In words, this property says that if we have been waiting for t units of time, then the probability that we wait s more units of time is the same as the probability that we wait s units of time right from the beginning.

This property can be proven as follow.

First, we note that

$$\{T > t + s\} \subseteq \{T > t\}$$

for all $t, s > 0$. By using the definition of conditional probability, we have

$$P(T > t+s|T > t) = \frac{P(\{T > t+s\} \cap \{T > t\})}{P(T > t)} = \frac{P(T > t+s)}{P(T > t)}$$
$$= \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = P(T > s).$$

Next, we discuss the exponential races.

Theorem 2.1.4. *Let $T_1, T_2, \ldots T_n$ be a sequence of i.i.d exponentially distributed random variables, $T_i \sim exp(\lambda_i)$, $1 \leq i \leq n$. Let $V = \min(T_1, T_2, \ldots, T_n)$ and I be the index of the smallest T_i. Then V and I are random variables, and*

$$P(V > t) = \exp\left(-\sum_{i=1}^{n} \lambda_i t\right),$$
$$P(I = i) = \frac{\lambda_i}{\sum_{i=1}^{n} \lambda_i},$$

for $t > 0$ and $1 \leq i \leq n$. Moreover, I and V are independent.

Directly from the definition of V, and by independence, for $t > 0$, we have

$$P(V > t) = P(\min(T_1, T_2, \ldots, T_n) > t) = P(T_1 > t, T_2 > t, \ldots, T_n > t)$$
$$= \prod_{i=1}^{n} P(T_i > t) = \prod_{i=1}^{n} e^{-\lambda_i t} = e^{-\sum_{i=1}^{n} \lambda_i t}.$$

Therefore, we can conclude that $V \sim \exp(-\sum_{i=1}^{n} \lambda_i)$.

We now prove the second property. Note that I is the discrete random variable, whose values are from 1 to n. Then for $1 \leq i \leq n$, we have

$$P(I = i) = P(T_i < T_1, T_i < T_2, \ldots T_i < T_n) = P(T_i < \min_{\substack{1 \leq j \leq n \\ j \neq i}} T_j).$$

Let $W = \min_{\substack{1 \leq j \leq n \\ j \neq i}} T_j$, and $\mu = \sum_{\substack{j=1 \\ j \neq i}}^{n} \lambda_j$. Moreover, let the density function of variable T_i is f_{T_i}, meaning that $f_{T_i}(x) = -\lambda_i e^{-\lambda_i x}$ for all $x > 0$. By the first property,

$W \sim \text{Exp}\,(\mu)$. By using the property of probability measure, we have

$$P(I=i) = \int_0^\infty f_{T_i}(t_i).P(t_i < W)dt_i = \int_0^\infty -\lambda_i e^{-\lambda_i t_i}.e^{-\mu t_i}dt_i$$
$$= \frac{\lambda_i}{\mu+\lambda_i}\int_0^\infty -(\lambda_i+\mu)e^{-(\lambda_i+\mu)t_i}dt_i = \frac{\lambda_i}{\mu+\lambda_i}.$$

Note that we have the last equality is because of the fact that $-(\lambda_i+\mu)e^{-(\lambda_i+\mu)t_i}$ is a density function, therefore must integrate to 1. Also, note that $\mu+\lambda_i = \sum_{i=1}^{n}\lambda_i$. We conclude that, for $i = 1, \ldots, n$,

$$P(I=i) = \frac{\lambda_i}{\sum_{i=1}^{n}\lambda_i}.$$

Now, we prove the independence of V and I using the definition. Let $t \leq 0$ and $1 \leq i \leq n$, we consider

$$P(V<t, I=i) = P(T_i < t, T_1 > t, \ldots, T_n > t) = \int_0^t f_{T_i}(x)\prod_{\substack{j=1\\j\neq i}}^{n} P(T_j > x)dx$$
$$= \int_0^t \lambda_i e^{-\lambda_i x}\prod_{\substack{j=1\\j\neq i}}^{n} e^{-\lambda_j x}dx = \frac{\lambda_i}{\sum_{i=1}^{n}\lambda_i}\int_0^t \sum_{i=1}^{n}\lambda_i e^{-\sum_{i=1}^{n}\lambda_i x}dx$$
$$= P(I=i).P(V<t).$$

The reason why we have the last inequality is because $V \sim \exp\left(\sum_{i=1}^{n}\lambda_i\right)$.

2.1.2 Poisson distribution

Definition 2.1.5. *A discrete random varible X is said to follow a Poisson distribution with rate λ, written as $X \sim Poisson(\lambda)$, if*

$$p(n) = P(X=n) = e^{-\lambda}\frac{\lambda^n}{n!}, \quad \textit{for all } n \in \mathbb{N},$$

where p is the probability mass function of the random variable X.

Remark 2.1.6. *Let X be a random variable and $X \sim Poisson(\lambda)$. Then, the expected value and variance of X is*

$$E(X) = \lambda, \quad Var(X) = \lambda.$$

Next, we discuss the connections between Poisson distribution with the binomial distribution.

Theorem 2.1.7. *Consider a binomial distribution with n trials with probability p of success on each trial. If n approaches infinity but the expected value remains the same (np is fixed), the binomial distribution converges to the Poisson distribution.*

Let $X \sim \text{Binomial}(n, p)$, where $n \in \mathbb{N}$ and $p \in (0, 1)$ and $\lambda = np$. Then for $k \in \{0, \ldots, n\}$

$$\begin{aligned} P(X = k) &= \binom{n}{k} p^k (1-p)^{n-k} = \frac{n(n-1)\ldots(n-k+1)}{k!} p^k (1-p)^{n-k} \\ &= \frac{(np)^k}{k!} \frac{n(n-1)\ldots(n-k+1)}{n^k} (1-p)^{n-k} \\ &= \frac{\lambda^k}{k!} \frac{n(n-1)\ldots(n-k+1)}{n^k} \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!} \frac{n(n-1)\ldots(n-k+1)}{n^k} \left(1 - \frac{\lambda}{n}\right)^{n} \left(1 - \frac{\lambda}{n}\right)^{-k}. \end{aligned}$$

Note that $\dfrac{\lambda^k}{k!}$ is independent of n, and

$$\begin{aligned} \frac{n(n-1)\ldots(n-k+1)}{n^k} = \frac{n}{n}.\frac{n-1}{n}\ldots\frac{n-k+1}{n} &\to 1, \\ \left(1 - \frac{\lambda}{n}\right)^{n} &\to e^{-\lambda}, \\ \left(1 - \frac{\lambda}{n}\right)^{-k} &\to 1^{-k} = 1 \end{aligned}$$

as n converges to infinity. Moreover, as n converges to infinity, the range of k gets

larger, $k \in \mathbb{N}$. Thus,

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad k \in \mathbb{N},$$

which is the probability mass function of a Poisson distribution with rate λ. Thus, the Binomial(n, p) distribution can be approximate as a Poisson(np) distribution as n gets large.

2.2 Continuous-time Markov chains

In this section, we introduce the continuous-time Markov chains. These are stochastic processes that have continuous time, $t \in [0, \infty)$, and discrete-valued states. Moreover, we show that the amount of time spending in one state before moving to the next state is exponentially distributied.

2.2.1 Definitions and notations

Let $\{X(t) : t \in [0, \infty)\}$ be a collection of random variables. Moreover, the random variables attain value in a finite set, $\{0, 1, 2, \dots N\}$, or an infinite set, $\mathbb{N}$.

Definition 2.2.1. *The stochastic process $\{X(t) : t \in [0, \infty) \in\}$ is called a continuous-time Markov chain if it satisfies the following condition:*
For any sequence of real numbers satisfying $0 \leq t_0 < t_1 < \cdots < t_n < t_{n+1}$, and for possible states $i_0, i_1, \dots, i_n, i_{n+1}$, we have

$$\begin{aligned} &P(X(t_{n+1}) = i_{n+1} | X(t_0) = i_0, X(t_1) = i_1, \dots, X(t_n) = i_n) \\ &\quad = P(X(t_{n+1}) = i_{n+1} | X(t_n) = i_n). \end{aligned}$$

Remark 2.2.2. *The latter condition in the definition is called the Markov property.*

In words, the probability of moving to a next state only depends on the value of

the most recent time and does not depend on the history of the process. From now on, unless stated otherwise, we assume that the state space is infinitely countable and take values in $\mathbb{N}$.

For $t \geq 0$, we denote the mass probability distribution of $X(t)$ as $\{p_i(t)\}_{i=0}^{\infty}$, where

$$p_i(t) = P(X(t) = i).$$

Denote the vector of probabilities as $p(t) = (p_0(t), p_1(t), \dots)^\top$. Since $X(t)$ attains value in $\mathbb{N}$, we have that

$$\sum_{i=0}^{\infty} p_i(t) = 1.$$

For $0 \leq s < t$, we define the relationship between the random variables $X(s)$ and $X(t)$ by the transition probabilities as follow

$$p_{ij}(s,t) = P(X(t) = j | X(s) = i),$$

for $i, j \in \mathbb{N}$.

Definition 2.2.3. *If the transition probabilities only depends on the length of interval, i.e. $t - s$, then they are called stationary or homogeneous transition probabilities; otherwise they are called nonstationary or nonhomogeneous transition probabilities. For simplicity, if the transition probabilities only depends on the length of interval, we write it as follow*

$$p_{ij}(t-s) = P(X(t) = j | X(s) = i) = P(X(t-s) = j | X(0) = i). \tag{2.2.1}$$

From now on, unless stated otherwise, we only consider homogeneous transition

probabilities. Denote the matrix of transition probabilities as

$$P(t) = (p_{ij}(t))_{i,j\in\mathbb{N}}.$$

As in the discrete case, the matrix of transition probabilities of the continuous-time Markov chain satisfies the Chapman-Kolmogorov Equation, which stated as

Theorem 2.2.4. *Let $s, t \geq 0$, and $i, j \in \mathbb{N}$,*

$$\sum_{k=0}^{\infty} p_{ik}(t)p_{kj}(s) = p_{ij}(s+t).$$

This can be written in matrix form

$$P(t)P(s) = P(s+t).$$

Remark 2.2.5. *The theorem states that in order to go from i to j in the time interval of the length $s+t$, we can go to some state k in the interval of the length s and then from that state k go to state j in the interval of the length t.*

We prove this theorem as follow

$$\begin{aligned}
p_{ij}(s+t) = P(X(s+t) = j|X(0) = i) &= \frac{P(X(s+t) = j, X(0) = i)}{P(X(0) = i)} \\
&= \sum_{k=0}^{\infty} \frac{P(X(s+t) = j, X(t) = k, X(0) = i)}{P(X(0) = i)} \\
&= \sum_{k=0}^{\infty} \frac{P(X(s+t) = j, X(t) = k, X(0) = i)}{P(X(t) = k, X(0) = i)} \frac{P(X(t) = k, X(0) = i)}{P(X(0) = i)} \\
&= \sum_{k=0}^{\infty} P(X(s+t) = j|X(t) = k, X(0) = i).P(X(t) = k|X(0) = i) \\
&= \sum_{k=0}^{\infty} P(X(s+t) = j|X(t) = k).P(X(t) = k|X(0) = i) \\
&= \sum_{k=0}^{\infty} p_{kj}(s).p_{ik}(t).
\end{aligned}$$

2.2.2 Generator matrix

Now, we construct the generator matrix Q, which describes the rate of change of the transition probabilities.

Consider the continuous-time Markov chain $\{X(t) : t \in [0, \infty)$, with the transition matrix $P(t)$. Moreover, assume that for $i, j \in \mathbb{N}$, $p_{ij}(t)$ is continuous and differentiable for $t \geq 0$. Note that at $t = 0$, they satisfy

$$p_{ij}(0) = 0, \quad j \neq i, \quad \text{and } p_{ii}(0) = 1.$$

The rate of change of the transition probability from i to j in a short time is then defined as follow

$$\begin{aligned} q_{ij} &= \lim_{\Delta t \to 0^+} \frac{p_{ij}(\Delta t) - p_{ij}(0)}{\Delta t} = \lim_{\Delta t \to 0^+} \frac{p_{ij}(\Delta t)}{\Delta t}, \quad i \neq j \\ q_{ii} &= \lim_{\Delta t \to 0^+} \frac{p_{ii}(\Delta t) - p_{ii}(0)}{\Delta t} = \lim_{\Delta t \to 0^+} \frac{p_{ii}(\Delta t) - 1}{\Delta t}. \end{aligned} \tag{2.2.2}$$

Note that for $i \in \mathbb{N}$, $\sum_{j=0}^{\infty} p_{ij}(\Delta t) = 1$ since we can go to any other state from state i. Following from this, we have that

$$p_{ii}(\Delta t) - 1 = -\sum_{\substack{j=0, \\ j \neq i}}^{\infty} p_{ij}(\Delta t) = -\sum_{\substack{j=0, \\ j \neq i}}^{\infty} [q_{ij} \Delta t + o(\Delta t)],$$

where the notation $o(\Delta t)$ is the Lauder order symbol. In general, we say that the function $f(\Delta t)$ is $o(\Delta t)$ means that as $\Delta t \to 0$,

$$\lim_{\Delta t \to 0} \frac{f(\Delta t)}{\Delta t} = 0.$$

From that, we derive the relationship between q_{ii} and q_{ij} as follow

$$q_{ii} = -\sum_{\substack{j=0, \\ j \neq i}}^{\infty} q_{ij}.$$

From the definition (2.2.2), we can define the matrix of transition Q as follow

Definition 2.2.6. *Let $P(\Delta t)$ be the transition matrix for the time interval Δt, then the matrix of transition Q is defined as*

$$Q = (q_{ij})_{i,j\in\mathbb{N}} = \begin{pmatrix} q_{00} & q_{01} & \dots \\ q_{10} & q_{11} & \dots \\ \dots & \dots & \dots \end{pmatrix} = \begin{pmatrix} -\sum_{i=1}^{\infty} q_{0i} & q_{01} & \dots \\ q_{10} & -\sum_{\substack{i=0,\\ i\neq 1}}^{\infty} q_{1i} & \dots \\ \dots & \dots & \dots \end{pmatrix}$$

$$= \begin{pmatrix} \lim_{\Delta t\to 0^+} \frac{p_{00}(\Delta t)-1}{\Delta t} & \lim_{\Delta t\to 0^+} \frac{p_{01}(\Delta t)}{\Delta t} & \dots \\ \lim_{\Delta t\to 0^+} \frac{p_{10}(\Delta t)}{\Delta t} & \lim_{\Delta t\to 0^+} \frac{p_{11}(\Delta t)-1}{\Delta t} & \dots \\ \dots & \dots & \dots \end{pmatrix} = \lim_{\Delta t\to 0^+} \frac{P(\Delta t)-I}{\Delta t},$$

where I is the matrix with same dimension as $P(\Delta t)$ but with ones on the diagonal and zeros on other components.

2.2.3 Kolmogorov Differential Equations

In this section, we construct the Kolmogorov Differential Equations. By using the Chapman-Kolmogorov, for $i, j \in \mathbb{N}$, we have that

$$p_{ij}(t+\Delta t) = \sum_{k=0}^{\infty} p_{ik}(t)p_{kj}(\Delta t).$$

Assume that the generator matrix Q exists, then we can rewrite this into

$$p_{ij}(t+\Delta t) = \sum_{k=0}^{\infty} p_{ik}(t)(\delta_{kj} + q_{kj}\Delta t + o(\Delta t)),$$

where δ_{jk} is the Kronecker symbol. By subtracting both side with $p_{ij}(t)$, dividing by Δt and using the fact that $\sum_{k=0}^{\infty} p_{ik}(t) = 1$, we derive

$$\frac{p_{ij}(t)-p_{ij}(t)}{\Delta t} = \sum_{k=0}^{\infty} p_{ik}(t)\left[q_{kj} + \frac{o(\Delta t)}{\Delta t}\right].$$

Letting $\Delta t \to 0$, we have that

$$\frac{dp_{ij}(t)}{dt} = \sum_{k=0}^{\infty} p_{ik}(t) q_{kj}. \tag{2.2.3}$$

Definition 2.2.7. *The equation* (2.2.3) *is known as the forward Kolmogorov differential equations, which can be expressed in matrix form as*

$$\frac{dP(t)}{dt} = P(t)Q.$$

By doing similarly, we can derive the backward Kolmogorov differential equations, which can be expressed in matrix form as

$$\frac{dP(t)}{dt} = QP(t).$$

2.2.4 Interevent Time and Stochastic Realizations

In a continuous-time Markov chain, assume that we start at state $X(0)$, we stay in that state for a random amount of time W_1. After that, it moves a new state $X(W_1)$. Then it stays at $X(W_1)$ for a random amount of time W_2 before moving to a new state at $X(W_2)$. By continue doing that, we define W_i is the random variable for the waiting time before the i-th jump. Set $W_0 = 0$, we then have the collection of random variables $\{W_i\}_{i=0}^{\infty}$, which can be referred as the waiting times of the process. By this, we can define $T_i = W_{i+1} - W_i$ as the interevent times between the jumps. We show that T_i is an exponentially distributed random variable.

