Epidemiological characterisation of 1918 pandemic influenza aboard ships

Lachlan Bubb

February 16, 2017

Thesis submitted for the degree of
Master of Philosophy
in
Applied Mathematics & Statistics
at The University of Adelaide
Faculty of Engineering, Computer and Mathematical Sciences
School of Mathematical Sciences

SUB CRUCE LUMEN

THE UNIVERSITY
of ADELAIDE

Contents

Signed Statement

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Signed: .... Date: ....

Acknowledgements

First and foremost, I would like to thank my supervisors, Dr Joshua Ross and Dr Andrew Black, for their consistent wisdom and encouragement. This thesis would not be a quarter of what it is today without them. I would also like to thank my family and friends for the gratuitous patience, goodwill and laughter I have received over the last two years. In particular, I would like to thank my mother, Leigh, and partner, Bonnie, for their endless support in getting me to this point. Lastly, thank you to everyone who has ever contributed to the kind words and thoughtful conversations that have inspired my academic pursuits.

Abstract

Influenza has been the principal cause of pandemic events over the last century. As such, strategies must be implemented to reduce the potential impact of future pandemics. These epidemic control measures should be informed by the epidemiological characteristics of the disease, but our current understanding of influenza is wanting. Here, we study one of the worst pandemics: the *1918 "Spanish flu" pandemic*. Viral descendants of the 1918 influenza strain are still in circulation today, such as the 2009 influenza pandemic virus. Hence, there is significant motivation to study the epidemiological characteristics of the strain responsible for the 1918 pandemic to best inform the development of control measures against future pandemics.

Past epidemiological studies of the 1918 pandemic have been restricted to data and epidemic models that fail to account for important dynamics, or ignore external factors which could potentially bias results. Here, we investigate a previously unstudied data source of contained influenza outbreaks from the 1918 pandemic that alleviates these issues. Data of 15 influenza outbreaks aboard naval and passenger vessels travelling to Australia has been collated. These *on-board epidemics* are natural *pseudo experiments* of influenza transmission; contained outbreaks replicating transmission experiments with a level of recorded detail unprecedented for the 1918 pandemic. To exploit the data, we develop a novel stochastic epidemic model that accounts for features salient to characterising the epidemiology of the 1918 pandemic strain; these include asymptomatic infections, the pre-symptomatic infectious period and prior immunity. To validate our approaches, an extensive investigation of inference methods and parameter identifiability is conducted.

Parallel inference across multiple ship outbreaks is used to characterise the 1918 pandemic influenza strain and enables comparison across pandemic waves. We find that 1918 pandemic was caused by a highly transmissible virus, and the reduced impact between the second and third pandemic waves was a result of significantly increased population immunity. We find evidence indicating individuals are infectious for a significant period of time (approximately 20 hours) before the development of symptoms. Most importantly, we find transmission from *non-symptomatic* individuals, that is, infectious individuals that are asymptomatic or in the period prior to onset of symptoms, was the dominant cause of infection aboard these ships.

Chapter 1

Introduction

Epidemics have been a consistent presence in humanity's history. In the modern day, diseases such as HIV/AIDS, measles, smallpox and the more recent Ebola and Zika viruses have caused epidemics and pandemics responsible for countless deaths [1, 16, 34, 59, 63]. As such, strategies must be implemented to reduce the potential impact of any future epidemics. Crucial to the development of these control measures is an in-depth understanding of the *epidemiological characteristics* of the virus [25].

The most common cause of pandemic events is the *influenza* virus. Commonly known as the flu, influenza is a highly contagious viral disease transmitted via water vapour emitted from the mouth and nose. Webster *et al.* [78] contains a detailed description of the influenza virus. Here, we discuss the influenza A type virus as associated with pandemic influenza. Influenza pandemics usually originate from the emergence of a novel strain of the virus, unique to previous endemic influenza strains [77, 78]. These strains develop in other species by a process of *antigenic shift* (genetic combination of multiple viruses) before being transmitted to humans; the new strain is largely unaffected by immune responses from previous attacks of the influenza virus. Hence, the virus can rapidly spread within the unprotected population with potentially global reach.

In the last century, there has been four influenza pandemics with true global reach [78]. The worst of these was the *1918 "Spanish flu" pandemic*, which will be focus of this thesis and is discussed further below. The 1918 influenza pandemic was a natural disaster of catastrophic proportions, in both scope and severity. Over the course of 1918-1919, the particular influenza virus is estimated to have infected a third of the world's population and caused approximately 50 million deaths [74]. However, the relevance of the 1918 pandemic to modern influenza cannot be understated. Alarmingly, almost all of today's influenza A strains are descendants of the 1918 virus; this includes the pandemic H1N1 virus, notably the cause of the 2009 influenza pandemic [56, 74]. Hence, there is significant motivation for the epidemiological factors responsible for the 1918 pandemic to be further understood.

1.1 1918 Influenza Pandemic

The 1918 influenza pandemic may be the 'mother of all pandemics' [74], but it is still a relatively unknown quantity for its many atypical pandemic qualities. The geographical origin of the pandemic is open to debate, but it is thought to have originated within the USA in March 1918 [61]. Over the next few months, the virus spread across Europe, Asia and North Africa; eventually reaching Australia in July 1918. This period is known as the *first wave* of the pandemic; the smallest of three waves. The first wave was characterised by mortality rates not dissimilar to seasonal influenza [40, 61]. A highly fatal *second wave* re-emerged in September and spread globally until November-December. The second wave demonstrated much greater virulence than the previous wave. This wave had a significant increase in morbidity and mortality rates, and was infamously deadly to healthy adults; atypical of pandemic influenza [77]. A *third wave* re-emerged in early 1919, although was not as severe as the second wave and only appeared within some countries [74]. The three waves of the 1918 pandemic are notable as while influenza pandemic waves are not uncommon, the rapid progression of all three waves in approximately 12 months and the differing levels of virulence is unprecedented. There are several theories as to the cause of the wave behaviour and the extreme morbidity of the disease. Factors such as the increasing presence of immunity and *antigenic drift* (viral genetic mutation) between waves are thought to have contributed, but the exact cause is unclear [74]. Hence, many questions about the epidemiology of the 1918 influenza pandemic remain unanswered. As many viral descendants of the pandemic strain are in circulation today, it remains a notable worst case scenario of influenza pandemics. Hence, the virus requires comprehensive epidemiological characterisation to best inform the development of control measures against potential future attacks.

There have been numerous mathematical studies of the 1918 pandemic due to its significance, to varying degrees of success. These previous studies typically use data from citywide case/mortality records [11, 17, 23, 54], household studies [26], or institutional records such as schools or military bases [19, 48, 49, 70]. As such, these investigations suffer from the typical pitfalls of epidemic characterisation: the complications of unknown external factors. Unidentified dynamics such as heterogeneous-mixing populations and externally-introduced transmission are notoriously difficult to account for, and require the use of questionable assumptions and complicated modelling approaches. Often inappropriate models are used that approximate or ignore these dynamics, but such methods could fail to capture, or possibly bias, the estimation of important characteristics. Here, we characterise the 1918 pandemic by investigating a previously unstudied data source of contained outbreaks that mitigates some of these difficulties.

The 1918 pandemic coincided with the end of WWI, and the mass transportation of troop and trading vessels during this period played a critical role in the global spread of the disease. In response, the Australian government implemented national maritime quarantine of all vessels attempting to enter Australia. *Influenza and maritime quarantine* [20] is a service publication of the Australian quarantine during the 1918 pandemic. The report contains a register of all ships entering Australia during this time, and includes detailed records from aboard vessels that have travelled to Australia while carrying influenza-infected passengers. The resulting *on-board epidemics* are natural *pseudo experiments* of pandemic influenza transmission; contained outbreaks replicating transmission experiments with a level of recorded detail unprecedented for the 1918 pandemic. As the Australian Director of Quarantine, J.H.L. Cumpston, states:

> *'The conditions on a ship at sea offer an almost ideal opportunity for studying the natural history of any infectious disease'* [20].

The contained nature and records of the outbreaks alleviates some of the external factors described previously and allows for better identification of the true dynamics. Here, we have collated data of 15 influenza outbreaks aboard ships (predominately naval vessels) with documentation including: daily resolution case counts, port arrival/departure dates, use of inoculations/quarantine measures and timings, and all landings of healthy and infected passengers. The collection of data gives an unparalleled picture of the evolution of a contained outbreak of the 1918 pandemic influenza strain. The data is used to epidemiologically characterise the 1918 pandemic by conducting inference using epidemic models.

1.2 Epidemic Models

Mathematical models for the spread of an infectious disease through a population are known as *epidemic models.* Epidemic models are commonly formulated as *compartmental models*, where an individual's infection status is denoted by their progression through a series of status-defined compartments [3, 4, 41]. The movement between compartments is governed by a set of parameters, which provides insight into the behaviour of an epidemic. The original and most common compartmental model is the *SIR model*, developed by Kermack and McKendrick [41, 43]. The SIR model represents transmission of a disease with immediate infectiousness upon infection and complete immunity post recovery. Hence, a population of individuals is divided into compartments of *Susceptible* (S) - able to be infected, *Infectious* (I) - infected and infectious, *Recovered* (R) - recovered and now immune to further infection. For clarity, a diagram of the possible compartments an individual can be within the SIR system is given in Figure 1.1.