Assume that the value of the state at the i-th jump is n. Moreover, we assume that

$$P(\text{moving to another state}) = \sum_{j \neq n} p_{nj}(\Delta t) = \alpha(n)\Delta t + o(\Delta t).$$

$$P(\text{staying at the same state}) = p_{nn}(\Delta t) = 1 - \alpha(n)\Delta t + o(\Delta t).$$

Let $t > 0$,

$$P(T_i > t) = P(W_{i+1} - W_i > t) = P(W_{i+1} > t + W_i) := G_i(t),$$

that is, the probability that the process remains in state n for a time interval $[W_i, W_i + t]$. The first property of $G_i(t)$ is that, $G_i(0) = P(T_i > 0) = 1$. For Δt small enough, we have that

$$\begin{aligned} & G_i(t + \Delta t) = G_i(t).p_{nn}(\Delta t) = G_i(t)(1 - \alpha(n)\Delta t + o(\Delta t)) \\ \Leftrightarrow & G_i(t + \Delta t) - G_i(t) = G_i(t)(-\alpha(n)\Delta t + o(\Delta t)) \\ \Rightarrow & \frac{G_i(t + \Delta t) - G_i(t)}{\Delta t} = G_i(t)\left(-\alpha(n) + \frac{o(\Delta t)}{\Delta t}\right) \end{aligned}$$

Let $\Delta t \to 0$, we obtain the ODE as follow

$$\frac{dG_i(t)}{dt} = -\alpha(n)G_i(t), \quad G_i(0) = 1.$$

The solution of the ODE is $G_i(t) = e^{-\alpha(n)t}$. Then $P(T_i \leq t) = 1 - G_i(t) = 1 - e^{-\alpha(n)t}$, which is the cumulative function of the exponentially distributed random variable with rate $\alpha(n)$. Thus T_i is the exponential random variable.

2.3 Simple Birth and Death Processes

2.3.1 Definition and the extinction probability

Consider the application of continuous-time Markov chain to biology. Let the initial state of the process is $X(0) = N$, which can be understood as the population of the beginning of the process. For small Δt, we then define the infinitesimal

transition probabilities as follow

$$p_{i,i+j}(\Delta t) = P(X(t+\Delta t) - X(t) = j | X(t) = i)$$
$$= \begin{cases} \mu i \Delta t + o(\Delta t), & j = -1, \\ \lambda i \Delta t + o(\Delta t), & j = 1, \\ 1 - (\mu + \lambda) i \Delta t + o(\Delta t), & j = 0, \\ o(\Delta t), & j \neq -1, 0, 1. \end{cases}$$

This means in a small time interval Δt, the process only move to the next greater state; from state i to state $i+1$ (with rate λi) or move the next smaller state; from state i to state $i-1$ (with rate μi). The backward Kolmogorov differential equations are

$$\frac{dp_i(t)}{dt} = \lambda(i-1)p_{i-1}(t) + \mu(i+1)p_{i+1}(t) - (\lambda+\mu)ip_i(t)$$
$$\frac{dp_0(t)}{dt} = \mu p_1(t),$$

for $i = 1, 2, \ldots$, with inital condition $p_i(0) = \delta_{iN}$.

For $i = 1, 2, \ldots$, by multiplying the differential equation by z^i and sum over i, we have

$$\begin{aligned} \frac{\partial \mathcal{P}(z,t)}{\partial t} &= \lambda \sum_{i=1}^{\infty} (i-1)p_{i-1}(t)z^i + \mu \sum_{i=0}^{\infty} (i+1)p_{i+1}(t)z^i - (\lambda+\mu)\sum_{i=0}^{\infty} ip_i(t)z^i \\ &= \lambda z^2 \frac{\partial \mathcal{P}(z,t)}{\partial z} + \mu \frac{\partial \mathcal{P}(z,t)}{\partial z} - (\lambda+\mu)z\frac{\partial \mathcal{P}(z,t)}{\partial z} \\ &= [\lambda z^2 + \mu - (\lambda+\mu)z]\frac{\partial \mathcal{P}(z,t)}{\partial z}, \end{aligned}$$

where $\mathcal{P}$ is the probability generating function and the initial condition is $\mathcal{P}(z,0) = z^N$.

By using the method of characteristic, we can solve for $\mathcal{P}(z,t)$.

$$\mathcal{P}(z,t) = \begin{cases} \left(\dfrac{e^{t(\mu-\lambda)}(\lambda z-\mu)-\mu(z-1)}{e^{t(\mu-\lambda)}(\lambda z-\mu)-\lambda(z-1)}\right)^N, & \text{if } \lambda \neq \mu, \\ \left(\dfrac{1-(\lambda t-1)(z-1)}{1-\lambda t(z-1)}\right)^N & \text{if } \lambda = \mu. \end{cases}$$

Plug $z = 0$ into $\mathcal{P}(z,t)$, we have the formula for $p_0(t)$, which is the probability that the population go to zero. We have

$$p_0(t) = \mathcal{P}(0,t) = \begin{cases} \left(\dfrac{\mu-\mu e^{(\mu-\lambda)t}}{\lambda-\mu e^{(\mu-\lambda)t}}\right)^N & \text{if } \lambda \neq \mu, \\ \left(\dfrac{\lambda t}{1+\lambda t}\right)^N & \text{if } \lambda = \mu. \end{cases}$$

Remark 2.3.1. *This is the probability of extinction until time t. To find the probability of extinction in the long run, we let $t \to \infty$,*

$$p_0(\infty) = \lim_{t\to\infty} p_0(t) = \begin{cases} 1 & , \text{ if } \lambda \leq \mu, \\ \left(\dfrac{\mu}{\lambda}\right)^N, & \text{if } \lambda > \mu. \end{cases}$$

This result will be used in the next chapter to calculate the probability that the outbreak in the stochastic SIR model happens.

2.3.2 Simulation

In this section, we simulate the birth and death process by using R. Assume that the initial population is 100. We simulate the process in two cases.

Case 1: $\lambda \leq \mu$

In the first case, we let $\lambda = 0.3$ and $\mu = 0.4$. The process is shown in the figure (2.1). We can see from figure (2.1) that the population is decreasing, and eventually will be 0. This is because the probability of extinction is 1.

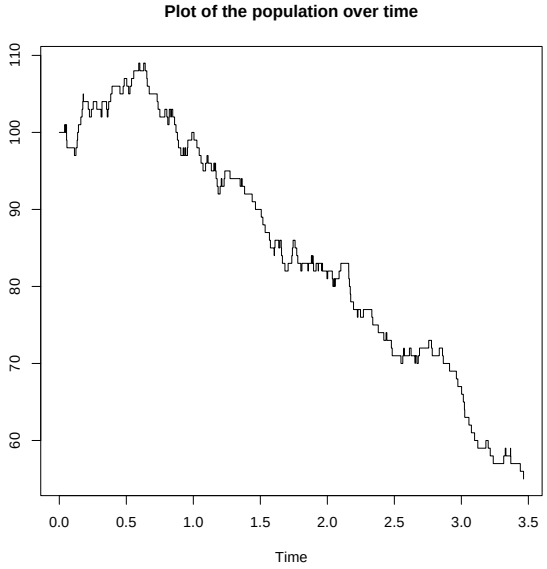

Figure 2.1: The stochastic birth and death process when $\lambda = 0.3$ and $\mu = 0.4$.

Case 2: $\lambda > \mu$

In the second case, we let $\lambda = 0.4$ and $\mu = 0.3$. The process is shown in the figure (2.2). We can see from figure (2.2) that the population is increasing. This

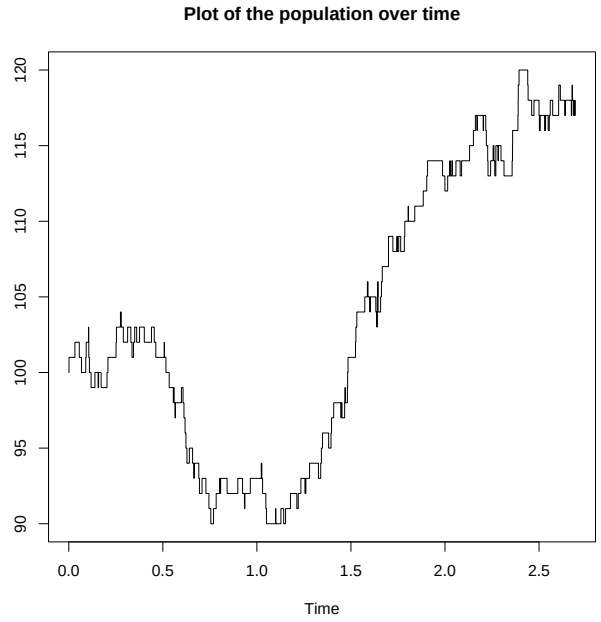

Figure 2.2: The stochastic birth and death process when $\lambda = 0.4$ and $\mu = 0.3$.

is because the probability of extinction is $\left(\frac{0.3}{0.4}\right)^{100} \approx 0$, so in most cases, the population increases.

Chapter 3

Deterministic and stochastic SIR model

In this chapter, we introduce the deterministic and stochastic SIR model. We also introduce the basic reproduction number $\mathcal{R}_0$. Moreover, in the case $\mathcal{R}_0 > 1$, we also show that the outbreak always happens for the deterministic model, however, that is not the case for the stochastic model. This is an introductory text referring to a book by Allen [1].

3.1 Deterministic SIR model

3.1.1 Introduction

The SIR model was discovered by Kermack and McKendrick in 1927 in order to model the dynamic of infectious diseases. Up to now, the model has been used to model the variety of diseases. In this section, we study the deterministic version of this model. Moreover, for the purpose of the thesis, we only consider the closed community, meaning that the total population remains constant throughout the process. Let N be the total population. For $t \geq 0$, let $S(t), I(t), R(t)$ be the susceptible susceptible, infected, and removed individuals at time t, respectively. For simplicity, we assume that $S(t), I(t), R(t)$ are continuous, meaning that they

can attain real value instead of integers only. Since N is the total population, $S(t) + I(t) + R(t) = N$ for all $t \geq 0$. Moreover, $S(t), I(t), R(t)$ is the solution of the ordinary differential equations:

$$\begin{aligned}
\frac{d}{dt}S &= -\beta\frac{SI}{N} \\
\frac{d}{dt}I &= \beta\frac{SI}{N} - \gamma I \\
\frac{d}{dt}R &= \gamma I,
\end{aligned}$$

where β is the diseases' transmission rate, and γ is the diseases' removal rate. For simplicity, we let β and γ to be constant positive numbers. We also assume that individuals meet any other individuals uniformly at random rate.

3.1.2 The basic reproduction number

Consider the SIR model, for all $t \geq 0$, β, S, I are non-negative, then $\frac{d}{dt}S(t) \leq 0$. Therefore, the proportion of susceptible individuals is always decreasing. By the same manner, the proportion of infected individual is always increasing.

Consider the dynamic of infected individuals, the number of infected individuals are increasing if $\frac{d}{dt}I(t) > 0$ or

$$\beta\frac{S(t)I(t)}{N} - \gamma I(t) > 0$$

Since $I(t) \geq 0$ for all $t \geq 0$, then $\beta\frac{S(t)}{N} - \gamma > 0$ or

$$\frac{\beta}{\gamma}\frac{S(t)}{N} > 1.$$

In the beginning of the process, $\frac{S(t)}{N} \approx 1$, so we consider the ratio $\frac{\beta}{\gamma}$. Let $\mathcal{R}_0 = \frac{\beta}{\gamma}$. This is called the basic reproduction number.

If $\mathcal{R}_0 > 1$, then the number of infected individuals increases, therefore we have a epidemic. On the other hand, if $\mathcal{R}_0 < 1$, then the number of infected individuals

decreases, thus the epidemic does not happen.

As time goes on, we define $\mathcal{R} := \mathcal{R}(t) = \frac{\beta}{\gamma}\frac{S(t)}{N}$. At each time, we calculate this quantity again. If $\mathcal{R} > 1$, then the number of infected individuals increase, and decrease otherwise.

3.1.3 Final size

Now, we study the dynamic of infected individuals and final size in the case of the epidemic, meaning that $\mathcal{R}_0 > 1$. We also assume that $I(0) = 1$ and $R(0) = 0$ as our initial condition. Then $S(0) = N - 1$. By using the chain rule, we have that

$$\frac{dI}{dS} = \frac{\beta\frac{SI}{N} - \gamma I}{-\beta\frac{SI}{N}} = -1 + \frac{\gamma N}{\beta S} = -1 + \frac{N}{\mathcal{R}_0 S} = -1 + \frac{1}{\mathcal{R}}. \tag{3.1.1}$$

Using this equation, we can derive a relationship between S and I. In the beginning of the process, we use the interpretation involving $\mathcal{R}_0$,

$$\frac{dI}{dS} < 0 \quad \Leftrightarrow \quad -1 + \frac{N}{\mathcal{R}_0 S} < 0 \quad \Leftrightarrow \quad \frac{S}{N} > \frac{1}{\mathcal{R}_0}.$$

In the case $\mathcal{R}_0 > 1$, then $\frac{1}{\mathcal{R}_0} < 1$. However, in the beginning of the process, $\frac{S}{N} \approx 1$, so $\frac{S}{N} > \frac{1}{\mathcal{R}_0}$, leading to $\frac{dI}{dS} < 0$. Since the number of susceptible individuals is always decreasing, then the number of infected individuals are increasing.

Now, we consider the latter interpretation in (3.1.1). As time goes on, $\frac{S(t)}{N}$ decreases, and there exists t such that $\mathcal{R} < 1$. By (3.1.1), $-1 + \frac{1}{\mathcal{R}} > 0$, so $\frac{dI}{dS} > 0$. Since S is always decreasing, then the number of infected individuals decrease as well. Moreover, the number of infected individuals decreases to 0.

In summary, for $\mathcal{R}_0 > 1$, in the beginning of the process, the number of infected individuals increase. Then it decreases after reach a certain amount of time.

Now, we study the final size of the outbreak. Denote $S(\infty) = \lim_{t\to\infty} S(t)$, $I(\infty) = \lim_{t\to\infty} I(t)$ and $R(\infty) = \lim_{t\to\infty} R(t)$. $R(\infty)$ is called the final size of the outbreak. Since the number of infected individuals decrease to 0, $I(\infty) = 0$. From (3.1.1),

we derive that

$$I(t) = -S(t) + \frac{N}{\mathcal{R}_0} \ln(S(t)) + C. \tag{3.1.2}$$

At $t = 0$, we substitute $S(0)$ and $I(0)$ into (3.1.2) to find C, denote C^*. Let $t \to \infty$, we have

$$S(\infty) - \frac{N}{\mathcal{R}_0} \ln(S(\infty)) = C^*. \tag{3.1.3}$$

We do not have the explicit formula for $S(\infty)$, but we can use numerical method to attain the value of $S(\infty)$. Then the final size of the outbreak is $R(\infty) = N - S(\infty)$. By replace $S(\infty) = N - R(\infty)$ into (3.1.3), we have

$$N - R(\infty) - \frac{N}{\mathcal{R}_0} \ln(N - R(\infty)) = C^*.$$

By using numerical method, we can plot the relation between $\mathcal{R}_0$ and $R(\infty)$ as in figure (3.1).

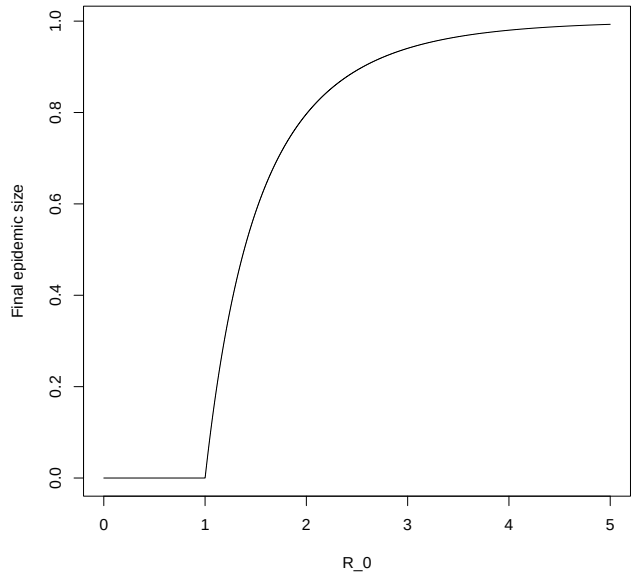

Figure 3.1: The relation between $\mathcal{R}_0$ and the final size of the outbreak.

Note that the final epidemic size in figure (3.1) is the proportion $\frac{R(\infty)}{N}$. The

figure shows that the epidemic does not happen if $\mathcal{R}_0 < 1$, and happens otherwise. Moreover, the figure shows that the higher $\mathcal{R}_0$ is, the higher the final epidemic size is. To explain this, recall that $\mathcal{R}_0 = \frac{\beta}{\gamma}$, where β is the transmission rate and γ is the removal rate. If $\mathcal{R}_0$ is large, then the transmission rate is larger than the removal rate, meaning that people gets infected faster than they are removed. This leads to the larger final size of the epidemic.

3.1.4 Simulation

Now, we simulate the deterministic SIR process in two cases. The first case is when $\mathcal{R}_0 < 1$ and the second case is when $\mathcal{R}_0 > 1$.

First case: $\mathcal{R}_0 < 1$

In this first case, we consider $N = 1000$, and $S(0) = 999, I(0) = 1, R(0) = 0$, $\beta = 0.15$ and $\gamma = 0.2$. Then the dynamics of the process is plotted in figure (3.2).

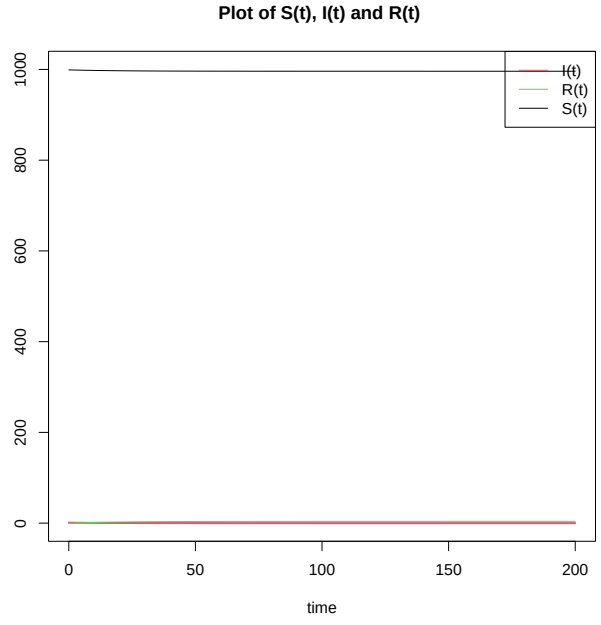

Figure 3.2: The dynamics of the deterministic SIR model when $\beta = 0.15$ and $\gamma = 0.2$

Remark 3.1.1. *In this case,* $\mathcal{R}_0 = \dfrac{\beta}{\gamma} = \dfrac{0.15}{0.2} = 0.75 < 1$. *Therefore, the infected individuals decrease over time. Thus, the outbreak does not happen. We can see in figure* (3.2) *that* $S(t), I(t), R(t)$ *almost stay the same through out the process.*

Second case: $\mathcal{R}_0 > 1$

In this first case, we consider $N = 1000$, and $S(0) = 999, I(0) = 1, R(0) = 0$, $\beta = 0.3$ and $\gamma = 0.2$. Then the dynamics of the process is plotted in figure (3.3).