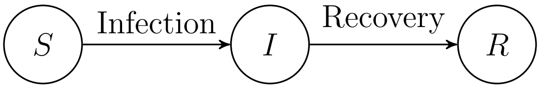


Figure 1.1: Compartment diagram of SIR system.

Compartmental models can be designed to replicate specific disease dynamics. To appropriately model influenza, the *course of infection* must be scrutinized to inform the model development [15, 69]. Once an individual is infected with influenza, there exists a period of time until the individual is infectious to others, denoted the *latency period.* They are then infectious for a period of time before developing clinically observable symptoms; this period is denoted the *pre-symptomatic infectious period.* After the development of observable symptoms, the individual can be observed as a *clinical illness.* The individual will remain infectious to others until they recover from the virus, where they will exhibit a duration of immunity from further infection.

Although, not all individuals who contract the virus will be observed as a clinical illness. There is also the possibility of *asymptomatic infections*, where infected individuals display mild or zero symptoms, or are unable to have their infection recognised, and are not clinically observed. The susceptibility of the population should also be scrutinized, via *prior immunity.* Prior immunity is the proportion of the host population with immunity to the disease prior to epidemic onset; this arises either from a previous wave of the pandemic or prior attacks of a related seasonal influenza strain. Asymptomatic infections and prior immunity are particularly important to pandemic influenza, where case recognition rates can be low and immunity can be gained from past pandemic waves [48, 49]. Past modelling approaches (including those previously mentioned) rarely account for all of these dynamics as the complexity of the model required inhibits analysis. Hence, to accurately characterise pandemic influenza, complicated disease dynamics must be included within the model.

1.3 Epidemic Inference

Epidemiological characterisation is informed by conducting parameter inference on outbreak data using epidemic models. Dependent on the model choice, there are numerous quantities of interest that can be derived from epidemic inference. Of particular importance, Fraser *et al.* [25] lists the factors needed to be estimated to determine effective control measures for an epidemic. Here, we detail and expand on these parameters, including some that have previously been studied for the 1918 pandemic but lack sufficient evidence. The *basic reproduction number*, R_0, is defined as the expected number of secondary infections caused by an average infectious individual in an otherwise susceptible population [41]. If R_0 is greater than one, it indicates the spread of disease may reach epidemic levels. As such, an accurate estimate of R_0 is critical in understanding the likelihood of disease invasion and its potential impact. Previous studies of the 1918 pandemic using city-wide data estimate R_0 between 1.4 - 2.0 [54], but as R_0 is population dependent, estimates from schools and military base outbreaks can be as high as 20 [25, 49]. Hence, there is much uncertainty about influenza transmission rates in contained environments.

There are several epidemiological periods of interest as they determine the rate of epidemic growth. The *incubation period* is the time between an individuals infection and development of symptoms [64]. Previous studies have estimated the incubation period (sometimes approximated by the latency period) as 1.3 - 2.0 days [48, 54]. The *generation time* is the time between a primary individual's infection and their first transmission to another individual [64]. There is minimal conclusive evidence of the generation time for the 1918 pandemic influenza, due to its R_0 dependence, but previous estimates have been around 2.3 - 3.0 days [26, 48, 54]. The incubation period and generation time can determine the rate of epidemic growth in the number of observed cases, and hence deserve consideration alongside R_0.

Prior immunity is also an important factor to consider for pandemic influenza. Previous studies have shown that prior immunity increased between waves [49], where the exact immune proportion is dependent on the population. This is not surprising as previous attacks of seasonal influenza can reduce susceptibility to pandemic influenza, and immunity can be gained from earlier waves of the pandemic [48]. Prior immunity estimates suggest the susceptible population prior to the first wave was the range of 50 - 80% [26, 48, 49]. Although, there is little in the literature estimating the proportion of susceptibility in later waves. The *symptomatic proportion*, the proportion of influenza infections that present observable symptoms, is also of interest. Past studies have found estimates are varied, ranging from as low as 38% to as high as 91% [48]. The number of asymptomatic infections is related to prior immunity in that it is difficult to determine if individuals not observed have escaped infection, are immune, or are asymptomatic but still infectious [26]. Hence, the difficulty in estimating prior immunity and the symptomatic proportion are linked.

An important factor of pandemic influenza which critically lacks evidence is the proportion of transmission arising from *non-symptomatic* individuals [60]; that is, the number of infections that occur by transmission from infectious individuals either asymptomatically or prior to onset of symptoms. As these individuals are not, or not yet, clinically observed but still contribute to transmission, identifying their proportion of transmission is crucial to identifying the best choice of intervention [25]. Previous estimates have found a plausible range for the proportion of transmission from non-symptomatic individuals for pandemic influenza to be 30 - 50%, but review literature has found little conclusive evidence [60]. Hence, the proportion of transmission arising from symptomatic and non-symptomatic individuals has yet to be accurately captured for influenza and the 1918 pandemic.

Within this thesis, we characterise the 1918 pandemic influenza strain by conducting inference on ship outbreak data collated from Cumpston [20]. We use a Bayesian approach to infer the joint distribution for the parameters of an epidemic model specifically designed for this study. Due to the complexity of the model required to capture the desired pandemic influenza characteristics, a recently developed simulation-based algorithm is used. Inference is conducted on collections of ship outbreaks corresponding to the second and third pandemic waves. The parallel inference allows for epidemiological characterisation of the 1918 influenza pandemic including estimates of quantities above, but also comparison across pandemic waves for key dynamics such as transmission rates and prior immunity.

1.4 Thesis Outline

In Chapter 2, the technical background relevant to the work in this thesis is presented. Chapter 3 details the development of an appropriate stochastic epidemic model for pandemic influenza. The basic SIR model is built upon using extensions derived from the literature to develop a novel epidemic model, denoted the *SEIpIsIaR* model. A rigorous investigation of inference methods is conducted to validate the use of an simulation-based inference algorithm and test the parameter identifiability of the model. This includes an investigation into the benefits of parallel inference in identifying key dynamics of interest. Also presented is an alternative inference method not appropriate for the current study but that may have applications within the field. In Chapter 4, the epidemiological characterisation of the 1918 pandemic is presented. The circumstances surrounding the ship outbreaks are thoroughly illustrated, and the *ship epidemic model* is developed to account for any changes in population/transmission dynamics on board. The ship epidemic model is used to conduct inference on the ship outbreak data in wave-based groups, and key findings are discussed in detail. Chapter 5 presents a summary discussion of the key findings, limitations and consequences of the work within this thesis. Included are possible further areas of study into the 1918 pandemic from the data within Cumpston [20].

Chapter 2

Technical Background

This chapter will give an overview of the technical background and literature relevant to this thesis. The topics include Markov chains, epidemic models and Bayesian inference.

2.1 Markov Chains

Markov chains are a class of random processes that obey a characteristic property known as the *Markov property.* The Markov property states that the random process retains no memory of its past and only the current state of the process influences its future. This property is extremely important as it allows the tractable calculation of many quantities of interest for random processes. As such, Markov chains are commonly used in applied mathematics for a variety of applications including epidemic modelling (see Section 2.2) and Bayesian inference (see Section 2.3). The following section is a discussion of the relevant sections of Markov chain theory for these applications. The theory of this section is explained in greater detail in Kroese *et al.* [44], Norris [58] and Ross [68].

Markov chains are formulated in *discrete-time* or *continuous-time* with similar definitions but distinct properties. Within this thesis, discrete-time Markov chain theory is relevant for Markov chain Monte Carlo (Section 2.3.1) and our stochastic epidemic models are formulated as continuous-time Markov chains (Section 2.2).

2.1.1 Discrete-time Markov Chains

A *discrete-time Markov chain* (DTMC) is a discrete collection of random variables $\{X_n\}_{n\in\mathbb{N}}$ on a countable state space $\mathcal{S}$ which obeys the Markov property as follows.

Definition 2.1.1. A discrete-time stochastic process $\{X_n\}_{n\in\mathbb{N}}$ on a countable state space $\mathcal{S}$ is a DTMC iff

$$P(X_{n+1} = j|X_1 = i_1, X_2 = i_2, ..., X_n = i_n) = P(X_{n+1} = j|X_n = i_n),$$
$$\forall\ n \in \mathbb{N},\ j, i_1, ..., i_n \in \mathcal{S}.$$

That is, if X_n denotes the current state of the process at time n, then the next state of the process X_{n+1} is only dependent on the current state and no prior history.

For this thesis, the special case of *time-homogeneous* Markov chains are considered. A DTMC $\{X_n\}_{n\in\mathbb{N}}$ is time-homogeneous iff

$$p_{ij} := P(X_{n+1} = j|X_n = i) = P(X_1 = j|X_0 = i),\ \forall\ n \in \mathbb{N},\ i, j \in \mathcal{S}.$$

The p_{ij} defined above are the *one-step transition probabilities* of $\{X_n\}_{n\in\mathbb{N}}$. That is, p_{ij} is the probability of $\{X_n\}_{n\in\mathbb{N}}$ leaving state i and entering state j in a single time-step. These probabilities can be represented in matrix form by the *transition probability matrix*, an $|\mathcal{S}| \times |\mathcal{S}|$ matrix defined by

$$P = (p_{ij} : i, j \in \mathcal{S}).$$

The *m-step transition probability matrix*, the probabilities of moving from state i to state j in m steps, is defined by the matrix power $P^{(m)} = P^m$.

For use in Markov chain Monte Carlo (Section 2.3.1), the long term behaviours of DTMCs are of interest. Denote the *stationary distribution* as the vector $\pi = (\pi_j : j \in \mathcal{S})$ which represents the equilibrium probability distribution of the Markov process, such that, $\pi = \pi P$. The stationary distribution is given by the unique solution to the following system of equations [44]:

$$\begin{aligned} &0 \leq \pi_j \leq 1\ , \\ &\sum_{j\in S} \pi_j = 1\ , \\ &\pi = \pi P\ . \end{aligned} \tag{2.1}$$

Given $\{X_n\}_{n\in\mathbb{N}}$ obeys a set of regularity conditions (see Norris [58], Theorem 1.8.3), it is possible to determine the stationary distribution by the process's *limiting distribution*. The limiting distribution is the long run proportion of time that $\{X_n\}_{n\in\mathbb{N}}$ spends in state $j \in \mathcal{S}$, independent of initial state $X_0 = i$. This can be expressed as

$$P(X_n = j) = \frac{1}{n}\sum_{m=1}^{n} P_{ij}^{(m)} \ , \ \ \forall j \in \mathcal{S},$$

where the limiting distribution will converge to the unique stationary distribution as $n \to \infty$.

Given a sequence of outcomes from $\{X_n\}_{n\in\mathbb{N}}$ denoted as a *sample-path*, the limiting distribution can be estimated by the sample-path long-run proportion of time that $\{X_n\}_{n\in\mathbb{N}}$ spends in each state $j \in \mathcal{S}$. Hence, it is possible to estimate the stationary distribution of a DTMC through sample-path simulation [28, 44]. This technique is exploited in Markov chain Monte Carlo (Section 2.3.1).

2.1.2 Continuous-time Markov Chains

A *continuous-time Markov chain* (CTMC) is analogous to a DTMC $\{X_n\}_{n\in\mathbb{N}}$, except the process evolves in continuous time, $\{X(t)\}_{t\geq 0}$ where $t \in \mathbb{R}$.

Definition 2.1.2. A continuous-time stochastic process $\{X(t)\}_{t\geq 0}$ on a countable state space $\mathcal{S}$ is a CTMC iff

$$P(X(t+s) = j|X(u) = k, X(s) = i, u < s) = P(X(t+s) = j|X(s) = i), \\ \forall\ s,t,u \geq 0,\ i,j,k \in \mathcal{S}.$$

That is, analogous to a DTMC, the Markov property is satisfied. Note, although defined on a countable state space, this thesis will focus on finite state space CTMCs.

We define a CTMC $\{X(t)\}_{t\geq 0}$ to be time-homogeneous iff

$$p_{ij}(t) := P(X(t+s) = j|X(s) = i) = P(X(t) = j|X(0) = i),\ \ \forall\, s,t \geq 0,\ i,j \in \mathcal{S}.$$

The $p_{ij}(t)$ defined above are the *transition function probabilities* of $\{X(t)\}_{t\geq 0}$, the probability of moving from state i to state j in elapsed time t. These probabilities can be represented in matrix form by the *transition function*, an $|\mathcal{S}| \times |\mathcal{S}|$ matrix defined by

$$P(t) = (p_{ij}(t) : i, j \in \mathcal{S}).$$

The *infinitesimal generator* or “Q-matrix” of a CTMC, which denotes the *rate* of the process moving from state to state, is defined by

$$Q = \lim_{h \to 0^+} \frac{P(h) - I}{h}.$$

Define $Q = (q_{ij} : i, j \in \mathcal{S})$. For $i \neq j$, q_{ij} is the *transition rate* of moving from state i to state j. For the diagonal elements, q_{ii}, define

$$q_i := -q_{ii} = \sum_{\substack{j \in \mathcal{S} \\ j \neq i}} q_{ij}$$

to be the rate of leaving state i.

The transition function can be evaluated as a solution to the *Kolmogorov forward equation* [58]. The Kolmogorov forward equation states

$$\frac{d}{dt} P(t) = P(t)Q. \tag{2.2}$$

Define the *probability mass function* of $\{X(t)\}_{t \geq 0}$,

$$p(t) = (P(X(t) = i) : i \in \mathcal{S}).$$

By Equation (2.2) and assuming initial probability distribution $p(0)$, the probability mass function at time t can be evaluated as the solution to the *master equation* [42],

$$\frac{d}{dt} p(t) = p(t)Q. \tag{2.3}$$

Hence, the probability mass function can be found from the numerical solution to a set of $|\mathcal{S}|$ linear ordinary differential equations. There are various numerical solver techniques for ordinary differential equations [55]. Although, depending on the complexity of the system and Q-matrix structure, these methods can become prohibitively computationally expensive as the state space, and therefore size of the ODE system, increases [39].

Alternatively, solving the Kolmogorov forward Equation (2.2) analytically gives the transition function solution as

$$P(t) = e^{Qt}. \tag{2.4}$$

Hence, by Equation (2.4) and assuming initial probability distribution $p(0)$, the probability mass function is given by the matrix exponential solution,

$$\begin{aligned} p(t) &= p(0)P(t) \\ &= p(0)e^{Qt}. \end{aligned} \tag{2.5}$$

Rarely is the solution to the matrix exponential analytically available. Hence to evaluate the probability mass function via Equation (2.5), a numerical approximation is required. An efficient tool for evaluating the numerical approximation is the Matlab software package `Expokit` [71], which uses Krylov subspace projection techniques. `Expokit` is a practical solution for evaluating the probability mass function for Markov processes with small state spaces, but again, as the state space increases the efficiency of the algorithm greatly decreases [55]. Hence, other methods of calculating the probability mass function are required for Markov processes with large state spaces. A common approach is to use simulation-based estimates.

As $\{X(t)\}_{t\geq 0}$ evolves continuously in time, the transition between being within a current state i and moving to a future state j can be mapped to a pair of random variables: the *holding time*, the time until the process leaves state i once entered, and the *jump chain* which denotes the probability of entering state j upon leaving state i. The value q_i determines the holding time of being in state i by a fundamental theorem of CTMCs [58, 68].

Theorem 2.1.3. *(Ross [68]) The holding time of $\{X(t)\}_{t\geq 0}$ being in state i, T_i, is exponentially distributed with mean*

$$E[T_i] = \begin{cases} \frac{1}{q_i}, & \text{if } q_i > 0, \\ \infty, & \text{if } q_i = 0. \end{cases}$$

Upon the expiration of the holding time T_i, it is possible to determine the probability distribution of the jump chain. That is, the probability of entering state $j \neq i$ upon leaving state i.

Theorem 2.1.4. *(Ross [68]) Upon leaving state i, $\{X(t)\}_{t\geq 0}$ will enter state $j \neq i$ with probability,*

$$P(X(t) \text{ enters state } j \neq i, \text{ when leaving } i) = \frac{q_{ij}}{q_i}.$$

The decomposition of a CTMC into a holding time and jump chain process allows for quick sample-path simulation by a stochastic simulation algorithm. For our purposes all simulations will be generated using the stochastic simulation algorithm as defined in Algorithm 1 (sometimes denoted the

Gillespie algorithm [29]). For these simulations the initial state of the process $X(0)$ and a predetermined *exit condition* must be specified.

```
Data: Q matrix, $Q$, initial state $X(0)$, exit condition.
Set $t = 0$;
while exit condition false do
    From current state $X(t) = i$, list possible events $E_1, ..., E_k$ and
    their rate of occurrence $q_{i1}, ..., q_{ik}$ ;
    Calculate mean holding time in state $X(t)$,
        $$q_i = \sum_{\substack{j \in \mathcal{S} \\ j \neq i}} q_{ij};$$
    Sample $U_1, U_2 \sim \text{Uniform}(0, 1)$;
    Sample random holding time $\delta t = \frac{1}{q_i} \log(U_1)$ ;
    Set $P = q_i \times U_2$;
    Event $j$ occurs if
        $$\sum_{m=1}^{j-1} q_{im} < P \leq \sum_{m=1}^{j} q_{im};$$
    Set $X(t + \delta t) = j$;
    Update time $t \to t + \delta t$;
end
Result: Realisation of CTMC $\{X(t)\}_{t \geq 0}$.
```

Algorithm 1: Stochastic simulation algorithm.