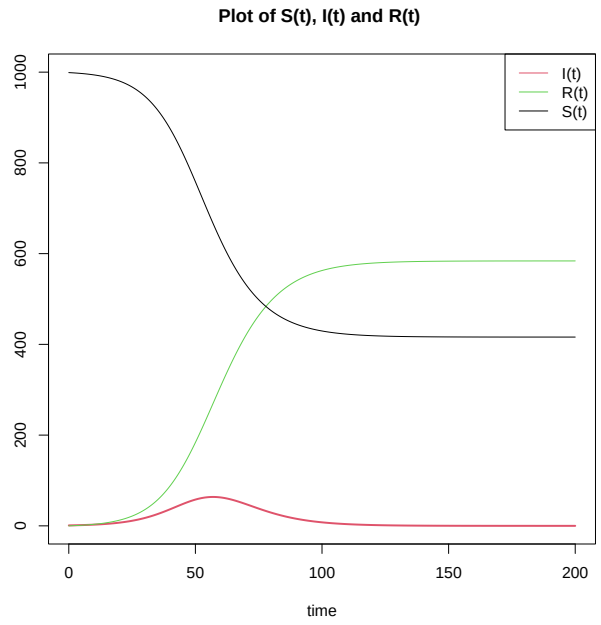

Figure 3.3: The dynamics of the deterministic SIR model when $\beta = 0.3$ and $\gamma = 0.2$

Remark 3.1.2. *In this case,* $\mathcal{R}_0 = \dfrac{\beta}{\gamma} = \dfrac{0.3}{0.2} = 1.5 > 1$. *Therefore, the infected individuals increase at first, then decrease to* 0 *over time. Thus, the outbreak happens, and the final size of the outbreak is approximately* 583. *We can also see in figure* (3.3) *that* $S(t)$ *decreases over time,* $R(t)$ *increases over time. Moreover,* $I(t)$ *increases in the beginning of the process, and then decreases to* 0 *afterward.*

3.2 Stochastic SIR Model

3.2.1 Introduction

Now, we consider the stochastic SIR model. For simplicity, we still consider the community with constant population N. For $t \geq 0$, we consider the process $\{S(t), I(t), R(t), t \in [0, \infty)\}$ as a continuous-time Markov chain. Since $S(t) + I(t) + R(t) = N$, then $R(t)$ depends on $S(t), I(t)$. Therefore, it is enough to consider $S(t)$ and $I(t)$. For the small interval of time $\Delta t > 0$, denote $\Delta S(t) = S(t+\Delta t) - S(t), \Delta I(t) = I(t+\Delta t) - I(t)$. We then define the transition probability to be

$$
\begin{aligned}
&P(\Delta S(t) = i, \Delta I(t) = j | S(t), I(t)) \\
&= \begin{cases}
\beta \dfrac{S(t)I(t)}{N} \Delta t + o(\Delta t), & (i,j) = (-1,1), \\
\gamma I(t) \Delta t + o(\Delta t), & (i,j) = (0,-1), \\
1 - \left[\beta \dfrac{S(t)I(t)}{N} + \gamma I(t) \right] \Delta + o(\Delta t), & (i,j) = (0,0) \\
o(\Delta t), & \text{otherwise.}
\end{cases}
\end{aligned}
$$

In short, we only allow one step happens in the short amount of time Δt. It is either a susceptible individual got infected and move to a infected group, or an infected individual is removed from the process, or nothing happens at all. Moreover, if $\Delta I(t) = -1$, then $\Delta R(t) = 1$.

For $t \geq 0$, let the pair $\{S(t) = i, I(t) = j\}$ be the state of the process. Ì $i \in \{0, 1, 2, \ldots, N\}$ then $j \in \{0, 1, 2, \ldots, N - i\}$ so that $i + j \leq N$.
For example, if $i = 0$, then j can be chosen from $0, 1, 2, \ldots, N$. If $i = 1$, then j can be chosen from $0, 1, 2, \ldots, N-1$. Therefore, the possible states in the process is $\sum_{k=1}^{N+1} k = \dfrac{(N+1)(N+2)}{2}$.

We can also derive the backward Kolmogorov equations for the transition prob-

abilities. By definition, for $i \in \{0, 1, 2, \ldots, N$ and $j \in \{0, 1, 2, \ldots, N - i\}$, we have

$$\begin{aligned}\frac{dp_{(i,j)}(t)}{dt} &= \frac{\beta}{N}(i+1)(j-1)p_{(i+1,j-1)}(t) \\ &\quad + \gamma(j+1)p_{(i,j+1)}(t) - \left[\frac{\beta}{N}ij + \gamma j\right] p_{(i,j)}(t).\end{aligned}$$

If there is any (i, j) that does not belong to the domain, $p_{(i,j)} = 0$.

3.2.2 Probability of no epidemic

We now consider the early stage of the process. First, assume that the initial state is $(S(0), I(0)) = (s_0, i_0)$, $s_0 \geq 0, i_0 > 0$ and $s_0 + i_0 = N$. Moreover, in reality, the epidemic start with a small number of infected individuals, so $s_0 \approx N$. Therefore, the rate of having new infected individuals is approximately β, and the rate of removing from the process is approximately γ. Thus, in the beginning of the process, the dynamics of infected individuals can be understood as the birth death process, where the initial population of the process is i_0, β is the birth rate and γ is the death rate. By using the result for the birth and death rate, if $\beta \leq \gamma$ (or $\mathcal{R}_0 \leq 1$), then the probability of extinction (meaning that there is no infected individual) is 1. On the other hand, if $\beta > \gamma$ (or $\mathcal{R}_0 > 1$), the probability of extinction is $\left(\frac{\gamma}{\beta}\right)^{i_0}$. In epidemiology, we call those occurrences as minor outbreak.

In conclusion, the result can be summarize using the $\mathcal{R}_0$ notation as

Theorem 3.2.1. *In the stochastic SIR model, let the initial number of infected individuals be $I(0) = i$. Then*

$$P(\textit{minor outbreak}) = \begin{cases} 1, & \mathcal{R}_0 \leq 1 \\ \left(\frac{1}{\mathcal{R}_0}\right)^i, & \mathcal{R}_0 > 1. \end{cases}$$

Remark 3.2.2. *In this subsection, we show that if $\mathcal{R}_0 > 1$, in the stochastic model, the epidemic might not happen, with probability $\left(\frac{1}{\mathcal{R}_0}\right)^i$, whereas in the*

deterministic model, the epidemic always happens.

Remark 3.2.3. *We can also see that the bigger the initial number of infected individuals is, the smaller the probability that the minor outbreak happens (since* $\frac{1}{\mathcal{R}_0} < 1$*). This makes sense in reality, since the more infected individuals you have, the easier they transmit the disease to other individuals.*

3.2.3 Simulation and the final size

In this subsection, we simulate the stochastic SIR process. For $t \geq 0$, let $S(t), I(t), R(t)$ be the susceptible, infected and removed individuals at time t, respectively that follows the stochastic SIR process. Let N be constant population of the community. We simulate the SIR process by using R.

First, to simulate the process, we need to calculate the probability that the process move to another state. Assume that the process is at state (i, j), meaning that $(S(t), I(t)) = (i, j)$. Then, let T_1 be the time that the process moves to the state $(i-1, j+1)$ and T_2 be the time that the process moves to that state $(i, j-1)$. Since this is also a continuous-time Markov chain, then $T_1 \sim \text{Exp}\left(\frac{\beta}{N}ij\right)$ and $T_2 \sim \text{Exp}(\gamma j)$. Denote T be the time that the process move to the next state. By the exponential races, $T \sim \exp\left(\frac{\beta}{N}ij + \gamma j\right)$. Moreover, let p the probability that the process move to the state $(i-1, j+1)$. Then by the exponential races, we have

$$p = \frac{\frac{\beta}{N}ij}{\frac{\beta}{N}ij + \gamma j} = \frac{\frac{\beta}{N}i}{\frac{\beta}{N}i + \gamma}.$$

This follows that the probability that the process moves to the state $(i, j-1)$ is

$$1 - p = \frac{\gamma j}{\frac{\beta}{N}ij + \gamma j} = \frac{\gamma}{\frac{\beta}{N}i + \gamma}.$$

Now, we simulate the stochastic SIR process using R. Next, we calculate the final size of the stochastic process in the case of $\mathcal{R}_0 > 1$ based on our simulation.

Case 1: $\mathcal{R}_\prime < 1$

In this first case, we consider $N = 1000$, and $S(0) = 999, I(0) = 1, R(0) = 0$, $\beta = 0.15$ and $\gamma = 0.2$. Then the dynamics of the process is plotted in figure (3.4).

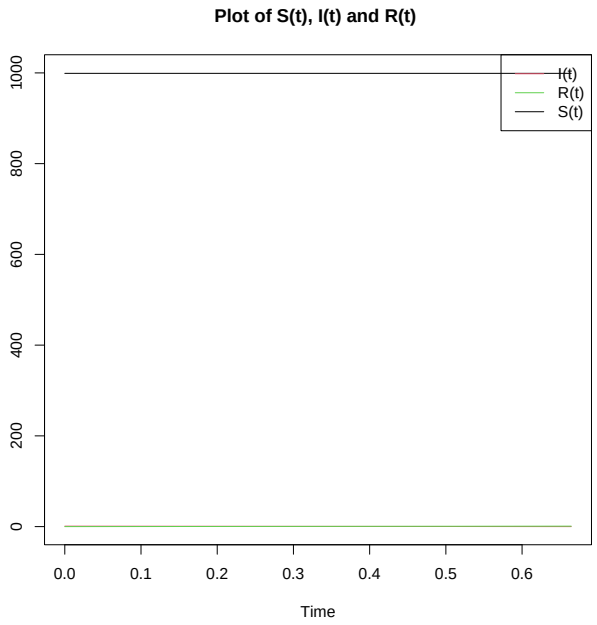

Figure 3.4: The dynamics of the stochastic model when $\beta = 0.15$ and $\gamma = 0.2$

Remark 3.2.4. *In this case,* $\mathcal{R}_0 = \dfrac{\beta}{\gamma} = \dfrac{0.15}{0.2} = 0.75 < 1$. *Therefore, the probability that the outbreak does not happen is* 1. *We can see in figure* (3.4) *that* $S(t), I(t), R(t)$ *almost stay the same through out the process.*

Case 2: $\mathcal{R}_\prime >$

In this first case, we consider $N = 1000$, and $S(0) = 999, I(0) = 1, R(0) = 0$, $\beta = 0.3$ and $\gamma = 0.2$. Then the dynamics of the process is plotted in figure (3.5).

Remark 3.2.5. *In this case,* $\mathcal{R}_0 = \dfrac{\beta}{\gamma} = \dfrac{0.3}{0.2} = 1.5 > 1$. *Therefore, the probability that the outbreak does not happen is* $\dfrac{1}{\mathcal{R}_0} = 0.67$. *We can see in figure* (3.5) *that the dynamics of* $S(t), I(t), R(t)$ *are almost the same from the deterministic case. The final size of the outbreak in this case is* 575.

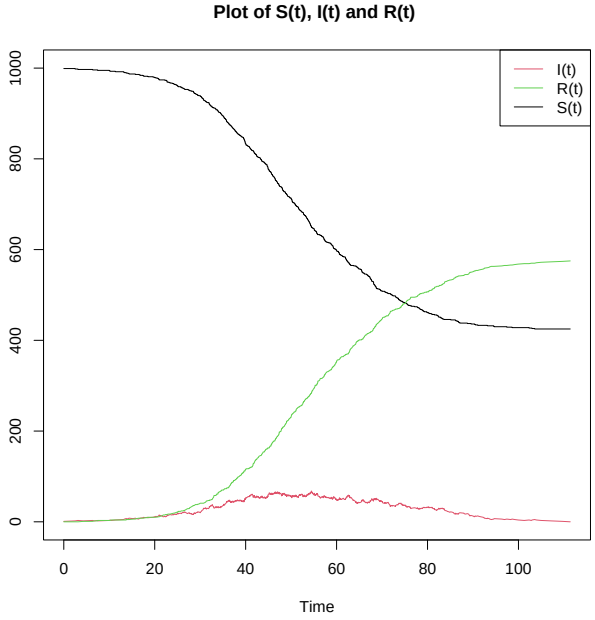

Figure 3.5: The dynamics of the stochastic model when $\beta = 0.3$ and $\gamma = 0.2$ when the major outbreak happens

However, if you run the simulation many times, you may recognize that there is a case where the minor outbreak happens. This is due to the fact that the probability does not happen is 0.67. Figure (3.6) shows the simulation in which the minor outbreak happens.

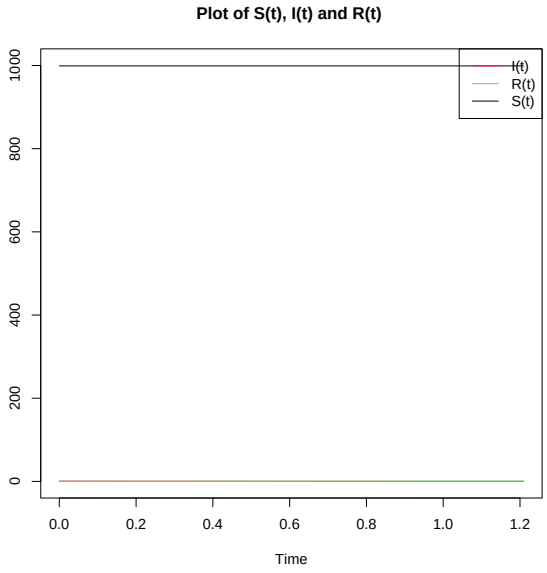

Figure 3.6: The dynamics of the stochastic model when $\beta = 0.3$ and $\gamma = 0.2$ when the minor outbreak happens

To study this further, we run the simulation multiples time, and collect the final size of the out break each time. By doing this, we can attain the distribution of the final size of the outbreak. We can also fine the mean of the final size. The distribution of the final size of the outbreak is shown below. The mean value of the final size is 195. In figure (3.7), we can see that the distribution is bimodal

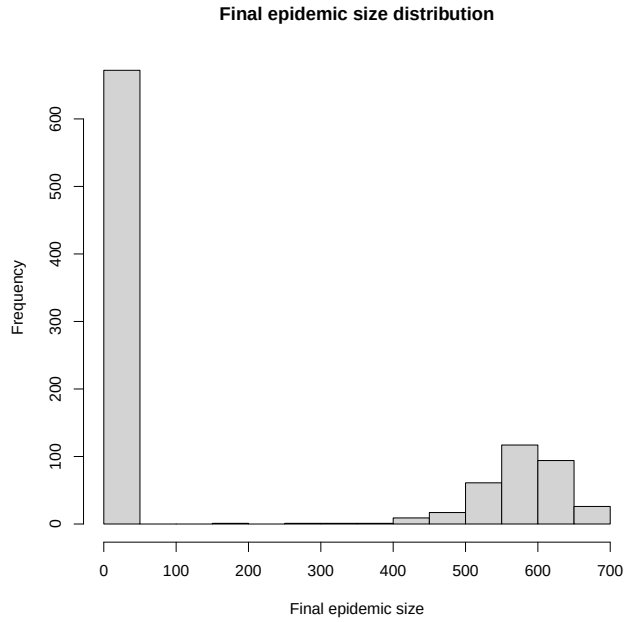

Figure 3.7: The distribution of the final size of the outbreak.

since there are two local peaks. The first one is for the minor case, and the second one is for the major case. Since we only have one initial infected individual, the probability that the minor outbreak happens is very high, 0.67. In face, there are over 600 simulations (over 1000 simulations) whose final epidemic size is less than 200. This also affects the mean of the final size so that it is only 195, which is less than the final size of the deterministic model. To see what is the mean of the final size, we remove the minor outbreaks and only keep the major outbreaks. Figure (3.8) show the distribution of the final size of the major outbreaks. The mean value of the final size is 578.

Note that the mean value of the final size if approximately 578, which is close to the final size of the outbreak in the deterministic model, 583.

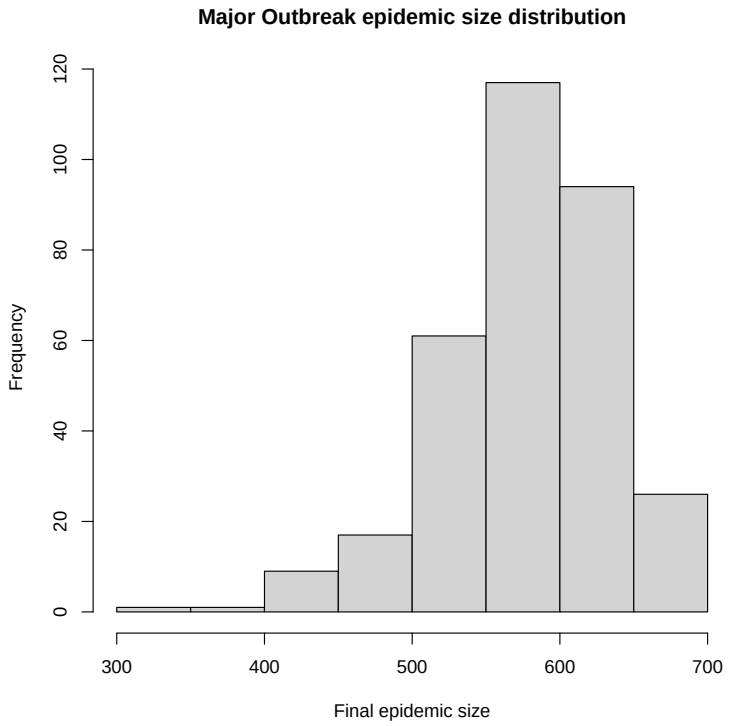

Figure 3.8: The distribution of the major final size of the major outbreak.