2.2 Epidemic Models

Epidemic models describe the spread of an infectious disease through a population. This process is approximated by a *compartmental model*, where an individual's infection status is denoted by their progression through a series of status-defined compartments [3, 24, 41].

Here, we focus on *stochastic compartmental models*. These models track the discrete number of individuals within each compartment and reflect the random dynamics of disease transmission [4]. These properties are important in modelling epidemics as the spread of disease from individual to individual is not deterministic, there exists a *probability* of transferring the disease between individuals. Hence, the probability of *epidemic die-out*, where the disease dies out from a population, can only be replicated by stochastic models. Here, we formulate all stochastic models as CTMCs (see Section 2.1.2) for ease of analysis, as common within the literature [3, 4, 41].

Stochastic models should always be used when possible but are often impractical to use for large population sizes. Alternatively, we can use deterministic approximations to a suitably scaled version of the stochastic model by a functional law of large numbers [45, 46]. The deterministic approximations allow for faster computational analysis and are preferable in some situations, but are not used here. Here, we give an introduction to stochastic compartmental models via the *SIR* model.

2.2.1 SIR Model

The SIR model is representative of a disease with immediate infectiousness upon infection and complete immunity upon recovery. A homogeneous population of individuals is divided into compartments defined by the infection status of each individual: *Susceptible* (S) - able to be infected, *Infectious* (I) - infected and infectious, or *Recovered* (R) - recovered and now immune to further infection. The model assumes homogeneous mixing between individuals and a closed population (no birth, death or migration) of fixed size N. Under these conditions, there are two possible transitions an individual can undergo: a susceptible individual becoming infected via contact with an infectious individual (*infection* event), or an infectious individual recovering from the disease (*recovery* event) [2, 13, 41]. A diagram of the possible compartments an individual can be within the SIR system is given in Figure 2.1.

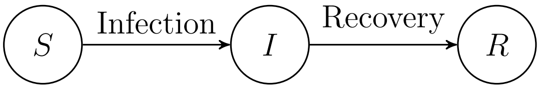


Figure 2.1: Compartment diagram of SIR system.

Other compartments and transitions can be added to the SIR model to better replicate the dynamics of disease transmission. These more complex models will be investigated in Chapter 3.

Let $S(t)$ and $I(t)$ denote the number of susceptible and infectious individuals at any time t, where $R(t) = N - S(t) - I(t)\ \forall t$. A two-variable CTMC that models the SIR process is $\{(S(t), I(t))\}_{t\geq 0}$ with state space

$$\mathcal{S} = \{(S, I) | S, I \in \mathbb{N}, 0 \leq S + I \leq N\}. \tag{2.6}$$

The transition rates from a state (S, I) are displayed in Table 2.1.

Transition Type	State Change	Transition Rate
Infection	$(S, I) \to (S-1, I+1)$	$\frac{\beta SI}{N-1}$
Recovery	$(S, I) \to (S, I-1)$	γI

Table 2.1: SIR compartmental transition rates.

Within Table 2.1, there are two key parameters that control infection and recovery, β and γ respectively. The first of these, β, the *transmission* parameter of the process, is defined as follows. The infection rate can be expressed as

$$q_{(S,I),(S-1,I+1)} = Scpv,$$

where c is the rate of contact between individuals, $p = \frac{I}{N-1}$ is the probability a contact is with an infectious individual, and v is the probability that a susceptible-infectious contact results in disease transmission. If we assume a *frequency-dependent* contact structure within the population, that is the rate of contact is only dependent on the number of individuals within the population, then the rate of contact between individuals c is constant. Hence, denote $\beta = cv$ to be the *frequency-dependent transmission term* and achieve the following result [8],

$$q_{(S,I),(S-1,I+1)} = \frac{\beta SI}{N-1}.$$

Frequency-dependent transmission is popular within epidemic modelling, although the assumption of a constant contact rate c is of much debate [8]. A more "intuitive" and arguably preferential contact structure is known as *density-dependent transmission*. Here, let $c = \frac{\kappa(N-1)}{A}$ so the contact rate between individuals is scaled in proportion to the *density* of individuals within known area A. Hence, denote $\beta^* = \kappa v(N-1)$ to be the *density-dependent transmission term* and achieve the following result [8],

$$q_{(S,I),(S-1,I+1)} = \frac{\beta^* SI}{A}.$$

For this thesis, frequency-dependent transmission is assumed but density-dependent will be utilized where stated. Specifically, density-dependent transmission will be used for inference with the 1918 pandemic ship data, using the available ship dimensions (see Chapter 4).

The remaining parameter γ is the recovery rate of an infectious individuals. By Theorem 2.1.3 of CTMCs, the infectious period of an individual is exponentially distributed with mean $1/\gamma$.

From β and γ defined previously, a critical epidemic indicator for the SIR model (and by extension other epidemic models) is derived. The *basic reproductive number*, R_0, is defined as the expected number of secondary infections caused by an infectious individual in an otherwise susceptible population. For the SIR model R_0 is given by

$$R_0 = \frac{\beta}{\gamma}. \tag{2.7}$$

Its importance is related to the *threshold phenomenon* (see Keeling and Rohani [41], Kermack and McKendrick [43]), where given a single infectious individual in an otherwise susceptible population a disease will not "invade" a population if $R_0 < 1$. This result proved R_0 to be arguably the most critical indicator for the prediction, and control of epidemics in the event of disease outbreaks.

2.2.2 Degree-of-Advancement Representation

Degree-of-Advancement (DoA) is a alternative representation of a compartmental model by counting the number of each transition event [39]. That is, the SIR process defined above can be expressed in an equivalent DoA representation using the counting processes of the number of infection and recovery events.

Let $Z_1(t)$ and $Z_2(t)$ denote the number of infection and recovery events by time t, where the population compartment numbers are given by

$$\begin{aligned} S(t) &= N - Z_1(t), \\ I(t) &= Z_1(t) - Z_2(t), \\ R(t) &= Z_2(t). \end{aligned} \tag{2.8}$$

The SIR process is expressed as a two-variable CTMC $\{(Z_1(t), Z_2(t))\}_{t\geq 0}$ with state space

$$\mathcal{S} = \{(Z_1, Z_2) | Z_1, Z_2 \in \mathbb{N}, 0 \leq Z_1 \leq Z_2 \leq N\}. \tag{2.9}$$

The transition rates from a state (Z_1, Z_2) are displayed in Table 2.2.

Transition Type	State Change	Transition Rate
Infection	$(Z_1, Z_2) \to (Z_1 + 1, Z_2)$	$\frac{\beta(N-Z_1)(Z_1-Z_2)}{N-1}$
Recovery	$(Z_1, Z_2) \to (Z_1, Z_2 + 1)$	$\gamma(Z_1 - Z_2)$

Table 2.2: SIR DoA transition rates.

Expressing a compartmental model in DoA allows the state space to be ordered lexicographically

$$\mathcal{S} = \{(0,0), (1,0), (1,1), (2,0), (2,1), (2,2), ..., (N,N)\}.$$

This produces a Q-matrix of the process that is upper triangular due to the non-decreasing counting processes [9]. The upper triangular structure can be exploited for efficiency in calculating the probability mass function by the implicit Euler method.

Implicit Euler

As discussed in Section 2.1.2, calculating the probability mass function of a Markov process using typical numerical methods such as `Expokit` is a computationally intensive task for even small state spaces and is often infeasible. An alternative method for compartmental models in DoA representation which exploits the upper triangular Q-matrix, as specifically used within this thesis, is the ordinary differential equation solver known as the implicit Euler method.

The *implicit Euler* method calculates the probability mass function by uses a numerical integration approach to solve the master equation (Equation (2.3)), as described in Jenkinson and Goutsias [39]. Let $\{X(t)\}_{t\geq 0}$ denote a CTMC with Q-matrix, Q, and probability mass function, $p(t)$. The implicit Euler method is for discrete time steps $\{t_j = \tau j : \ j = 0, 1, 2, ..., T\}$, the numerical solution of the probability mass function $\hat{p}(t_j)$ can be found from $\hat{p}(t_{j-1})$ by solving the set of linear equations

$$(I - \tau Q)\hat{p}(t_j) = \hat{p}(t_{j-1}). \tag{2.10}$$

As above, a compartmental model in DoA representation produces in a Q-matrix that is upper triangular. Hence, due to Q's upper triangular structure, the solving of Equation (2.10) can be computed efficiently by forward substitution. This implicit Euler method has been shown to be greatly more efficient than other numerical solutions such as `Expokit` for many CTMC epidemic models and is a feasible method of probability mass calculation for larger state space systems [9].