Chapter 4

Bayesian parameter estimation with data augmentation in the stochastic SIR model

In this chapter, we consider the Abakaliki smallpox data, which is obtained from a smallpox outbreak in a closed community of 120 individuals in Abakaliki, Nigeria. In real life, the disease has a latent period. However, for simplicity, we use the stochastic SIR model to estimate the transmission rate and removal rate of the disease. In particular, we first introduce the data, then introduce the likelihood function of the data. Next, we discuss the use of Bayesian estimation by using Markov Chain Monte Carlo (MCMC) and data augmentation algorithms. This is given by O'Neill [5], O'Neill [6] and O'Neill and Roberts [7].

4.1 Data and the likelihood function

4.1.1 Data interpretation

First, we introduce the dataset. This dataset was compiled from an outbreak of smallpox in a close community of 120 individuals in Abakaliki, Nigeria. The data consist of 29 inter-removal times between detection of cases, measured in days.

In the beginning, there is one initial case. Then, there are 29 subsequent cases, making a total of 30 cases. The data consists of 29 inter-removal times, measured in days

$$13, 7, 2, 3, 0, 0, 1, 4, 5, 3, 2, 0, 2, 0, 5, 3, 1, 4, 0, 1, 1, 1, 2, 0, 1, 5, 0, 5, 5.$$

A zero indicates that the a new removed case happened the same day as for the preceding case. We also assume that the first individual became infectious at time 0 and will be removed at time 14 (days) (so this agree with the interpretation made in [ref]). Then the 30 removal times (in days) are

$$\begin{aligned} &14, 27, 34, 36, 39, 39, 39, 40, 44, 49, 52, 54, 54, 56, 56 \\ &61, 64, 65, 69, 69, 70, 71, 72, 74, 74, 75, 80, 80, 85, 90. \end{aligned} \tag{4.1.1}$$

So the total duration of the outbreak is thus $T = 90$ days. So in this dataset, we only know the initial infected time (at time 0), and the removal times.

4.1.2 Construction of the likelihood function

To construct the likelihood function, we first assume that the data is complete, meaning that we know every time when an individual got infected and removed. In general, assume that there are N individuals in a closed community, and n cases in total. Additionally, we assume that each event, one of three outcomes occurs: a person becomes infected, an infected individual is removed from the infection process, or there are no changes. Let $\mathbf{i} = (i_1, i_2, \ldots, i_n)$ be the times when an individual got infected, and $\mathbf{r} = (r_1, r_2, \ldots, r_n)$ be the times when an individual got removed. Moreover, denote T to be the final removal time, so $r_n = T$. Moreover, by previous assumption, then $N = 120$, $n = 30$, $i_1 = 0$ and

$T = r_n = 90$. By O'Neill [5], the likelihood function is

$$f(\mathbf{i},\mathbf{r}|\beta,\gamma) = \left(\prod_{j=2}^{n} \frac{\beta}{N} S(i_j-)I(i_j-)\right)\left(\prod_{j=1}^{n} \gamma I(r_j-)\right) \times \exp\left(-\int_0^T \frac{\beta}{N} S(t)I(t) + \gamma I(t)dt\right), \tag{4.1.2}$$

where $S(u-) = \lim_{t\to u-} S(t), I(u-) = \lim_{t\to u-} I(t)$.

We explain how we can derive the likelihood function. Since this is a continuous-time stochastic process, then the time between two infection times and the time between two removal times follow a exponential distribution.

For $2 \le j \le n$, the rate to move from the j-th infection to the $(j+1)$-th infection is

$$\frac{\beta}{N} I(i_j)S(i_j).$$

Therefore, the interevent times between the j-th and $(j+1)$-th jump follow the exponential distribution with rate $\frac{\beta}{N} I(i_j)S(i_j)$. Thus, the probability density function of infections occurring at $i_2, \ldots, i_n$ can be written as following

$$\prod_{j=2}^{n} \left(\frac{\beta}{N} S(i_j-)I(i_j-)\right) \times \exp\left(-\frac{\beta}{N} S(i_j-)I(i_j-)(i_j - i_{j-1})\right). \tag{4.1.3}$$

By the same manner, the probability density function of removals occurring at $r_1, r_2, \ldots, r_n$ can be written as following

$$\prod_{j=1}^{n} \left(\gamma I(r_j-)\right) \times \exp\left(-\gamma I(r_j-)(r_j - r_{j-1})\right), \tag{4.1.4}$$

where $r_0 := 0$.

By combining the equation in (4.1.3) and (4.1.4), we obtain the likelihood function

for the complete-data

$$f(\mathbf{i},\mathbf{r}|\beta,\gamma) = \left(\prod_{j=2}^{n} \frac{\beta}{N} S(i_j-)I(i_j-)\right)\left(\prod_{j=1}^{n} \gamma I(r_j-)\right) \times \exp\left(-\frac{\beta}{N}\sum_{j=2}^{n} S(i_j-)I(i_j-)(i_j - i_{j-1}) - \gamma\sum_{j=1}^{n} I(r_j-)(r_j - r_{j-1})\right).$$

Rewrite likelihood function by replacing the summations with integrals, we can derive the likelihood function as in (4.1.2).

4.2 Methodology

Now, we assume that only the removal times $\mathbf{r} = (r_1, r_2, \ldots, r_n)$ have been observed. Since we do not have the infection times, we use the data augmentation and Gibbs sampling approach. In particular, we augment the set of unknown parameters β and γ with the infection times, which is $\mathbf{i}$.

Regarding the prior of β and γ, we choose that they follow the Gamma distribution. In particular,

$$\beta \sim \text{Gamma}(\mu_\beta, \lambda_\beta)$$
$$\gamma \sim \text{Gamma}(\mu_\gamma, \lambda_\gamma).$$

In other words, their prior distribution are proportional to

$$f(\beta) \propto \beta^{\mu_\beta - 1}\exp(-\lambda_\beta \beta)$$
$$f(\gamma) \propto \gamma^{\mu_\gamma - 1}\exp(-\lambda_\gamma \gamma)$$

Finally, we estimate β, γ and $\mathbf{i}$ based on their posterior distribution

$$f(\beta,\gamma,\mathbf{i}|\mathbf{r}) \propto f(\beta,\gamma,\mathbf{i},\mathbf{r}) = f(\mathbf{i},\mathbf{r}|\beta,\gamma)f(\beta)f(\gamma).$$

First, we have to initialize the parameters $\beta, \gamma, \mathbf{i}$. In our work, this is the initial

parameters:

- $\beta = 0.01$
- $\gamma = 0.01$
-

$$\begin{aligned}\mathbf{i} =&(0, 7, 17, 25.5, 30.75, 34.88, 36.94, 37.97, 38.98, 41.49,\\ &45.25, 48.62, 51.31, 52.66, 54.33, 55.16, 58.08, 61.04, 63.02, 66.01,\\ &67.51, 68.75, 69.88, 70.94, 72.47, 73.23, 74.12, 77.06, 78.53, 81.76)\end{aligned} \tag{4.2.1}$$

Note that in (4.2.1), (i_j, r_j) are the infection and removal times for the same individual j. Moreover, for each individual, the initial value of the j-th infection time is taken to be halfway between the $(j-1)$-th infection times and the $(j-1)$-th removal times.

At each iteration, parameters are sequentially updated with the Gibbs sampling procedure detailed as follows.

- Given the current value of $\mathbf{i}, \mathbf{r}, \gamma$, we update β using the posterior distribution. Note that

$$\begin{aligned} f(\beta|\mathbf{i}, \mathbf{r}, \gamma) &\propto f(\beta, \gamma, \mathbf{i}, \mathbf{r}) \propto f(\mathbf{i}, \mathbf{r}|\beta, \gamma) f(\beta) \\ &\propto \beta^{n-1} \exp\left(-\frac{\beta}{N} \int_0^T I(t)S(t)dt\right) \beta^{\mu_\beta - 1} \exp(-\lambda_\beta \beta) \\ &= \beta^{n+\mu_\beta - 2} \exp\left[-\left(\frac{1}{N} \int_0^T I(t)S(t)dt + \lambda_\beta\right)\beta\right] \end{aligned}$$

  Therefore, $\beta|\mathbf{i}, \mathbf{r}, \gamma \sim \text{Gamma}\left(n + \mu_\beta - 1, \frac{1}{N} \int_0^T I(t)S(t)dt + \lambda_\beta\right)$. This means that we update β by drawing from the Gamma posterior distribution.

- Given the current value of $\mathbf{i}, \mathbf{r}, \beta$, we update γ using the posterior distribution.

Note that

$$\begin{aligned} f(\gamma|\mathbf{i},\mathbf{r},\beta) &\propto f(\gamma,\beta,\mathbf{i},\mathbf{r}) \propto f(\mathbf{i},\mathbf{r}|\beta,\gamma)f(\gamma) \\ &\propto \gamma^n \exp\left(-\gamma \int_0^T I(t)dt\right) \gamma^{\mu_\gamma - 1} \exp(-\lambda_\gamma \gamma) \\ &= \gamma^{n+\mu_\gamma - 1} \exp\left[-\left(\int_0^T I(t)dt + \lambda_\gamma\right)\gamma\right] \end{aligned}$$

Therefore, $\gamma|\mathbf{i},\mathbf{r},\beta \sim \text{Gamma}\left(n+\mu_\gamma, \int_0^T I(t)dt + \lambda_\gamma\right)$. This means that we update γ by drawing from the Gamma posterior distribution.

- Given the value of β, γ, and the removal times $\mathbf{r}$, we update $\mathbf{i}$. We have

$$f(\mathbf{i}|\beta,\gamma,\mathbf{r}) \propto f(\mathbf{i},\beta,\gamma,\mathbf{r}) \propto f(\mathbf{i},\mathbf{r}|\beta,\gamma)$$

Note that $\mathbf{i}$ have n components, but we only update $i_2, i_3, \ldots, i_n$ since i_1 is always set to be 0. Without loss of generality, we show how to update i_2. Assume that the current value is i_2. Then the new value of i_2, $\tilde{i}_2$ is drawn from a uniform distribution on $(0, r_2)$. Let $\mathbf{i}$ be the current infection times vector, and $\tilde{\mathbf{i}}$ be the new vector, which is the same as $\mathbf{i}$, except for the second infection times i_2. The proposal is then accepted, with probability

$$\min\left\{1, \frac{f(\tilde{\mathbf{i}},\mathbf{r}|\beta,\gamma)}{f(\mathbf{i},\mathbf{r}|\gamma,\beta}\right\}.$$

After update i_2, we continue to update i_j, $3 \leq j \leq n$.

4.3 Result

Consider the smallpox data, we apply the MCMC algorithm proposed in section "Methodology". We run 600 iterations, and then discard 200 iterations as burn-in. The posterior distribution for β, γ and the basic reproduction number $\mathcal{R}_0$ is

depicted in figure (4.1), (4.2). The mean value of β is 0.093638. The mean value of γ is 0.083267. The mean value of $\mathcal{R}_0$ is 1.161784.

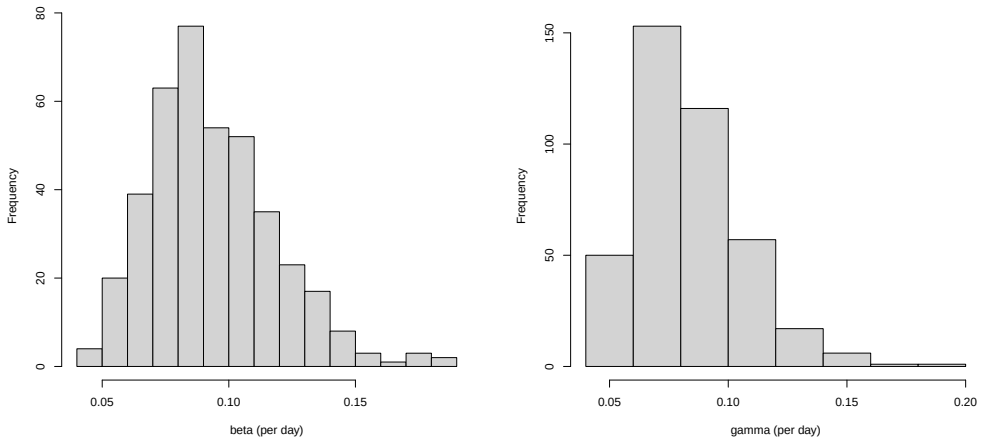

Figure 4.1: [Left] The distribution of β. [Right] The distribution of γ.

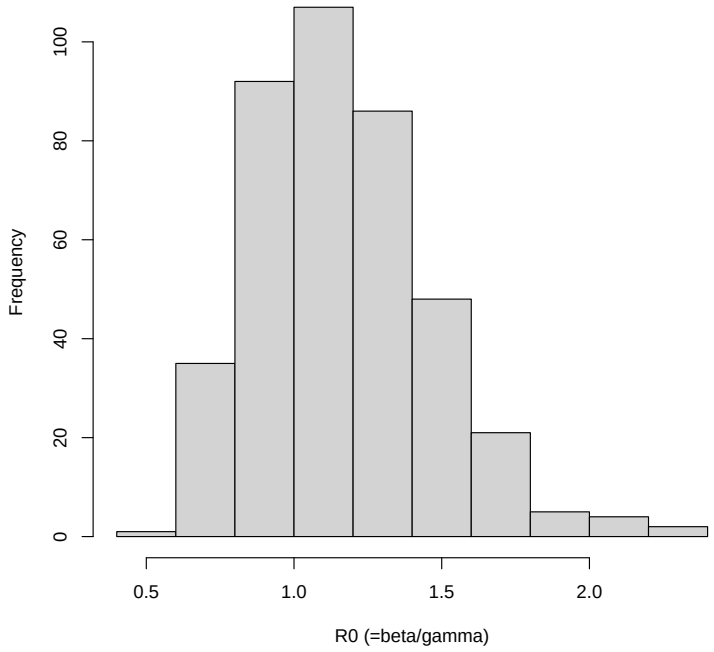

Figure 4.2: The distribution of the basic reproduction number $\mathcal{R}_0$.

Chapter 5

Bayesian posterior prediction with daily updated data

In this chapter, we introduce the main work of our Master thesis. In practice, we usually observe new cases daily. Therefore, we cannot use the model introduced in chapter 4. We adjust the model so that the new cases follow the Poisson distribution or the Binomial distribution. We then estimate the transmission rate and removal rate. Moreover, we also forecast new infected cases and the final size of the outbreak. This is an simplified model compared to the model from Gu and Yin [4] and Ward et al. [9].

5.1 Methodology

5.1.1 Data and notations

For the dataset, we consider the population of N individuals. At the beginning of the process, $t = 0$, the initial states are S_0, I_0, R_0. Note that $S_0 + I_0 + R_0 = N$. At time k, $k \in \mathbb{N}, k \geq 1$, the number of susceptible, infected and removed individuals are S_k, I_k, R_k, respectively. We observe the new infected cases and new removed cases daily. Denote the new infected cases in day k as IN_k and the new removed cases as RN_k. Since we observe n days of the epidemic, we let

$\mathbf{IN} = (IN_1, IN_2, \ldots, IN_n)$ and $\mathbf{RN} = (RN_1, RN_2, \ldots, RN_n)$ to be sequences of daily reported numbers of new infected cases and removed cases, respectively. It is obvious that

$$IN_k = S_{k-1} - S_k$$

$$RN_k = R_k - R_{k-1}$$

We consider 4 data sets. There are 2 simulated datasets, and the Ebola and Covid-19 data set. For the likelihood of the data, we consider two cases. The first one considers the binomial distribution, and the other one considers the Poisson distribution.

5.1.2 The binomial likelihood

For $k = 1, 2, \ldots, n$ we make the following assumption for each susceptible individual at day $k-1$: The number of in-person contact with infected individuals and then became infected follows a Poisson distribution with rate $\lambda = -\beta\frac{I_{k-1}}{N}$. Then the probability that there is new infected cases is $1 - \exp\left(-\beta\frac{I_{k-1}}{N}\right)$. Therefore, we can assume that the new cases follow a binomial distribution as

$$IN_k \sim \text{Binomial}\left(S_{k-1}, 1 - \exp\left(-\beta\frac{I_{k-1}}{N}\right)\right). \tag{5.1.1}$$

By the same manner, the new removed cases follow the Binomial distribution with rate γ

$$RN_k \sim \text{Binomial}\left(I_{k-1}, \gamma\right). \tag{5.1.2}$$

Regarding the prior of β and γ, we choose so that they follow the Gamma distribution. In particular,

$$\beta \sim \text{Gamma}(\mu_\beta, \lambda_\beta)$$
$$\gamma \sim \text{Gamma}(\mu_\gamma, \lambda_\gamma).$$

We then use MCMC to derive the posterior distribution. However, we can use theorem (2.1.7) to approximate the binomial distribution with the Poisson distribution so that we can update the posterior distribution easier via conjugate prior.

5.1.3 The Poisson likelihood

For $k \in \mathbb{N}, k \geq 1$, by theorem (2.1.7), we can approximate the Binomial distribution with size S_{k-1} and success rate $1 - \exp\left(-\beta\frac{I_{k-1}}{N}\right)$ with the Poisson distribution with rate $S_{k-1}\left(1 - \exp\left(-\beta\frac{I_{k-1}}{N}\right)\right)$. Moreover, by Taylor expansion, we have that

$$1 - \exp\left(-\beta\frac{I_{k-1}}{N}\right) \sim \beta\frac{I_{k-1}}{N}.$$

Thus, (5.1.1) becomes

$$IN_k \sim \text{Poisson}\left(\beta\frac{S_{k-1}I_{k-1}}{N}\right) \tag{5.1.3}$$

By the same manner, (5.1.2) becomes

$$RN_k \sim \text{Poisson}\left(\gamma I_{k-1}\right) \tag{5.1.4}$$

By choosing the prior distribution of β and γ to be Gamma distribution,

$$\beta \sim \text{Gamma}(\mu_\beta, \lambda_\beta)$$
$$\gamma \sim \text{Gamma}(\mu_\gamma, \lambda_\gamma),$$

we can derive the posterior distribution of β and γ

$$\beta|IN_1, IN_2, \ldots, IN_n, S_0, I_0 \sim \text{Gamma}\left(\mu_\beta + \sum_{i=1}^{n} IN_i, \lambda_\beta + \sum_{i=1}^{n} \frac{S_{i-1}I_{i-1}}{N}\right), \tag{5.1.5}$$

$$\gamma|RN_1, RN_2, \ldots, RN_n, I_0 \sim \text{Gamma}\left(\mu_\gamma + \sum_{i=1}^{n} RN_i, \lambda_\gamma + \sum_{i=1}^{n} I_{i-1}\right). \tag{5.1.6}$$

5.1.4 Bayesian prediction using the Poisson likelihood

From the community of N individuals, we observe n data points, and now we want to predict what happens in the next m days. To to this, we first define how can we simulate the epidemic with known β, γ, N and initial value S_0, I_0. Assume that we want to simulate m days of the epidemic, we define the simulation function with inputs $N, \beta, \gamma, S_0, I_0, m$ as follow

```
simulation <- function(N, beta, gamma, S0, I0, m){
  new_cases = c(NA)
  S = S0
  I = I0
  for (i in 1: m){
    IN_new = rpois(1, lambda = S*I*beta/N)
    RN_new = rpois(1, lambda = I*gamma)
    S = S - IN_new
    I = I + IN_new - RN_new
    new_cases = c(result, new_cases)
```

```
  }
  return(new_cases[-1])
}
```

Now, we describe our prediction. This is the iterative methods, and for each iteration, we

- Draw β from the posterior distribution (5.1.5). Draw γ from the posterior distribution (5.1.6).
- Now we run the simulation of m days of the epidemic with β and γ from the previous step. For the initial condition, we let S_n, I_n (the number of susceptible individuals and infected individuals at the n-th day, respectively) to be the initial condition of the epidemic.
- From the simulation, we obtain the number of new cases and the number of removed individuals from each day.
- Repeat the procedure again.

After that, we plot the median and the 90% credible interval for the real and prediction new infected cases from day 1 to day $n + m$. Moreover, we also plot our prediction for the final size of the epidemic.

5.2 First simulated data set

5.2.1 Introduce the data set

The first data set is the simulated data set. The population is 200 and there are 5 infected individuals in the beginning. The epidemic ends after 110 days and the outbreak size is 200 individuals. Figure (5.1) illustrates the new cases in the first 30 days and the process in the first 50 days.

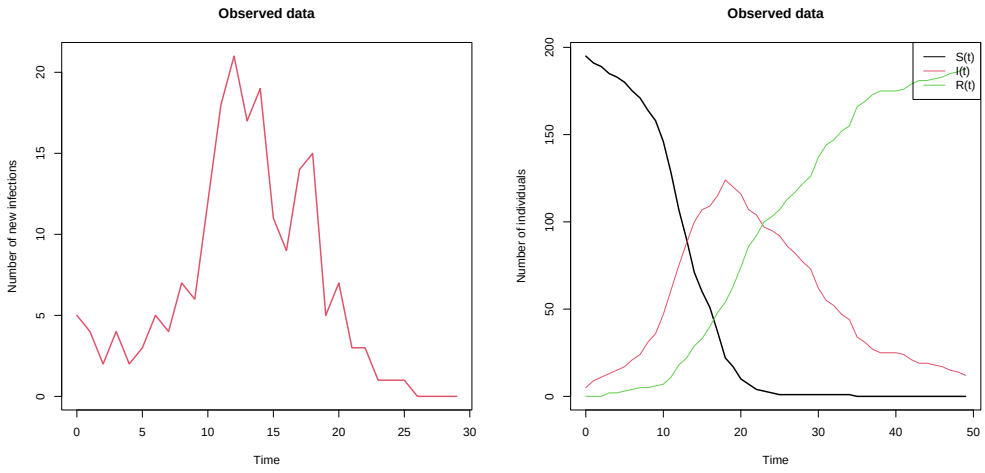

Figure 5.1: First data set: [Left] New cases reported in 30 days. [Right] The process in 50 days.

5.2.2 Bayesian analysis and posterior prediction

First, we determine our prior distribution. We choose $\beta \sim \text{Gamma}(0.01, 0.01)$ and $\gamma \sim \text{Gamma}(0.01, 0.01)$. We then update our belief in the parameters using the data. Our goal is to estimate the basic reproduction number $\mathcal{R}_0$ and predict the new cases until day 30 and the final size of the epidemic.

First case: We have data of 5 days

For now, we only have the data of 5 days. The posterior distribution of β and γ is

$$\beta \sim \text{Gamma}(15.01, 49.625),$$
$$\gamma \sim \text{Gamma}(3.01, 53.01).$$

The Bayesian estimation of β is 0.3025 and the 90% credible interval of β is $(0.18648, 0.44128)$. The Bayesian estimation of γ is 0.0568 and the 90% credible interval of γ is $(0.01552, 0.11905)$. By using the definition, $\mathcal{R}_0 = \dfrac{\beta}{\gamma}$, we can derive the distribution of the basic reproduction number $\mathcal{R}_0$, which is shown in figure (5.2). The red line in the figure indicates $\mathcal{R}_0 = 1$. The Bayesian estimation of $\mathcal{R}_0$ is 8.0401, and the 90% credible interval of $\mathcal{R}_0$ is $(2.18, 20.7)$.

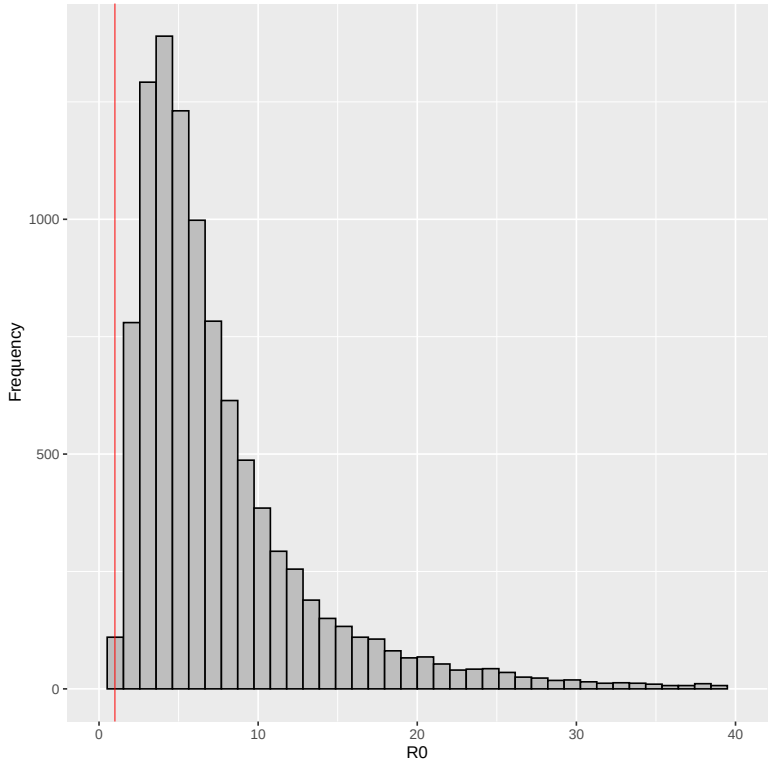

Figure 5.2: Data set 1, first case: The distribution of the basic reproduction number $\mathcal{R}_0$.

The Bayesian prediction for the new cases daily and the final size of the epidemic is shown in figure (5.3). In the early stage after just 5 days of data, the predictions had substantial uncertainty, reflected by the wide 90% credible intervals for the basic reproduction number $\mathcal{R}_0$ and the projected epidemic curves. The estimate of $\mathcal{R}_0$ had a mean of 8.04 but a broad credible range spanning 2.18 to 20.7.

Second case: We have data of 15 **days**

For now, we have the data of 15 days. The posterior distribution of β and γ is

$$\beta \sim \text{Gamma}(135.01, 346.255),$$

$$\gamma \sim \text{Gamma}(33.01, 553.01).$$

The Bayesian estimation of β is 0.389 and the 90% credible interval of β is $(0.336, 0.447)$. The Bayesian estimation of γ is 0.0597 and the 90% credible in-

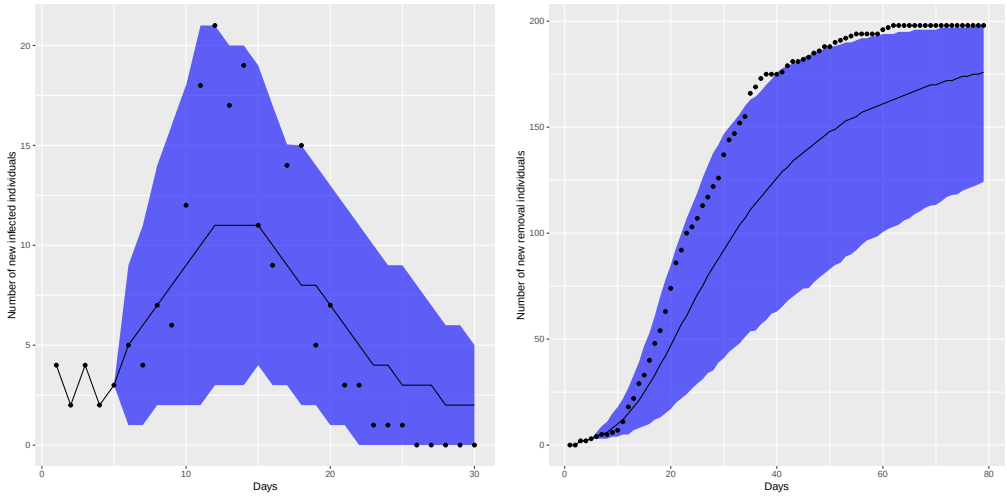

Figure 5.3: Data set 1, first case: The mean and 90% credible intervals for Bayesian prediction of future infected cases (left) and the final size (right) compared to the true value of the epidemic.

terval of γ is $(0.0437, 0.078)$. By using the definition, $\mathcal{R}_0 = \dfrac{\beta}{\gamma}$, we can derive the distribution of the basic reproduction number $\mathcal{R}_0$, which is shown in figure (5.4). The red line in the figure indicates $\mathcal{R}_0 = 1$. The Bayesian estimation of $\mathcal{R}_0$ is 6.761, and the 90% credible interval of $\mathcal{R}_0$ is $(4.85, 9.28)$. The credible intervals tightened substantially. The $\mathcal{R}_0$ estimate has a mean of 6.76 with a range of 4.85 to 9.28.

The Bayesian prediction for the new cases daily and the final size of the epidemic is shown in figure (5.5). The epidemic trajectories also began to align more accurately with the true value of the removed individuals outbreak.

Third case: We have data of 20 days

For now, we have the data of 20 days. The posterior distribution of β and γ is

$$\beta \sim \text{Gamma}(185.01, 451.265),$$
$$\gamma \sim \text{Gamma}(74.01, 1128.01).$$

The Bayesian estimation of β is 0.41 and the 90% credible interval of β is $(0.362, 0.46)$. The Bayesian estimation of γ is 0.066 and the 90% credible interval of γ is

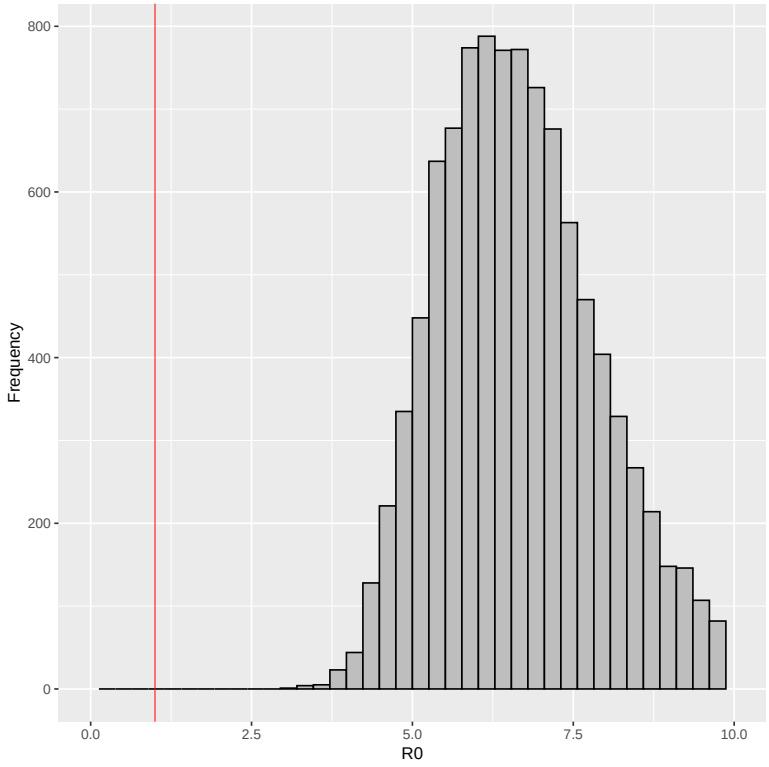

Figure 5.4: Data set 1, second case: The distribution of the basic reproduction number $\mathcal{R}_0$.

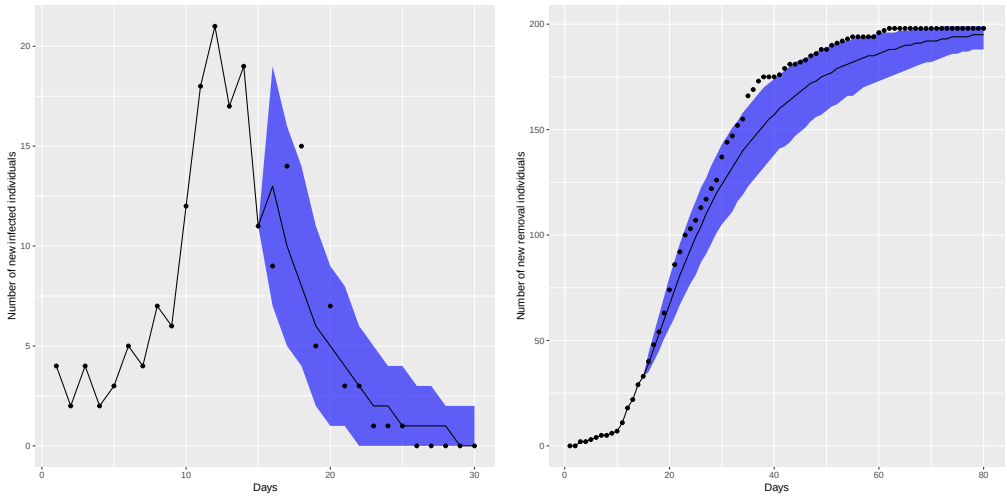

Figure 5.5: Data set 1, second case: The mean and 90% credible intervals for Bayesian prediction of future infected cases (left) and the final size (right) compared to to the true value of the epidemic.

$(0.054, 0.079)$. By using the definition, $\mathcal{R}_0 = \dfrac{\beta}{\gamma}$, we can derive the distribution of the basic reproduction number $\mathcal{R}_0$, which is shown in figure (5.6). The red line in the figure indicates $\mathcal{R}_0 = 1$. The Bayesian estimation of $\mathcal{R}_0$ is 6.272, and the 90% credible interval of $\mathcal{R}_0$ is $(4.99, 7.93)$.

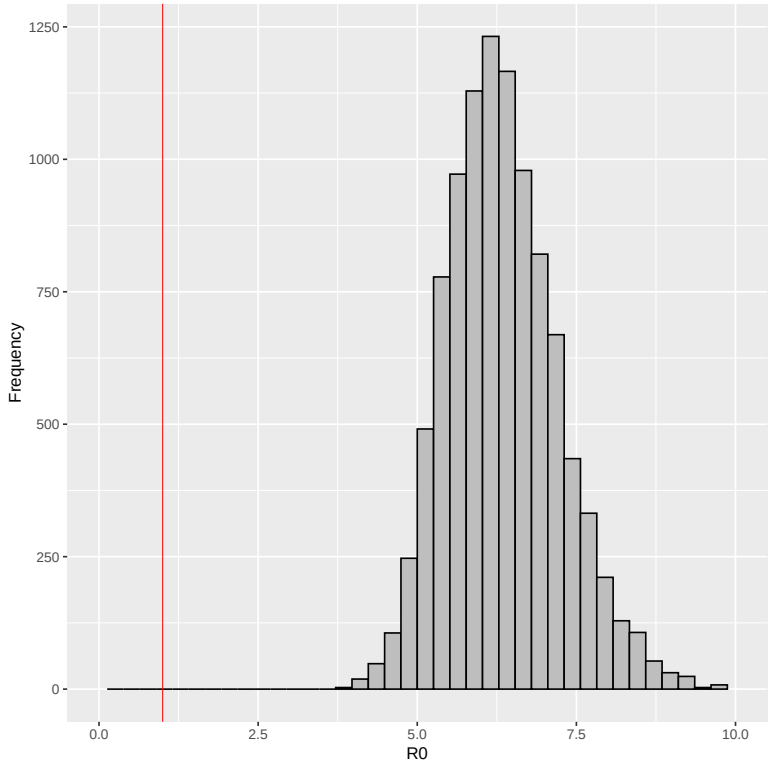

Figure 5.6: Data set 1, third case: The distribution of the basic reproduction number $\mathcal{R}_0$.