2.3 Bayesian Inference

Within this thesis, we wish to infer the unknown parameters of an epidemic model from the observed case data combined with prior knowledge about disease epidemiology from past studies. Hence, we use Bayesian inference. *Bayesian inference* is a statistical framework derived from Bayes' Theorem where a prior distribution and data likelihood are combined to evaluate the *posterior* distribution. Define the unknown parameters of a model to be $\theta \in \Theta$. Let $L(\theta) = p(D|\theta)$ denote the likelihood of obtaining data D from θ, $p(\theta)$ denote the prior distribution of θ and $p(\theta|D)$ denote the posterior distribution of θ given the data D. Bayes' theorem states

$$\begin{aligned} p(\theta|D) &= \frac{p(D|\theta)p(\theta)}{p(D)}, \\ &= \frac{L(\theta)p(\theta)}{\int L(\theta)p(\theta)d\theta}. \end{aligned}$$

For models of any reasonable complexity, the *normalising constant* $p(D) = \int L(\theta)p(\theta)d\theta$ is infeasible to calculate [28]. Hence, we rely on inference methods that use the proportionality of Bayes' rule,

$$p(\theta|D) \propto L(\theta)p(\theta), \tag{2.11}$$

to find an estimate of the posterior distribution that is used for analysis. An estimate of the posterior distribution can be taken by sampling through *Markov Chain Monte Carlo* (MCMC) methodology [28, 44]. MCMC is a class of algorithms that allow approximate sampling from any desired distribution, by implementing a Markov chain with stationary distribution equal to the desired distribution. The *Metropolis-Hastings* algorithm is a commonly-used method within MCMC methodology to draw samples from a posterior distribution.

2.3.1 Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm is able to draw samples from a given target distribution, π, by creating a DTMC (referred to as the *MH chain*) with a stationary distribution equal to π [33, 53]. Once the MH chain has sufficiently converged to its stationary distribution all further states visited by the chain are samples from π.

Let $\{X_n\}_{n\in\mathbb{N}}$ denote the MH chain, π denote the target distribution and $q(.|.)$ denote the *proposal distribution.* The Metropolis-Hastings algorithm is given in Algorithm 2.

```
Data: Initial state $X_0$, number of iterations $K$, proposal distribution
      $q$, burn-in $m$.
Set $n = 0$;
while $n \leq K$ do
    Sample candidate point $Y \sim q(Y|X_n)$;
    Calculate acceptance probability
        $$\alpha(X_n, Y) = \min\left\{1, \frac{\pi(Y)q(X_n|Y)}{\pi(X_n)q(Y|X_n)}\right\};$$
    Sample $U_1 \sim \text{Uniform}(0, 1)$;
    if $U_1 < \alpha(X_n, Y)$ then
        $X_{n+1} = Y$ ;
    else
        $X_{n+1} = X_n$ ;
    end
    Set $n = n + 1$;
end
Discard burn-in time $\{X_0, ..., X_{m-1}\}$ ;
Result: Samples of $\pi$
```

Algorithm 2: Metropolis-Hastings algorithm.

Given the state of the MH chain is X_n, then the next state X_{n+1} is chosen by assessing a *candidate point* Y sampled from the *proposal distribution*, $q(Y|X_n)$. (Here q is dependent on the current state X_n, this can be forgone for an independent sampler.) The candidate point Y is then accepted as the next state of the MH chain with probability,

$$\alpha(X_n, Y) = \min\left\{1, \frac{\pi(Y)q(X_n|Y)}{\pi(X_n)q(Y|X_n)}\right\}. \tag{2.12}$$

Otherwise $X_{n+1} = X_n$. Under the regularity conditions of a DTMC (see Theorem 1.8.3 Norris [58]), the limiting distribution of the MH chain will converge to the unique stationary distribution, regardless of initial state X_0. This is true given the following conditions on q [44]. There exists a positive probability of: not accepting a proposed move,

$$P(\alpha(X_n, Y) < 1|X_n) > 0,$$

and of proposing a move to any state,

$$q(Y|X_n) > 0, \ X_n, Y \in \mathcal{S}.$$

Given $\{X_n\}_{n\in\mathbb{N}}$ has run for n steps, large enough to conclude it has converged to the stationary distribution (known as the *burn-in* time), then it is possible to evaluate the target distribution by the *ergodic average*. Take a burn-in time to be m steps of a MH chain run for t steps, then the ergodic average can be calculated as

$$E[f(X)] = \frac{1}{t-m}\sum_{i=m+1}^{t} f(X_i).$$

For Bayesian inference, let $\{\theta_n\}_{n\in\mathbb{N}}$ denote the MH chain of the process targeting the posterior distribution $p(\theta|D)$. Using the proportionality of Bayes' Theorem (2.11) allows the acceptance ratio of candidate point θ' to be expressed as such,

$$\begin{aligned}\alpha(\theta_n, \theta') &= \min\left\{1, \frac{p(\theta'|D)q(\theta_n|\theta')}{p(\theta_n|D)q(\theta'|\theta_n)}\right\},\\ &= \min\left\{1, \frac{L(\theta')p(\theta')q(\theta_n|\theta')}{L(\theta_n)p(\theta_n)q(\theta'|\theta_n)}\right\}. \end{aligned} \tag{2.13}$$

Hence, to alter the Metropolis-Hastings algorithm for Bayesian inference of parameters, the acceptance probability used in Algorithm 2 is replaced with Equation (2.13).

For the optimal implementation of the Metropolis-Hastings algorithm some design decisions must be made. Choices of proposal distribution, burn-in time and initial state can effect the performance of the algorithm. Performance is measured by the *mixing*, how well the MH chain explores the support of the posterior distribution, and the *convergence time*, the number of iterations required to ensure convergence to the stationary distribution [28]. The ideal proposal distribution is the target distribution but since the target distribution is unknown, one often aims for close normal approximations. A good proposal choice results in quick convergence from any initial state and adequate exploration of the target distribution, leading to a shorter burn-in time and a representative sample. A poor proposal or initial state could lead to slow convergence times or the process becoming "stuck" on local maxima. A method of judging the effectiveness of a proposal distribution is to study the *acceptance ratio*, the number of candidate points that have been accepted as a proportion of the total number of proposed moves; the optimal acceptance ratio is approximately 0.234 [7]. The MH chain trace plots should also be inspected to assess mixing. Randomly chosen initial states over a number of MH chains can be used to measure the convergence of a MH chain and possibly detect multi-modal posterior distributions. Hence, a typical strategy is to run a smaller number of chains with randomly chosen initial values and run each chain until the estimator values are comparable, indicating convergence of each chain to a reasonable degree [14, 28].

Consider the Metropolis-Hastings algorithm has produced posterior samples $\{\theta_1, ..., \theta_N\}$. The choice of summary statistics/intervals and the number of samples required deserve consideration. The following discussions within can be found in more detail within Carlin and Louis [14]. Here, the *posterior sample mean*, $\bar{\theta}$, is used as the point estimate of the posterior distribution. Although, the posterior samples are not technically IID, as are correlated by the MH chain. Hence, it is possible to underestimate the variance of the sample mean, and therefore the number of samples required. This can be taken into account by the *effective sample size* (ESS), the corresponding sample size of IID samples needed to have the same variance as the posterior samples [14, 66]. The ESS is defined as

$$\text{ESS} = \frac{N}{1 + 2\sum_{k=1}^{\infty} \rho_k(\theta)},$$

where $\rho_k(\theta)$ is the autocorrelation at lag k of $\{\theta_1, ..., \theta_N\}$. ESS can be used to estimate the variance of the posterior sample mean,

$$\widehat{\text{Var}}(\bar{\theta}) = \frac{s^2}{\text{ESS}},$$

where s^2 is the sample variance of $\{\theta_1, ..., \theta_N\}$. This variance can be used to calculate the $100(1-\alpha)\%$ confidence interval of the posterior sample mean,

$$\bar{\theta} \pm z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\bar{\theta})}.$$

Here, we determine the required number of samples has been reached when reaching a benchmark minimum $\widehat{\mathrm{Var}}(\bar{\theta})$.

Alternatively, posterior distributions can be summarised by *Bayesian credible regions*. Credible regions are used to convey the percentiles of the posterior distribution [14, 44]. Define an $100(1-\alpha)\%$ *credible region* as a subset $C \in \Theta$, such that ,

$$P(\theta \in C|D) \geq 1 - \alpha. \tag{2.14}$$

That is, the *"the probability that θ lies in C, given data D, is at least* $(1-\alpha)$*"* [14]. A $100(1-\alpha)\%$ credible region can be calculated by the *highest posterior density* (HPD) interval such that

$$C = \{\theta \in \Theta \mid p(\theta|D) \geq k(\alpha)\}$$

where $k(\alpha)$ is chosen so Equation (2.14) is satisfied. The HPD interval can summarise the posterior range and confidence in the plausible values of θ.

2.3.2 Likelihood

The Metropolis-Hastings algorithm requires the likelihood, $L(\theta)$, and prior density, $p(\theta)$, to be evaluated for each candidate point θ. The prior distribution is defined by previous knowledge about parameters and usually expressed as a parametric function. The likelihood can often prove more difficult to evaluate, as discussed in this section.

The likelihood of a set of observed data from an underlying Markov process can be expressed by a *Hidden Markov Model* (HMM).