The Bayesian prediction for the new cases daily and the final size of the epidemic is shown in figure (5.7). The predicted epidemic curves matched the true outbreak well, precisely capturing the turning points of the curves.

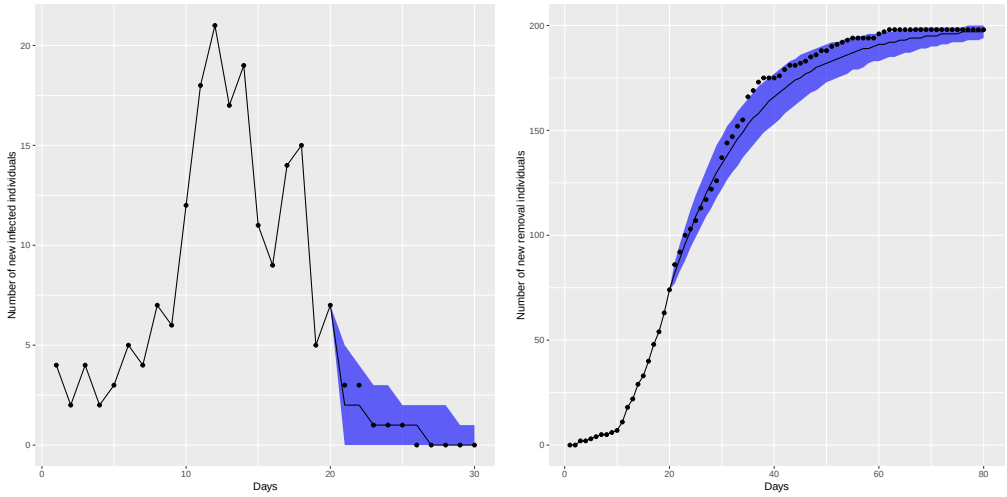

Figure 5.7: Data set 1, third case: The mean and 90% credible intervals for Bayesian prediction of future infected cases (left) and the final size (right) compared to to the true value of the epidemic.

5.3 Ebola outbreak in Kikwit, Democratic Republic of the Congo, 1995

5.3.1 Introduce the data set

In this section, we consider the Ebola disease in Kikwit, Democratic Republic of the Congo in 1995. The data on daily cases reported by Khan et al. (1999) cover the period 1995-01-06 to 1995-07-16, over which time there were 291 cases and 236 deaths. In their data, the first case became ill on 1995-01-06 and was death on 1995-03-02. However, since they only start collect data daily from 1995-03-01, we assume that the epidemic starts from 1995-03-06. To summarize, in our work, the epidemic starts from 1995-03-06 and there is one infected individual in the beginning. When an individual was infected or death, they would be moved to the infected state or removed state, respectively. We consider the epidemic in 132 days. Information about the epidemic is depicted in figure (5.8).

Remark 5.3.1. *Note that in the Ebola outbreak in Kikwit, the recovered individuals were not recorded. Therefore, the infected individuals curve does not decline to* 0. *However, for the simplicity, we assume that the epidemic continues after*

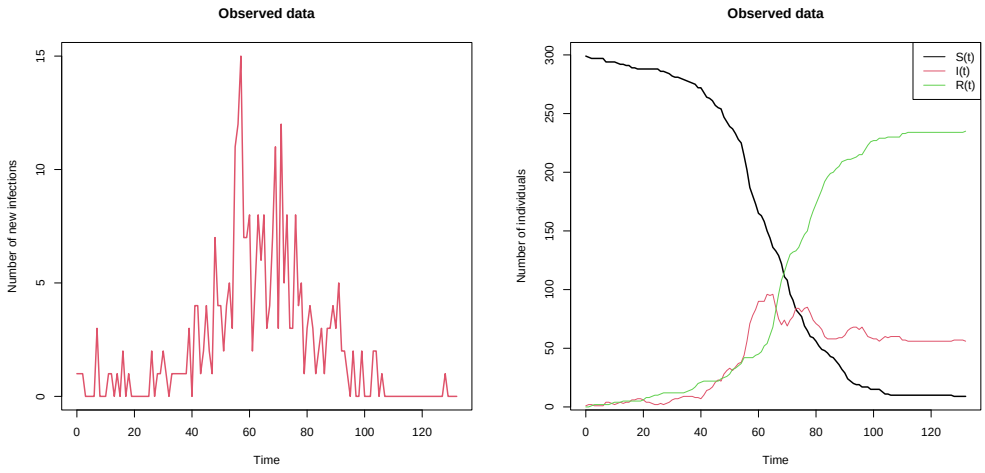

Figure 5.8: Ebola data set: [Left] New cases reported in 132 days. [Right] The process in 132 days.

1995-07-16, and we only consider the epidemic in the period of 132 *days.*

5.3.2 Bayesian posterior prediction

First, we determine our prior distribution. We choose $\beta \sim \text{Gamma}(0.01, 0.01)$ and $\gamma \sim \text{Gamma}(0.01, 0.01)$. We then update our belief in the parameters using the data. Our goal is to estimate the basic reproduction number $\mathcal{R}_0$ and predict the new cases until day 132 and the final size of the epidemic.

First case: We have data of 20 **days**

First, we assume that we have the data of 20 days. The posterior distribution of β and γ is

$$\beta \sim \text{Gamma}(11.01, 64.177),$$
$$\gamma \sim \text{Gamma}(6.01, 66.01).$$

The Bayesian estimation of β is 0.1716 and the 90% credible interval of β is $(0.0962, 0.2645)$. The Bayesian estimation of γ is 0.091 and the 90% credible interval of γ is $(0.0397, 0.1595)$. By using the definition, $\mathcal{R}_0 = \dfrac{\beta}{\gamma}$, we can derive the distribution of the basic reproduction number $\mathcal{R}_0$, which is shown in figure

(5.9). The red line in the figure indicates $\mathcal{R}_0 = 1$. The Bayesian estimation of $\mathcal{R}_0$ is 2.267, and the 90% credible interval of $\mathcal{R}_0$ is $(0.857, 4.73)$. In the early stage after just 20 days of data, the predictions were quite uncertain. The mean of $\mathcal{R}_0$ is 2.27 but the 90% credible interval ranging from 0.86 to 4.73, which is very wide.

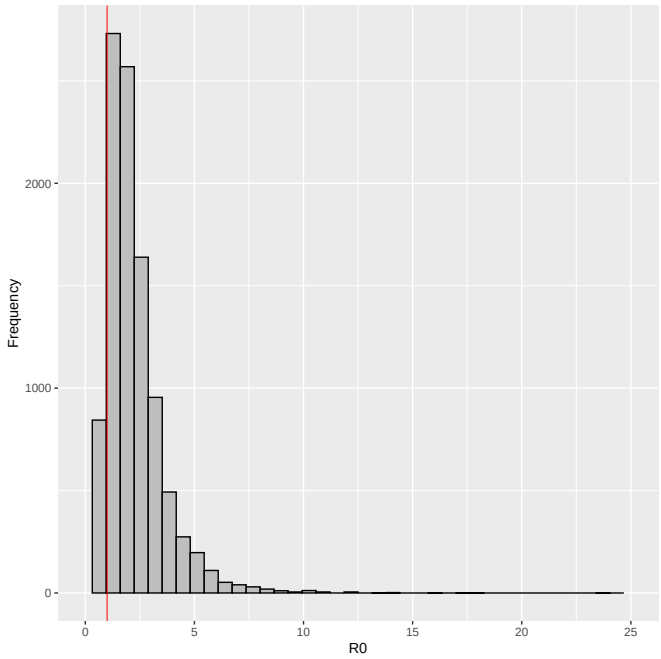

Figure 5.9: Data set 2, first case: The distribution of the basic reproduction number $\mathcal{R}_0$.

The Bayesian prediction for the new cases daily and the final size of the epidemic is shown in figure (5.10). The prediction showed reasonable potential scenarios, but cannot accurately capture the true value of the outbreak.

Second case: We have data of 50 days

Now, we assume that we have the data of 50 days. The posterior distribution of β and γ is

$$\beta \sim \text{Gamma}(60.01, 329.72),$$

$$\gamma \sim \text{Gamma}(28.01, 367.01).$$

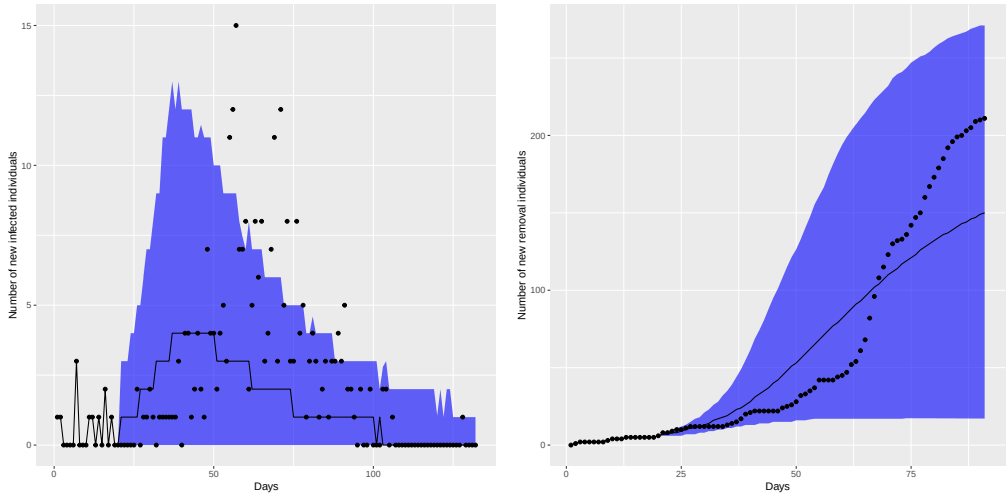

Figure 5.10: Ebola data set, first case: The mean and 90% credible intervals for Bayesian prediction of future infected cases (left) and the final size (right) compared to to the true value of the epidemic.

The Bayesian estimation of β is 0.182 and the 90% credible interval of β is $(0.1452, 0.2222)$. The Bayesian estimation of γ is 0.0763 and the 90% credible interval of γ is $(0.0543, 0.1015)$. By using the definition, $\mathcal{R}_0 = \dfrac{\beta}{\gamma}$, we can derive the distribution of the basic reproduction number $\mathcal{R}_0$, which is shown in figure (5.11). The red line in the figure indicates $\mathcal{R}_0 = 1$. The Bayesian estimation of $\mathcal{R}_0$ is 2.4756, and the 90% credible interval of $\mathcal{R}_0$ is $(1.66, 3.56)$. As more data accumulated up to 50 days, the parameter estimates became significantly more precise.

The Bayesian prediction for the new cases daily and the final size of the epidemic is shown in figure (5.12). The prediction aligns much better with the true value of the outbreak.

Third case: We have data of 80 **days**

Now, we assume that we have the data of 80 days. The posterior distribution of β and γ is

$$\beta \sim \text{Gamma}(243.01, 1286.833),$$

$$\gamma \sim \text{Gamma}(173.01, 2525.01).$$

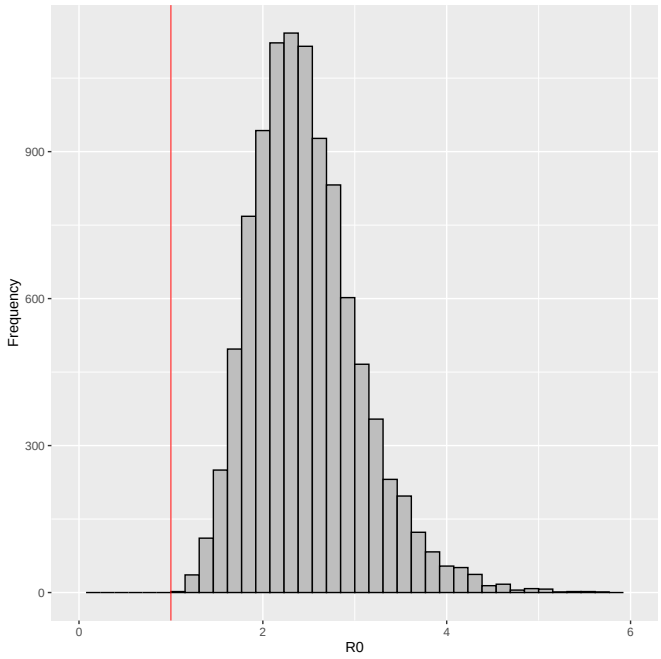

Figure 5.11: Data set 2, second case: The distribution of the basic reproduction number $\mathcal{R}_0$.

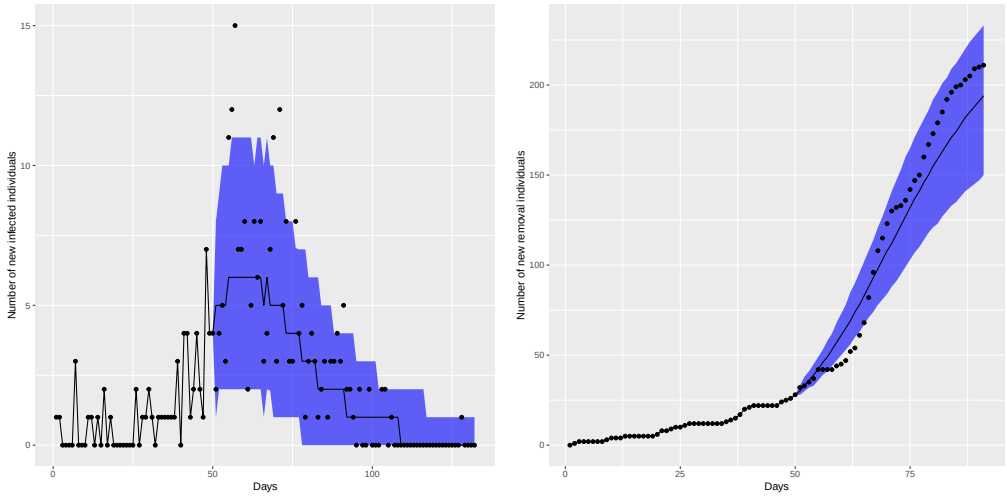

Figure 5.12: Ebola data set, second case: The mean and 90% credible intervals for Bayesian prediction of future infected cases and the final size compared to to the true value of the epidemic in 132 days.

The Bayesian estimation of β is 0.1888 and the 90% credible interval of β is $(0.1694, 0.2092)$. The Bayesian estimation of γ is 0.0685 and the 90% credible interval of γ is $(0.0602, 0.0773)$. The By using the definition, $\mathcal{R}_0 = \dfrac{\beta}{\gamma}$, we can derive the distribution of the basic reproduction number $\mathcal{R}_0$, which is shown in figure (5.13). The red line in the figure indicates $\mathcal{R}_0 = 1$. The Bayesian estimation of $\mathcal{R}_0$ is 2.774, and the 90% credible interval of $\mathcal{R}_0$ is $(2.34, 3.25)$. As we gather more information, the mean of $\mathcal{R}_0$ was 2.77 with a tight 90% credible interval of 2.34 to 3.25.

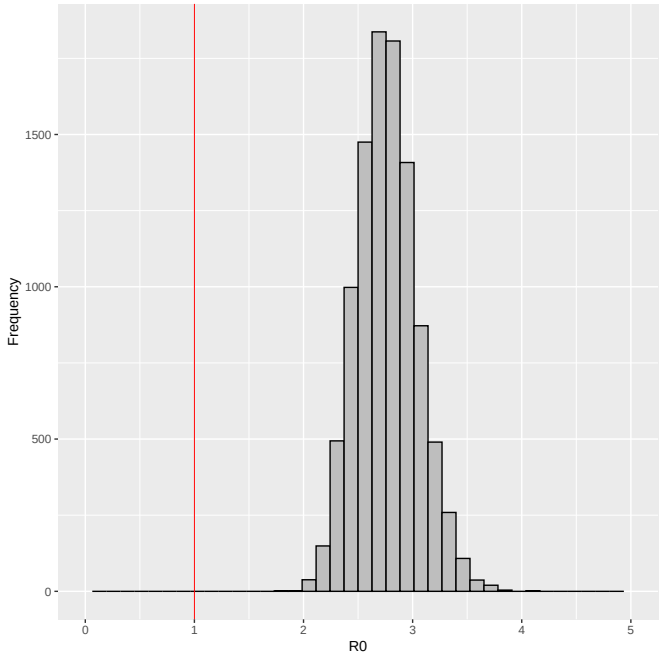

Figure 5.13: Data set 2, third case: The distribution of the basic reproduction number $\mathcal{R}_0$.

The Bayesian prediction for the new cases daily and the final size of the epidemic is shown in figure (5.14). The prediction can accurately predict the entire outbreak well, closely matching the observed data.

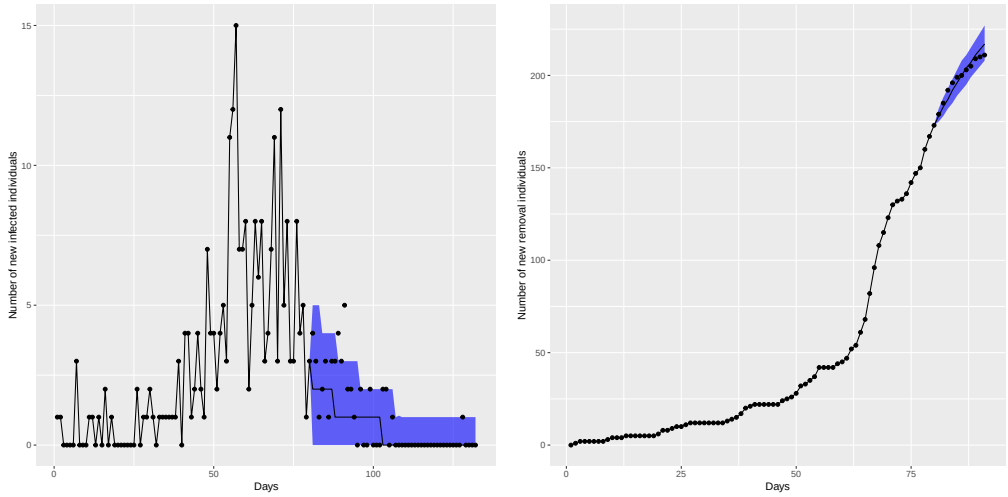

Figure 5.14: Ebola data set, third case: The mean and 90% credible intervals for Bayesian prediction of future infected cases (left) and the final size (right) compared to to the true value of the epidemic.

5.4 Second simulated data set

5.4.1 Introduce the data set

The third data set is the simulated data set. The population is 41100 and there are 10 infected individuals in the beginning. The epidemic ends after 154 days and the outbreak size is 24301 individuals. Figure (5.15) illustrates the new cases in the first 30 days and the process in 154 days. The difference with the first simulated data set is that not everyone got the disease. In particular, the number of population is 41100 but the final size of the epidemic is 24301, which is much smaller than 41100.

5.4.2 Baysian analysis and posterior prediction

First, we determine our prior distribution. We choose $\beta \sim \text{Gamma}(0.01, 0.01)$ and $\gamma \sim \text{Gamma}(0.01, 0.01)$. We then update our belief in the parameters using the data. Our goal is to estimate the basic reproduction number $\mathcal{R}_0$ and predict the new cases and the final size of the epidemic.

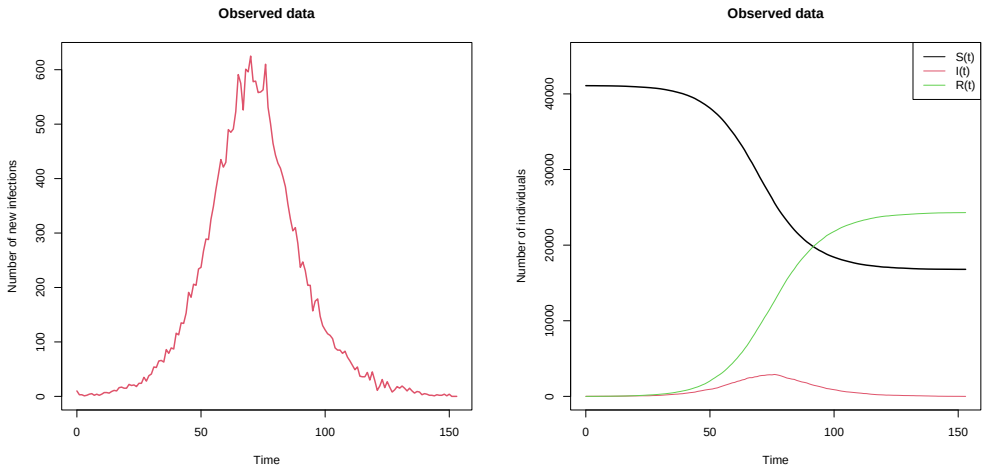

Figure 5.15: Third data set: [Left] New cases reported in 154 days. [Right] The process in 154 days.

First case: We have data of 5 days

First, we assume that we have the data of 5 days. The posterior distribution of β and γ is

$$\beta \sim \text{Gamma}(9.01, 45.99),$$

$$\gamma \sim \text{Gamma}(6.01, 46.01).$$

The Bayesian estimation of β is 0.1959 and the 90% credible interval of β is $(0.1023, 0.3141)$. The Bayesian estimation of γ is 0.1306 and the 90% credible interval of γ is $(0.0569, 0.2288)$. By using the definition, $\mathcal{R}_0 = \dfrac{\beta}{\gamma}$, we can derive the distribution of the basic reproduction number $\mathcal{R}_0$, which is shown in figure (5.16). The red line in the figure indicates $\mathcal{R}_0 = 1$. The Bayesian estimation of $\mathcal{R}_0$ is 1.78, and the 90% credible interval of $\mathcal{R}_0$ is $(0.64, 3.75)$. In the early stage after just 5 days of data, the predictions were again highly uncertain, with a broad 90% credible interval ranging from 0.64 to 3.75. Note that the credible interval includes the value of $\mathcal{R}_0 < 1$, so the model suggests that the epidemic might not happen.

The Bayesian prediction for the new cases daily and the final size of the epidemic is shown in figure (5.17). Beside the fact that the prediction is extremely

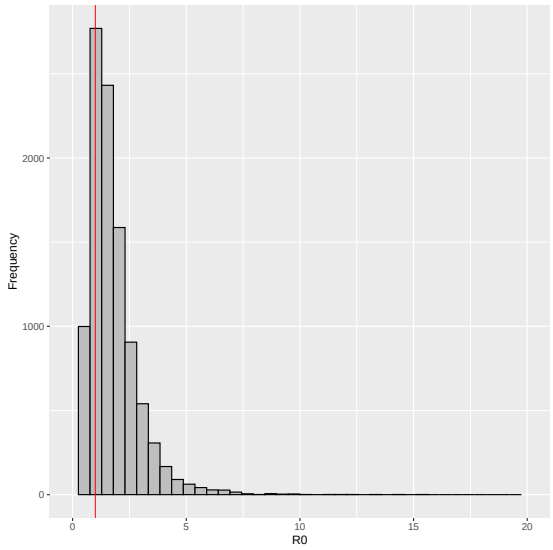

Figure 5.16: Data set 3, first case: The distribution of the basic reproduction number $\mathcal{R}_0$.

uncertain, we also notice that there is a chance that the epidemic might not happen.

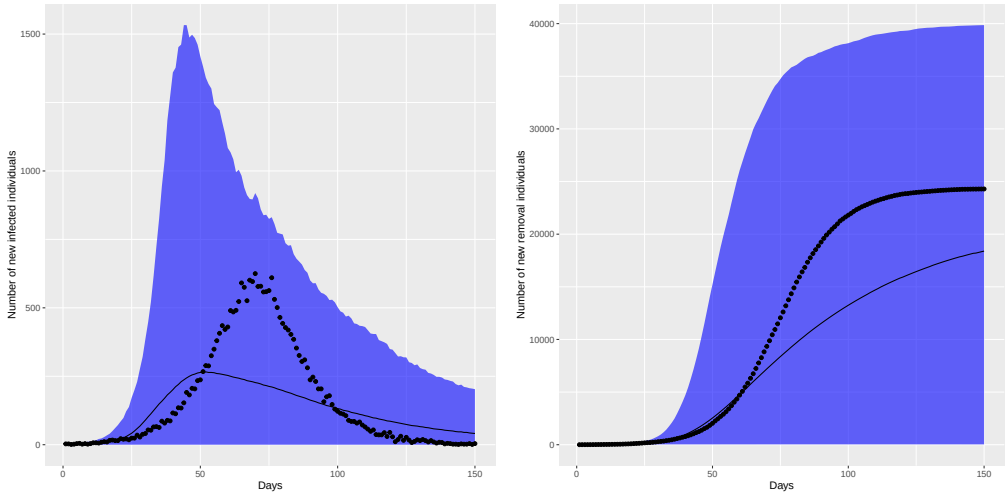

Figure 5.17: Data set 3, first case: The mean and 90% credible intervals for Bayesian prediction of future infected cases (left) and the final size (right) compared to the true value of the epidemic.

Second case: We have data of 30 days

Now, we assume that we have the data of 30 days. The posterior distribution of β and γ is

$$\beta \sim \text{Gamma}(373.01, 1232.5)$$

$$\gamma \sim \text{Gamma}(255.01, 1238.01).$$

The Bayesian estimation of β is 0.3026 and the 90% credible interval of β is $(0.2773, 0.3289)$. The Bayesian estimation of γ is 0.206 and the 90% credible interval of γ is $(0.1852, 0.2277)$. By using the definition, $\mathcal{R}_0 = \frac{\beta}{\gamma}$, we can derive the distribution of the basic reproduction number $\mathcal{R}_0$, which is shown in figure (5.18). The red line in the figure indicates $\mathcal{R}_0 = 1$. The Bayesian estimation of $\mathcal{R}_0$ is 1.475, and the 90% credible interval of $\mathcal{R}_0$ is $(1.29, 1.68)$.

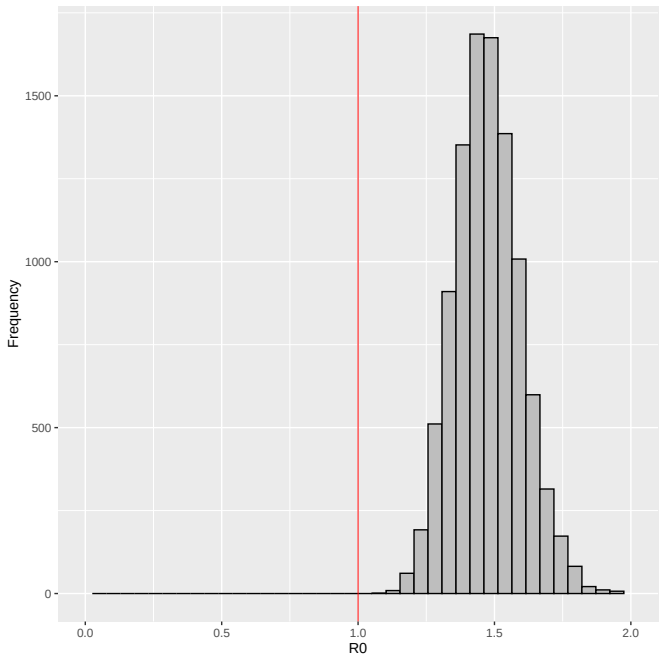

Figure 5.18: Data set 3, second case: The distribution of the basic reproduction number $\mathcal{R}_0$.

The Bayesian prediction for the new cases daily and the final size of the epidemic is shown in figure (5.19).

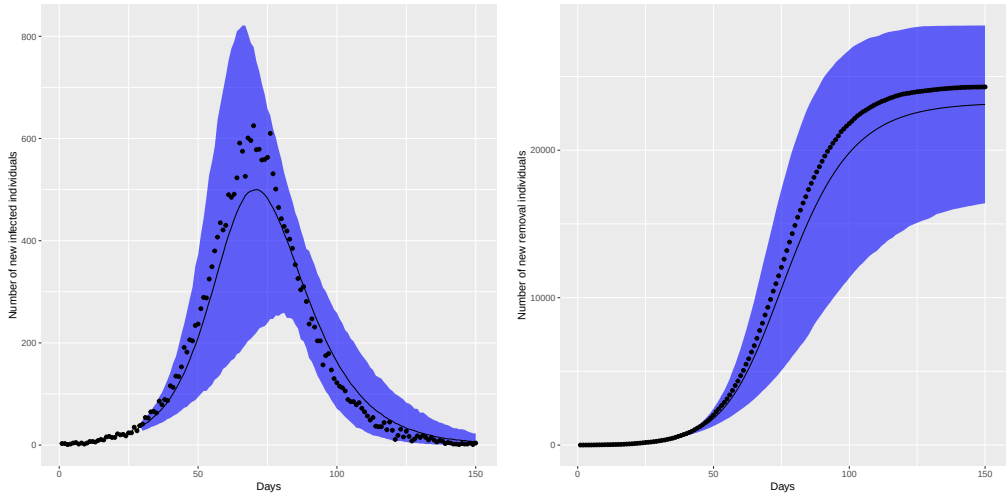

Figure 5.19: Data set 3, first case: The mean and 90% credible intervals for Bayesian prediction of future infected cases (left) and the final size (right) compared to the true value of the epidemic.

Third cases: We have data of 60 **days**

Now, we assume that we have the data of 60 days. The posterior distribution of β and γ is

$$\beta \sim \text{Gamma}(6122.01, 19953.66),$$

$$\gamma \sim \text{Gamma}(4344.01, 21534.01).$$

The Bayesian estimation of β is 0.3068 and the 90% credible interval of β is $(0.3, 0.3133)$. The Bayesian estimation of γ is 0.2017 and the 90% credible interval of γ is $(0.1967, 0.2068)$. By using the definition, $\mathcal{R}_0 = \dfrac{\beta}{\gamma}$, we can derive the distribution of the basic reproduction number $\mathcal{R}_0$, which is shown in figure (5.20). The red line in the figure indicates $\mathcal{R}_0 = 1$. The Bayesian estimation of $\mathcal{R}_0$ is 1.521, and the 90% credible interval of $\mathcal{R}_0$ is $(1.47, 1.57)$. With 60 days of data, the estimates became even more precise. The credible interval for $\mathcal{R}_0$ is very tight, suggesting that we can predict the final size of the epidemic well.

The Bayesian prediction for the new cases daily and the final size of the epidemic is shown in figure (5.21). We can see that the model can accurately predict the final epidemic size of the outbreak. A notable advantage highlighted by this

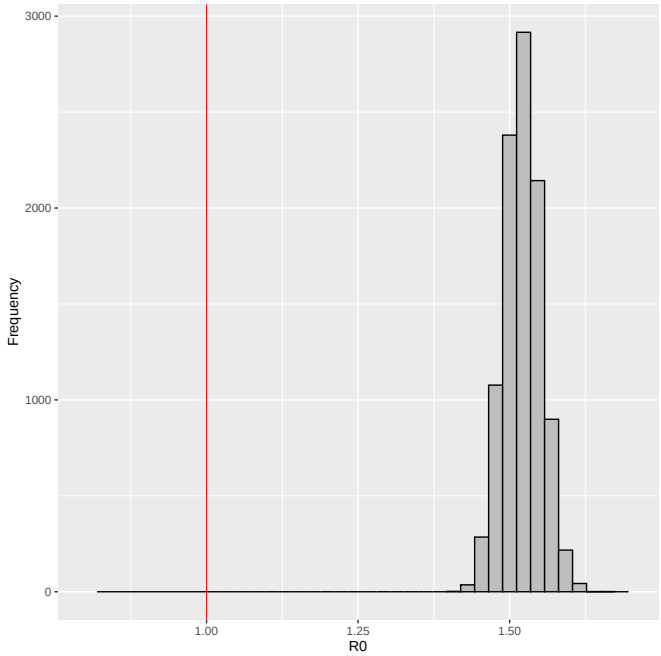

Figure 5.20: Data set 3, second case: The distribution of the basic reproduction number $\mathcal{R}_0$.

analysis was the flexibility of the framework to handle outbreaks that only impact a subset of the total population. Even though not everyone became infected, the model can still provide good prediction given sufficient data.

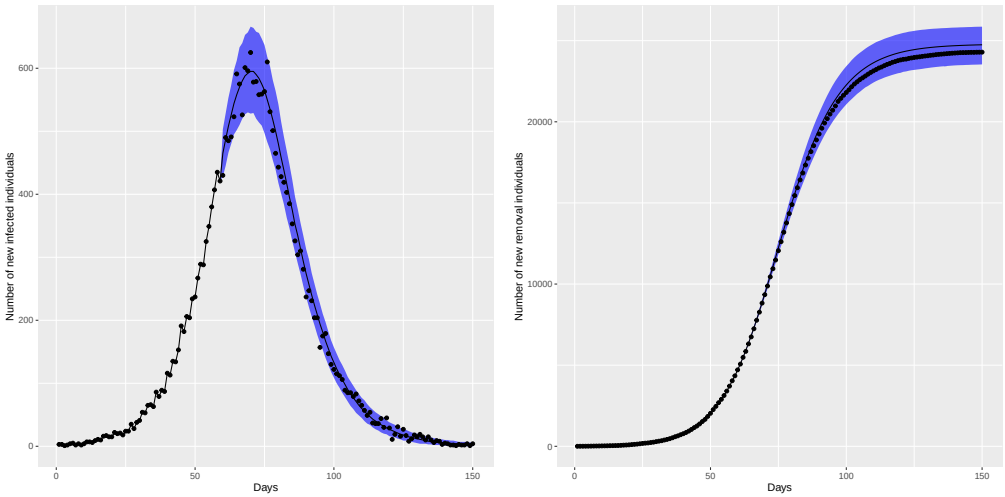

Figure 5.21: Data set 3, third case: The mean and 90% credible intervals for Bayesian prediction of future infected cases (left) and the final size (right) compared to the true value of the epidemic.

5.5 Covid-19 data set in New Orleans, Louisiana

5.5.1 Introduce the data

In this section, we consider the Covid 19 in New Orleans, Louisiana, US in 2020. The data on daily cases provided by Wahltinez et al. [8], covering the period 2020-03-09 to 2020-06-16 (before the second wave of Covid in New Orleans), over which time there were 8506 cases and 7418 deaths. In New Orleans, the first case was reported on 2020-03-09. The number of population is 436126. This is the first wave of Covid in New Orleans, and the recovery time was 20 days. Figure (5.22) illustrates the new cases and the whole process in 100 days.

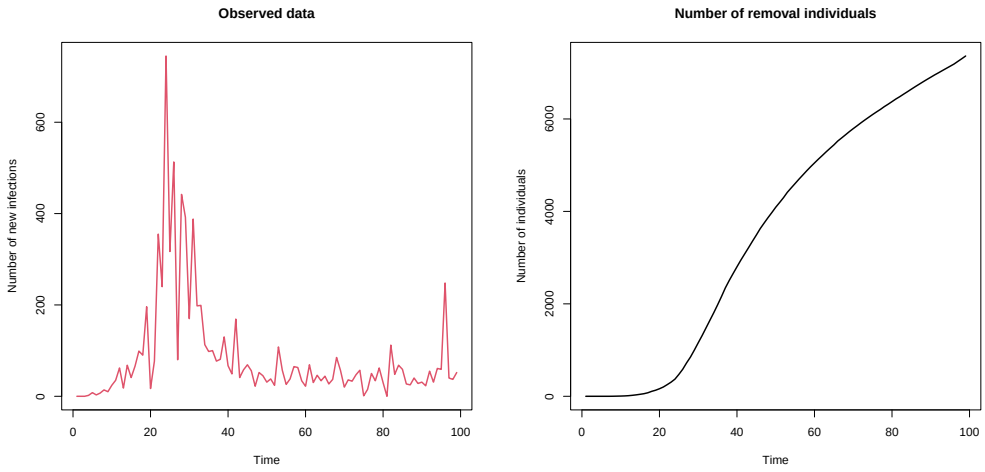

Figure 5.22: Covid-19 data set: [Left] New cases reported in 100 days. [Right] The removed individuals in 100 days.