Definition 2.3.1. A HMM is a set of Markov random variables $\{(X(t), Y(t))\}_{t\geq 0}$, where $X(t)$ denotes the hidden underlying Markov process and $Y(t)$ denotes the observation process dependent on $X(t)$, such that

$$P(\mathbf{Y}_{1:t} = \mathbf{y}_{1:t}|\mathbf{X}_{1:t} = \mathbf{x}_{1:t}) = \prod_{i=1}^{t} P(Y_i = y_i|X_i = X_i),$$

where at discrete time points $\{1, ..., t\}$, $\mathbf{X}_{1:t} = \{X(1), ..., X(t)\}$, $\mathbf{x}_{1:t} = \{x_1, ..., x_t\}$, $\mathbf{Y}_{1:t} = \{Y(1), ..., Y(t)\}$, $\mathbf{y}_{1:t} = \{y_1, ..., y_t\}$.

Let $\{X(t)\}_{t\geq 0}$ denote the underlying Markov process with parameters θ. Let $P(x_0)$ denote the initial distribution of the Markov process at time $t = 0$ and $\mathbf{y}_{1:T} = \{y_1, ..., y_T\}$ denote the series of observations that comprise the observed dataset D. The likelihood $L(\theta)$ is given by

$$L(\theta) = P(\mathbf{y}_{1:T}) = P(y_1) \prod_{t=2}^{T} P(y_t|\mathbf{y}_{1:t-1}), \tag{2.15}$$

by the law of total probability, where

$$P(y_1) = P(y_1|x_0)P(x_0),$$

and

$$P(y_t|\mathbf{y}_{1:t-1}) = \sum_{x_t \in S} P(y_t|x_t)P(x_t|\mathbf{y}_{1:t-1}).$$

Hence, we can calculate the likelihood by breaking it down into sequential calculations of: the probability of an observation given the state of the Markov process, $P(y_t|x_t)$, and the probability of being in that state of the Markov process conditioned on all previous information, $P(x_t|\mathbf{y}_{1:t-1})$. Take time $t = 0$,

$$\begin{aligned} P(x_1|y_1, x_0) &= \frac{P(y_1|x_1, x_0)P(x_1|x_0)}{P(y_1|x_0)} \\ &= \frac{P(y_1|x_1)P(x_1|x_0)}{\sum_{x_1 \in S} P(y_1|x_1)P(x_1|x_0)}, \end{aligned} \tag{2.16}$$

where $P(y_1|x_1)$ can be calculated from the observation process and $P(x_1|x_0)$ is found by the probability mass function of $\{X(t)\}_{t\geq 0}$ with initial distribution $P(x_0)$. For any time $t \geq 1$;

$$\begin{aligned} P(x_t|\mathbf{y}_{1:t}) &= \frac{P(y_t|x_t, \mathbf{y}_{1:t-1})P(x_t|\mathbf{y}_{1:t-1})}{P(y_t|\mathbf{y_{t-1}})} \\ &= \frac{P(y_t|x_t)P(x_t|\mathbf{y}_{1:t-1})}{\sum_{x_t \in S} P(y_t|x_t)P(x_t|\mathbf{y}_{1:t-1})}, \end{aligned} \tag{2.17}$$

where $P(y_t|x_t)$ can be calculated from the observation process and $P(x_t|\mathbf{y}_{1:t-1})$ is found by the probability mass function of $\{X(t)\}_{t\geq 0}$ with initial distribution $P(x_{t-1}|\mathbf{y}_{1:t-1})$.

For practical purposes, the observation process is usually a trivial or easily-evaluated distribution, but often the probability mass function of $\{X(t)\}_{t\geq 0}$

is far more difficult to compute. As discussed in Section 2.2.2, this thesis is focused on stochastic epidemic models that can be expressed in DoA representation. The DoA representation allows for efficient calculation of the probability mass function of the CTMC via the implicit Euler method. Hence, the implicit Euler method will be the standard for "exact" likelihood calculations within this thesis. Although like `Expokit`, the efficiency of the implicit Euler method decreases with the size of the state space. As the efficiency of the Metropolis-Hastings algorithm is dictated by the computational efficiency of the likelihood, the use of the implicit Euler solution is restrictive for some practical purposes. In such cases alternative methods for Bayesian inference are implemented that utilize simulation-based approximations.

Several likelihood-free, simulation-based alternatives for Bayesian inference are commonly used within the literature. They predominately originate from approximate Bayesian computation [50, 75], particle MCMC and sequential Monte Carlo methodologies [22]. We will focus on the latter techniques.

2.3.3 Particle Marginal Metropolis-Hastings

Particle MCMC refers to a class of MCMC algorithms where exact likelihood calculations are replaced by simulation-based estimates. Here, we focus on the *Particle Marginal Metropolis-Hastings* (PMMH) algorithm, which uses a *sequential Monte Carlo* (SMC) estimate of the likelihood within the Metropolis-Hastings algorithm [6]. For an explanation of the SMC estimate, see Section 2.3.4. The PMMH algorithm is shown in Algorithm 3.

PMMH avoids the main problem with the Metropolis-Hastings algorithm of computationally expensive (or infeasible) likelihood calculation, and instead uses a simulation-based estimate of the likelihood. This comes at the trade-off of using a number of simulations within the SMC algorithm to calculate a reliable likelihood estimate. As given in Section 2.3.4, the SMC likelihood estimate (Theorem 2.3.3) is unbiased, but the use of an estimate of the likelihood does have some limitations. This means introducing an amount of error into our posterior distribution, and the variability of the estimate can have a negative effect on the mixing of the MH chain [6, 30]. The estimator variance can be reduced by increasing the number of simulations, but at greater computational cost. Although, as PMMH is an "exact approximation" to the Metropolis-Hastings algorithm [6], PMMH is a "gold-standard" method for simulation-based parameter inference. The following section introduces the SMC likelihood estimate and discusses its effect on the PMMH algorithm.

```
Data: Data $\mathbf{y}_{1:T}$, initial state $\theta_0$, number of iterations $K$, proposal
      distribution $q$, burn-in $m$.
Set $n = 0$;
Run SMC scheme targeting $p(\mathbf{x}_{0:T}|\mathbf{y}_{0:T})$ with $\theta_0$ and extract marginal
likelihood estimate $\hat{L}(\theta_0) = \hat{P}(\mathbf{y}_{1:T})$;
while $n \leq K$ do
    Sample candidate point $\theta' \sim q(\theta'|\theta_n)$;
    Run SMC scheme targeting $p(\mathbf{x}_{0:T}|\mathbf{y}_{0:T})$ with $\theta'$ and withdraw
    extract likelihood estimate $\hat{L}(\theta') = \hat{P}(\mathbf{y}_{1:T})$;
    Calculate acceptance probability

        $$\alpha(\theta_n, \theta') = \min\left\{1, \frac{\hat{L}(\theta')p(\theta')q(\theta_n|\theta')}{\hat{L}(\theta_n)p(\theta_n)q(\theta'|\theta_n)}\right\};$$

    Sample $U_1 \sim \text{Uniform}(0, 1)$;
    if $U_1 < \alpha(\theta_n, \theta')$ then
        $\theta_{n+1} = \theta'$ ;
    else
        $\theta_{n+1} = \theta_n$ ;
    end
    Set $n = n + 1$;
end
Discard burn-in time $\{\theta_0, ..., \theta_{m-1}\}$;
Result: Samples of $p(\theta|\mathbf{y}_{1:T})$
```

Algorithm 3: PMMH algorithm for Bayesian inference.

2.3.4 Sequential Monte Carlo

Define a *particle* to be a possible state (or sequence of possible states) of a Markov process at time t. *Sequential Monte Carlo* (SMC) (or *particle filters*) refers to a class of algorithms built upon sequentially filtering a set of particles through conditioning on observations of the dataset. There are two schools of SMC algorithms: *state space inference* and *state space/parameter inference* [22]. Here, we will focus solely on a particular state space inference SMC algorithm due to its ability to produce unbiased likelihood estimates for use in PMMH. The "bootstrap filter" developed by Gordon *et al.* [31], is an intuitive SMC algorithm able to estimate the distribution of a Markov process and calculate a likelihood estimate from this approximate distribution. The bootstrap algorithm and likelihood estimate is given in Algorithm 4.

The bootstrap algorithm works as follows. Define a particle as a sequence of possible states of the Markov process conditioned on the data up to some time t. That is, a particle $\mathbf{x}_{0:t}$ is random sample drawn from $p(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})$. Assume a population of N_p independent particles, $\{\mathbf{x}_{0:t}^{(i)}|\ \forall\ i = 1, ..., N_p\}$, are distributed according to $p(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})$,

$$\{\mathbf{x}_{0:t}^{(1)}, ..., \mathbf{x}_{0:t}^{(N_p)}\} \sim p(\mathbf{x}_{0:t}|\mathbf{y}_{1:t}).$$

By *Prediction* and *Update* steps, it is possible to generate a set of particles distributed according to the state of the Markov process conditioned on each observation up to time $t+1$, $p(\mathbf{x}_{0:t+1}|\mathbf{y}_{1:t+1})$,

$$\{\mathbf{x}_{0:t+1}^{(1)}, ..., \mathbf{x}_{0:t+1}^{(N_p)}\} \sim p(\mathbf{x}_{0:t+1}|\mathbf{y}_{1:t+1}),$$

and calculate a likelihood estimate $\hat{P}(y_{t+1}|\mathbf{y}_{1:t})$. The two steps work as follows.