5.5.2 Bayesian analysis and posterior prediction

First, we determine our prior distribution. Since Covid-19 epidemic was very complicated, we let our $\gamma = 1/20 = 0.05$. We choose $\beta \sim \text{Gamma}(0.01, 0.01)$. We then update our belief in the β using the data. Our goal is to estimate the basic reproduction number $\mathcal{R}_0$ and predict the new cases and the final size of the epidemic.

First case: We have 20 days of data

First, assume that we have 20 days of data. The posterior distribution of β is

$$\beta \sim \text{Gamma}(743.01, 2200.341)$$

The Bayesian estimation of β is 0.3377 and the 90% credible interval of β is $(0.3176, 0.3583)$. By using the definition, $\mathcal{R}_0 = \dfrac{\beta}{\gamma}$, we can derive the distribution of the basic reproduction number $\mathcal{R}_0$. The Bayesian estimation of $\mathcal{R}_0$ is 6.756, and the 90% credible interval of $\mathcal{R}_0$ is $(6.35, 7.17)$. Figure (5.23) depicts the distribution of $\mathcal{R}_0$. The red line in the figure indicates $\mathcal{R}_0 = 1$.

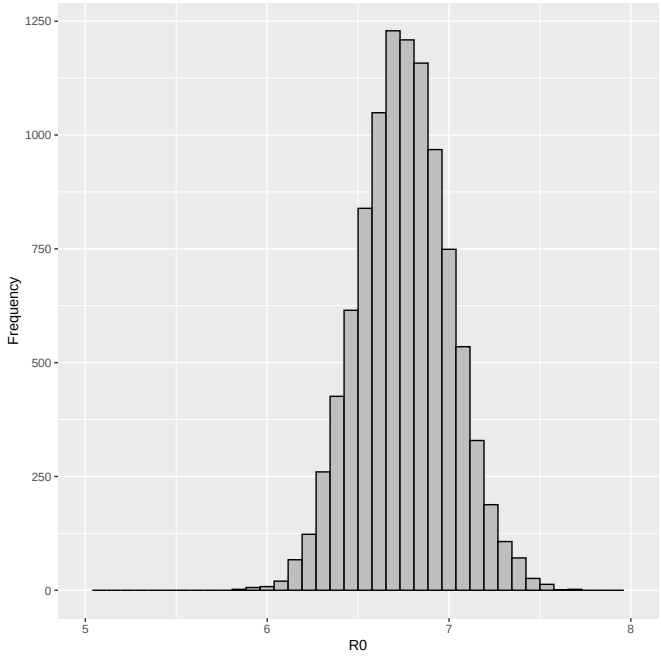

Figure 5.23: Covid-19 data set: We use 20 days of data. The distribution of $\mathcal{R}_0$

First, we use 20 days of data to predict new cases in the next 5 days. Figure (5.24) shows that the prediction can only catches one data point out of five data points. Next, we use these 20 days of data to predict new cases in the next 80 days of the epidemic. Figure (5.24) shows that the we over-estimate the new cases in the next 80 days. In particular, on day 29, there were only 170 cases, where as our model suggests that there are 2513 cases. This is because the New Orleans government has issued a stay-at-home order for all non-essential workers on March

23, 2020. This leads to the decrease in new cases, and the decrease in final size of the outbreak as well. Figure (5.25) shows that the we overestimate the final size of the epidemic, in which most of the population gets the disease. In particular, the model estimates the final size to be 416850, and the 90% credible interval is $(409591.05, 422567.30)$. However, the true value of the final size is 7418, which is significantly smaller than the prediction. Therefore, in the Covid-19 case, our model fails to estimate the transmission rate of the disease and fails to predict the final size of the epidemic.

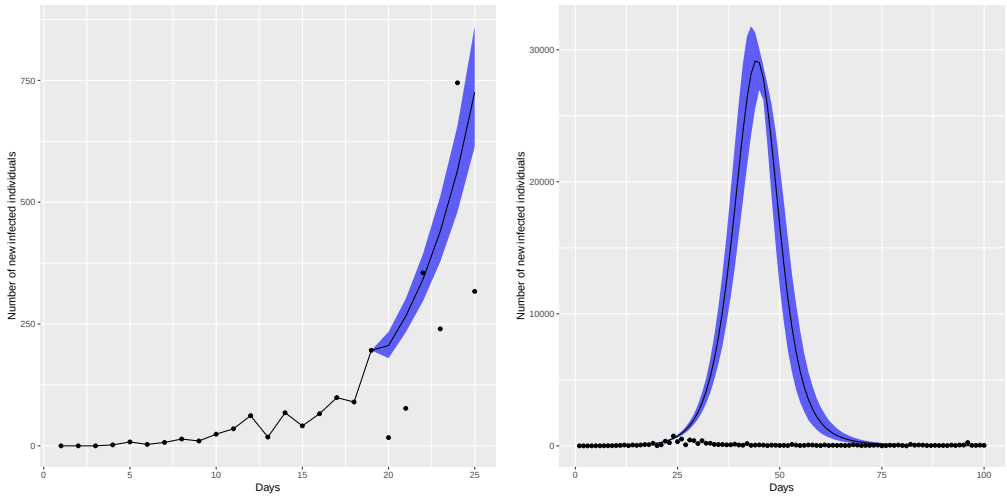

Figure 5.24: Covid-19 data set: We use 20 days of data to predict the next 5 days (left) and the next 80 days. The mean and 90% credible intervals for Bayesian prediction of future infected cases compared to the true value of the epidemic.

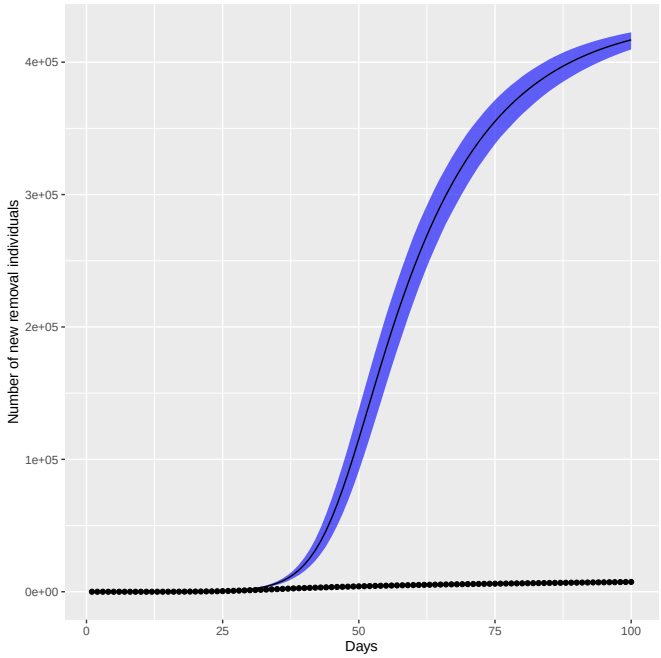

Figure 5.25: Covid-19 data set: We use 20 days of data to predict the final size. The mean and 90% credible intervals for Bayesian prediction of future infected cases compared to the true value of the epidemic.

Second case: We have 50 days of data

Now, we assume that we have 50 days of data. The posterior distribution of β is

$$\beta \sim \text{Gamma}(6103.01, 72436.51)$$

The Bayesian estimation of β is 0.0843 and the 90% credible interval of β is $(0.0825, 0.086)$. By using the definition, $\mathcal{R}_0 = \dfrac{\beta}{\gamma}$, we can derive the distribution of the basic reproduction number $\mathcal{R}_0$. The Bayesian estimation of $\mathcal{R}_0$ is 1.685, and the 90% credible interval of $\mathcal{R}_0$ is $(1.65, 1.72)$. Figure (5.26) depicts the distribution of $\mathcal{R}_0$ in this case. The red line in the figure indicates $\mathcal{R}_0 = 1$.

We still predict the new cases in 5 days and then in the next 50 days. Figure (5.27) shows that the prediction fail to predict the new cases. Figure (5.28) also shows that the we overestimate the final size of the epidemic. However, the final size is more accurate if we have 50 days of data instead of 20 days of data. In particular, the estimation of the final size is 16686 and the 90% credible interval is

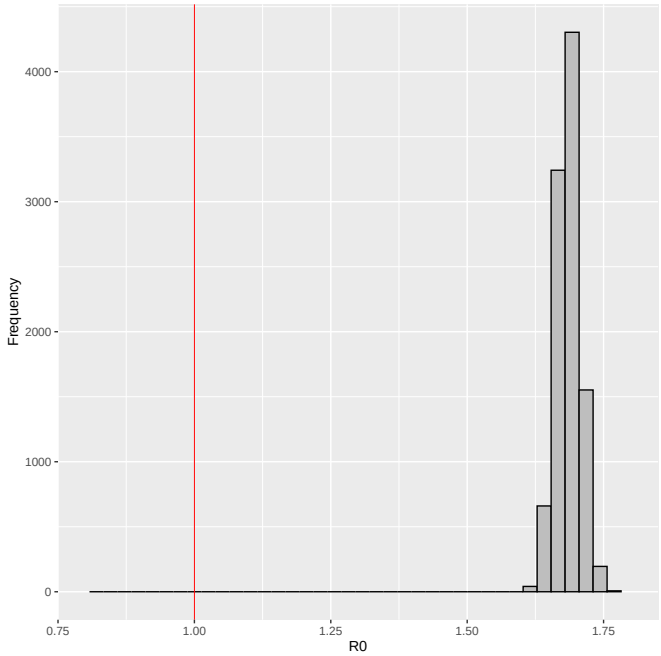

Figure 5.26: Covid-19 data set: We use 50 days of data. The distribution of $\mathcal{R}_0$

$(15783.20, 17599.05)$, which means that not everyone got the infected. Therefore, the model still fails even if we have more data.

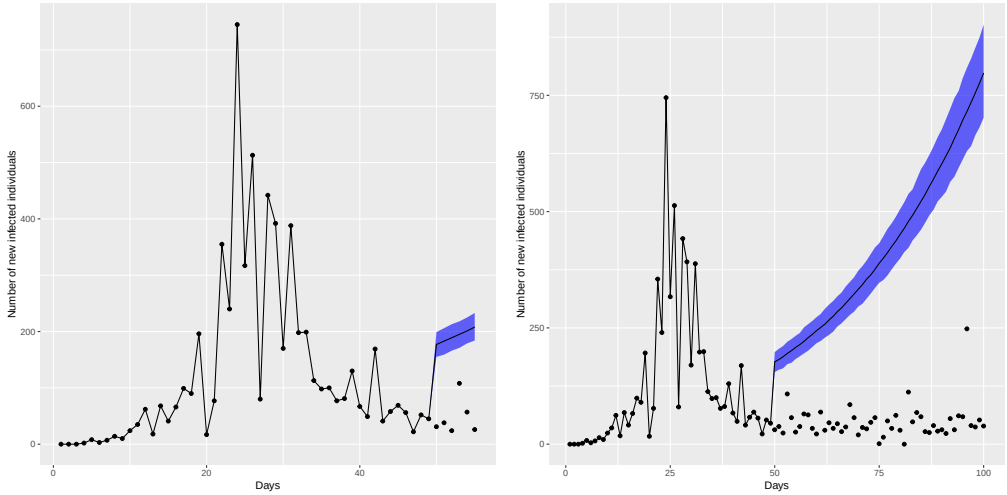

Figure 5.27: Covid-19 data set: We use 50 days of data to predict the next 5 days (left) and 50 days. The mean and 90% credible intervals for Bayesian prediction of future infected cases compared to the true value of the epidemic.

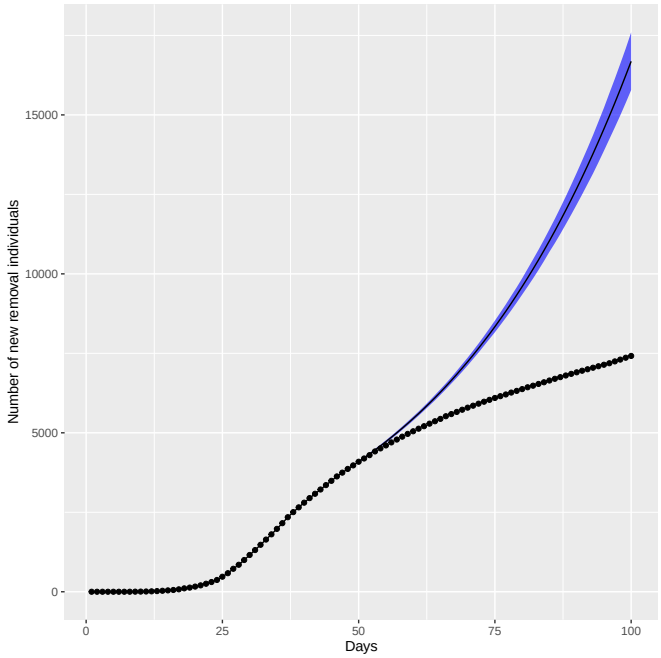

Figure 5.28: Covid-19 data set: We use 50 days of data to predict the final size. The mean and 90% credible intervals for Bayesian prediction of future infected cases compared to the true value of the epidemic.

5.6 Limitations of the methodology

Drawing from the analysis in the Covid-19 analysis, it is evident that the methodology we introduced has some limitations. This methodology, while providing predictions of new daily cases and final size of the epidemic, is subject to several limitations that must be considered. We conclude the limitations as follow

1 Our study for the Covid-19 has a major problem because we did not include the latent period in our model. This is important because the latent period is very common in studying infectious diseases. Including the latent period helps us understand how the process develops over time.

2 The limitation of our study is that we used constant parameters, β and γ, in our model. However, during an epidemic, especially Covid-19, behaviors and conditions change rapidly and frequently. For example, New Orleans detected the first case in March 9th, 2020. On March 14th, all schools in New Orleans closed. On March 23rd, Gov. John Bel Edward's issued a stay-at-home order

for all non-essential workers. On May 16th, New Orleans phase one reopening plan goes into effect. Since our parameters do not adjust to these changes, our model may not accurately reflect what really happens during an epidemic. This means our results only predicts the process if there was no behavior or policy changes. The changes in policy also reduces the number of infectious individual. Therefore the final size of the epidemic was relatively small (7418) compared to the population of New Orleans. Since $\gamma = 0.05$ and the transmission rate is quite large in the beginning of Covid-19, the basic reproduction number $\mathcal{R}_0$ becomes extremely large. Therefore, our model predicts that everyone in the epidemic would get the disease.

Chapter 6

Conclusions and future work

This project provides a comprehensive overview and application of deterministic and stochastic SIR models for studying infectious disease outbreaks. Our main work introduce a model for Bayesian posterior prediction and updating of new daily cases and final outbreak size as new data becomes available over time. While the method works well in some outbreak scenarios, analyzing the COVID-19 data highlights key limitations of the simple SIR modeling assumptions, such as not accounting for time-varying parameters, population behavioral changes, and lacking a latent disease state. These limitations lead to several promising future research building upon this work:

- We can assume that the transmission rate and the removal rate are time-varying by introduce the indicator vector. This indicator vector partition the pandemic wave into several stages. Instead of being constant throughout the whole period, the parameters in the model are homogenous within each stage and change after moving to another period. This method is given in Gu and Yin [4].

- Extend the framework to more advanced epidemic models like SAIR (Susceptible-Asymptomatic-Infectious-Removed) that account for exposed/latent periods before an individual becomes infectious. This would make it more applicable to diseases like COVID-19. An proper analysis using the deterministic SAIR model is given in [2]. We can develop a stochastic version for the SAIR model.

Bibliography

[1] L. J. S. Allen. *An Introduction to Stochastic Processes with Applications to Biology*. Taylor and Francis, second edition, 2011.

[2] S. Ansumali, S. Kaushal, A. Kumar, M. K. Prakash, and M. Vidyasagar. Modelling a pandemic with asymptomatic patients, impact of lockdown and herd immunity, with applications to sars-cov-2. *Annual Reviews in Control*, 50:432–447, 2020. doi: 10.1016/j.arcontrol.2020.10.003.

[3] R. Durrett. *Essentials of Stochastic Processes*. Springer International Publishing, Cham, 2016.

[4] J. Gu and G. Yin. Bayesian sir model with change points with application to the omicron wave in singapore. *Scientific Reports*, 12(1):20864, 2022. doi: 10.1038/s41598-022-25473-y.

[5] P. D. O'Neill. Inference for an epidemic when susceptibility varies. *Biostatistics*, 2(1):99–108, 2001. doi: 10.1093/biostatistics/2.1.99.

[6] P. D. O'Neill. A tutorial introduction to bayesian inference for stochastic epidemic models using markov chain monte carlo methods. *Mathematical Biosciences*, 180(1-2):103–114, 2002. doi: 10.1016/s0025-5564(02)00109-8.

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

[8] O. Wahltinez et al. Covid-19 open-data a global-scale spatially granular meta-dataset for coronavirus disease. 2022. URL `https://goo.gle/covid-19-open-data`.

[9] C. Ward, R. Deardon, and A. M. Schmidt. Bayesian modeling of dynamic behavioral change during an epidemic. *Infectious Disease Modelling*, 8(4):947–963, 2023. doi: 10.1016/j.idm.2023.08.002.

About the author

The author completed his undergraduate education at Ho Chi Minh University of Education, graduating in May 2022 with a Bachelor in Mathematics, and a minor in Education. He receives his Master of Science degree in Applied Mathematics in May 2024. In the future, he continues his PhD in Mathematics at Boston University. In particular, he continues his research in the field of dynamical systems and Bayesian statistics.