1. Prediction:
   A simulation from each particle "predicts" a possible state of the Markov process conditioned on all previous observations. Each particle $\mathbf{x}_{0:t}^{(i)}$ is used as a initial state in simulation over time $[t, t+1]$ to obtain $\mathbf{x}_{0:t+1}^{*(i)}$, which is a sample of $p(\mathbf{x}_{0:t+1}|\mathbf{y}_{1:t})$ by the Markov property. The resulting set of particles is distributed as

$$\{\mathbf{x}_{0:t+1}^{*(1)}, ..., \mathbf{x}_{0:t+1}^{*(N_p)}\} \sim p(\mathbf{x}_{0:t+1}|\mathbf{y}_{1:t}).$$

```
Data: Data $\mathbf{y}_{1:T}$, initial distribution $p(x_0)$, number of particles $N_p$.
Set $t = 0$;
Sample $\{x_0^{(1)}, ..., x_0^{(N_P)}\} \sim p(x_0)$ ;
for $i = 1, ..., N_p$ do
    Simulate $\mathbf{x}_{0:1}^{*(i)} \sim p(x_1|x_0^{(i)})$ ;
    Assign weight $W_1^{(i)} = P(y_1|x_1^{*(i)})$ ;
end
Calculate likelihood
    $$\hat{P}(y_1) = \frac{1}{N_p}\sum_{i=1}^{N_p} W_1^{(i)};$$
Re-sample $\{\mathbf{x}_{0:1}^{(1)}, ..., \mathbf{x}_{0:1}^{(N_P)}\}$ with weights
    $$P(\mathbf{x}_{0:1} = \mathbf{x}_{0:1}^{*(j)}) = \frac{W_1^j}{\sum_k W_1^k};$$
for  $t = 2, ..., T$ do
    for $i = 1, ..., N_p$ do
        Simulate $\mathbf{x}_{0:t}^{*(i)} \sim p(x_t|x_{t-1}^{(i)})$ ;
        Assign weight $W_t^{(i)} = P(y_t|x_t^{*(i)})$ ;
    end
    Calculate likelihood
        $$\hat{P}(y_t|\mathbf{y}_{1:t-1}) = \frac{1}{N_p}\sum_{i=1}^{N_p} W_t^{(i)};$$
    Re-sample $\{\mathbf{x}_{0:t}^{(1)}, ..., \mathbf{x}_{0:t}^{(N_P)}\}$ with weights
        $$P(\mathbf{x}_{0:t+1} = \mathbf{x}_{0:t+1}^{*(j)}) = \frac{W_{t+1}^j}{\sum_k W_{t+1}^k};$$
end
Marginal likelihood given by
    $$\hat{P}(\mathbf{y}_{1:T}) = \hat{P}(y_1)\prod_{t=2}^{T}\hat{P}(y_t|\mathbf{y}_{1:t-1})$$
Result: Marginal likelihood estimate $\hat{P}(\mathbf{y}_{1:T})$, Markov process state space distribution $\{\mathbf{x}_{0:T}^{(1)}, ..., \mathbf{x}_{0:T}^{(N_p)}\}$.
```

Algorithm 4: Bootstrap algorithm.

Assign particle weights $W_t^i = P(y_{t+1}|\mathbf{x}_{0:t+1}^{*(i)})$ and estimate the likelihood as

$$\hat{P}(y_{t+1}|\mathbf{y}_{1:t}) = \frac{1}{N_p}\sum_{i=1}^{N_p} P(y_{t+1}|x_{t+1}^{*(i)}).$$

2. Update:
A re-sampling scheme based on the probability of observing the new datum then updates the set of particles to conditioned on y_{t+1}. Re-sampling $\{\mathbf{x}_{0:t+1}^{*(1)}, ..., \mathbf{x}_{0:t+1}^{*(N_p)}\}$ with weighted probability

$$P(\mathbf{x}_{0:t+1} = \mathbf{x}_{0:t+1}^{*(j)}) = \frac{W_{t+1}^j}{\sum_k W_{t+1}^k} \tag{2.18}$$

returns the resulting set of particles distributed as

$$\{\mathbf{x}_{0:t+1}^{(1)}, ..., \mathbf{x}_{0:t+1}^{(N_p)}\} \sim p(\mathbf{x}_{0:t+1}|\mathbf{y}_{1:t+1}).$$

The re-sampling scheme is a result from Smith and Gelfrand [72] which states the following:

Theorem 2.3.2. *(Smith and Gelfand [72]) Given N independent samples of distribution $G(x)$, $\{x^{*(i)}|\forall i = 1, ..., N\}$, and known function $L(x)$, then re-sampling with probability weight*

$$P(x = x^{*(j)}) = \frac{L(x^{*(j)})}{\sum_k L(x^{*(k)})}$$

returns samples $\{x^{(i)}|\forall i = 1, ..., N\}$ which converge in distribution to $L(x)G(x)$ as $N \to \infty$.

Re-sampling by Equation (2.18) will converge in distribution (by Theorem 2.3.2) as

$$\begin{aligned}\{\mathbf{x}_{0:t+1}^{(1)}, ..., \mathbf{x}_{0:t+1}^{(N_p)}\} &\sim p(\mathbf{x}_{0:t+1}|\mathbf{y}_{0:t})p(y_{t+1}|x_{t+1})\\ &= p(\mathbf{x}_{0:t+1}|\mathbf{y}_{0:t})p(y_{t+1}|\mathbf{x}_{0:t+1}, \mathbf{y}_{0:t})\\ &\propto p(\mathbf{x}_{0:t+1}|\mathbf{y}_{0:t}, y_{t+1})\\ &= p(\mathbf{x}_{0:t+1}|\mathbf{y}_{0:t+1}).\end{aligned}$$

The algorithm continues, updating t to $t+1$ and using the set of particles $\{\mathbf{x}_{0:t+1}^{(1)}, ..., \mathbf{x}_{0:t+1}^{(N_p)}\}$ for the *prediction* step.

As stated in Andrieu *et al.* [6], the above process can generate an unbiased estimate of the marginal likelihood within the bootstrap algorithm by the following theorem.

Theorem 2.3.3. *(Andrieu et al. [6]: Proved by Del Moral [21], and Pitt et al. [62]) Within the bootstrap algorithm using a population of N_p particles, an unbiased estimate of the likelihood is given by,*

$$\hat{P}(\mathbf{y}_{1:T}) = \hat{P}(y_1) \prod_{t=2}^{T} \hat{P}(y_t|\mathbf{y}_{1:t-1}) \tag{2.19}$$

where

$$\begin{aligned} \hat{P}(y_t|\mathbf{y}_{1:t-1}) &= \frac{1}{N_p} \sum_{i=1}^{N_p} W_t^{(i)} \\ &= \frac{1}{N_p} \sum_{i=1}^{N_p} P(y_t|x_t^{*(i)}), \end{aligned} \tag{2.20}$$

and

$$\hat{P}(y_1) = \frac{1}{N_p} \sum_{i=1}^{N_p} P(y_1|x_1^{*(i)}). \tag{2.21}$$

The effectiveness of PMMH is directly related to effectiveness of the SMC algorithm in producing a likelihood estimate with efficiency and accuracy. Unfortunately, for the bootstrap algorithm the variability of the likelihood estimate is directly related to the number of particles, and hence the computational cost of simulations used. Increasing the number of particles will reduce the variability of the estimate at the trade-off of a slower likelihood calculation. Often, a large number of particles are required as a highly variable estimate will have a negative effect on the mixing of the PMMH chain. The variability of the estimate is roughly constant as the number of observations increases if the number of particles increases linearly [30], but noisy datasets where datums have low probability of observation will require a high number of particles to obtain a reasonable estimate. These *rare events* can cause *particle degeneracy*, where a low number of particles are consistent with the data and so the bulk of the particles are “filtered” out at the re-sampling stage. The replenished set of particles now originate from a small number of dominate particles, and so does not adequately represent the subsequent distribution of the Markov process.

A common method to alleviate this issue, by increasing the number of particles that are consistent with the data, is to use *importance sampling.* Importance sampling is a technique where realisations are generated from an *importance process* that guides particle simulations towards states with a higher probability of matching the data [6, 22, 44]. Particles weights are then discounted by the difference between the original model and the importance process. That is, given a set of particles

$$\{\mathbf{x}_{0:t}^{(1)}, ..., \mathbf{x}_{0:t}^{(N_p)}\} \sim p(\mathbf{x}_{0:t}|\mathbf{y}_{1:t}),$$

instead of simulating according to the model process (sampling from density $p(x_{t+1}|x_t)$), realisations are generated from the importance process (sampling from density $q(x_{t+1}|x_t)$). Resulting particles $\{\mathbf{x}_{0:t+1}^{*(1)}, ..., \mathbf{x}_{0:t+1}^{*(N_p)}\}$ are assigned weights that are the product of observation probability and the *importance factor*, the ratio of the model-to-importance probability mass of that particle's realisation, such that

$$W_{t+1}^{i} = P(y_{t+1}|x_{t+1}^{*(i)})\frac{p(x_{t+1}^{*(i)}|x_t^{*(i)}, \mathbf{y}_{1:t})}{q(x_{t+1}^{*(i)}|x_t^{*(i)})}. \tag{2.22}$$

Re-sampling with normalised weights as in Equation (2.18) results in a distribution of particles sampled from $p(\mathbf{x}_{0:t+1}|\mathbf{y}_{1:t+1})$ as required [6]. The importance sampling algorithm is identical to Algorithm 4, but where simulations occur according to the importance process $q(x_{t+1}|x_t)$ and re-sampling weights are calculated by Equation (2.22). The use of importance sampling can greatly increase the number of particles that are consistent with the data, but requires the computational cost of calculating the probability mass of each particles' sample path under both the model and importance processes.

The bootstrap algorithm is a special case of the importance sampling SMC algorithm where the importance process is equal to the model process; hence, the importance factor cancels out of Equation (2.22). The computational cost in calculating the importance factor exceeds the variance reduction benefits of importance sampling for the epidemic models considered here. Therefore, the bootstrap algorithm will be the standard SMC algorithm used within this thesis. In Chapter 3, we propose an alternative method to alleviate particle degeneracy issues, but it is not applicable to this study.

Chapter 3

Epidemic Model Inference

This chapter details the development of the SEIpIsIaR model for use in epidemiologically characterising the 1918 pandemic influenza strain (see Chapter 4). Inference methods are tested on simulated datasets to validate our approaches. A comparison of the Metropolis-Hastings and PMMH algorithms finds the PMMH performance is effectively identical in far less computation time. Inherent biases are discovered in the SEIpIsIaR model when conducting parameter inference from a single dataset; these issues are minimised by the use of parallel inference. We demonstrate that parallel inference allows for identification of key dynamics such as prior immunity and symptomatic/non-symptomatic transmission.

3.1 Epidemic Model Development

A stochastic epidemic model formulated as a CTMC will be fitted to the ship data. As discussed in Chapter 2, the discrete and random properties of a stochastic epidemic model replicate the highly variable nature of an influenza outbreak. These properties are particularly important for outbreaks with small populations sizes and low numbers of infected individuals as typical of the ship data. Although, the use of a CTMC model requires several assumptions; for the simulation studies within this chapter we take the following to be true, as motivated by the ship data. We have a homogeneously-mixing population of fixed size, N, where the population is *closed* (no births, deaths or migration). We assume all *symptomatic* cases of the disease are clinically observed upon immediate development of symptoms, and recorded at a daily resolution. See Chapter 4 for a discussion of the validity of these assumptions for the ship data. The task remains to develop the most appropriate model to reflect the influenza transmission dynamics within the ship-bound outbreaks. As with Chapter 2, we begin with the simplest case of the SIR model and build upon it.

Note, all models will be expressed in the *Degree-of-Advancement* (DoA) representation for comparison between the implicit Euler and SMC likelihood calculation methods. As such, let $\mathbf{y}_{1:T} = \{y_1, ..., y_T\}$ denote the dataset of *cumulative* counts of daily observed cases for any generic outbreak.

3.1.1 SIR Model

The SIR model is investigated for comparison to more complex and realistic models. To recap the example given in Chapter 2, in the SIR model individuals can be susceptible, infectious or recovered and can transition between compartments by infection and recovery events. A diagram of the possible compartments an individual can be in is given in Figure 3.1.

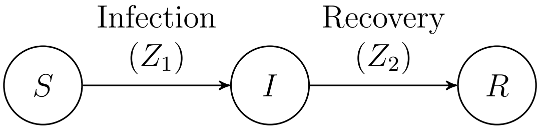


Figure 3.1: Compartment diagram of SIR system with DoA events.

Let $Z_1(t)$ and $Z_2(t)$ denote the number of *infection* and *recovery* events by time t, where the population compartment numbers can be retrieved by

$$\begin{aligned} S(t) &= N - Z_1(t), \\ I(t) &= Z_1(t) - Z_2(t), \\ R(t) &= Z_2(t). \end{aligned} \tag{3.1}$$

The SIR process is expressed as a two-variable CTMC $\{(Z_1(t), Z_2(t))\}_{t \geq 0}$ with state space,

$$\mathcal{S} = \{(Z_1, Z_2) | Z_1, Z_2 \in \mathbb{N}, 0 \leq Z_2 \leq Z_1 \leq N\}. \tag{3.2}$$

The transition rates from a state (Z_1, Z_2) are displayed in Table 3.1.

Transition Type	State Change	Transition Rate
Infection	$(Z_1, Z_2) \to (Z_1 + 1, Z_2)$	$\frac{\beta(N-Z_1)(Z_1-Z_2)}{N-1}$
Recovery	$(Z_1, Z_2) \to (Z_1, Z_2 + 1)$	$\gamma(Z_1 - Z_2)$

Table 3.1: SIR events and transition rates.

The use of this model requires several assumptions. A consequence of the CTMC formulation is that the length of the infectious period is exponentially distributed with mean $1/\gamma$. While this is not a clinically-motivated assumption, it is commonly allowed to facilitate easier mathematical analysis [4]. Another consequence of the SIR model is all individuals are immediately infectious upon contracting the disease, with all infectious individuals equally likely to transmit the disease. For the observation process, we assume the onset of symptoms coincides with an individual becoming infectious. As all symptomatic individuals are clinically observed, we therefore assume perfect observation of all infected individuals and define the observation process as

$$P(y_t|(Z_1, Z_2)) = \begin{cases} 1 & \text{if } y_t = Z_1, \\ 0 & \text{else.} \end{cases} \tag{3.3}$$

Other compartments and transitions are commonly added to the SIR model to better replicate the dynamics of disease transmission. For example, influenza requires time for the viral load to develop within an individual before they are infectious to others [11, 54, 77]. This time, known as the latency period, can be replicated by an additional compartment within an epidemic model known as the SEIR model.

3.1.2 SEIR Model

The SEIR model is an extension of the SIR model where a compartment is included to represent the *latency period*. A latency period is the period of time between an individual becoming infected with a disease and the individual becoming infectious to others. The infection status of each individual are now divided into compartments of: *Susceptible* (S) - able to be infected, *Exposed* (E) - infected but not yet infectious, *Infectious* (I) - infected and infectious, *Recovered* (R) - recovered and now immune to further infection. There are three possible transitions an individual can undergo: a susceptible individual becoming infected via contact with an infectious individual (*infection* event), the viral load developing within an exposed individual to where they are infectious to others (*become infectious* event) or an infectious individual recovering from the disease (*recovery* event) [41]. A diagram of the possible compartments an individual can be in is given in Figure 3.2.

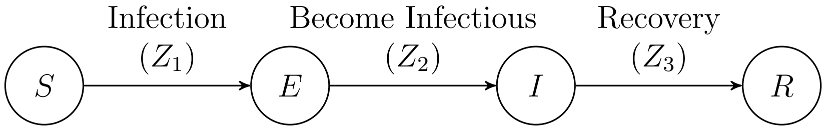


Figure 3.2: Compartment diagram of SEIR system with DoA events.

Let $Z_1(t)$, $Z_2(t)$ and $Z_3(t)$ denote the number of *infection*, *become infectious* and *recovery* events by time t, where the population numbers are given by

$$\begin{aligned} S(t) &= N - Z_1(t), \\ E(t) &= Z_1(t) - Z_2(t), \\ I(t) &= Z_2(t) - Z_3(t), \\ R(t) &= Z_3(t). \end{aligned} \tag{3.4}$$

The SEIR process is expressed as a three-variable CTMC $\{(Z_1(t), Z_2(t), Z_3(t))\}_{t\geq 0}$ with state space

$$\mathcal{S} = \{(Z_1, Z_2, Z_3) | Z_1, Z_2, Z_3 \in \mathbb{N}, 0 \leq Z_3 \leq Z_2 \leq Z_1 \leq N\}. \tag{3.5}$$

The transition rates from a state (Z_1, Z_2, Z_3) are displayed in Table 3.2.

Transition Type	State Change	Transition Rate
Infection	$(Z_1, Z_2, Z_3) \to (Z_1 + 1, Z_2, Z_3)$	$\frac{\beta(N-Z_1)(Z_2-Z_3)}{N-1}$
Become Infectious	$(Z_1, Z_2, Z_3) \to (Z_1, Z_2 + 1, Z_3)$	$\sigma(Z_1 - Z_2)$
Recovery	$(Z_1, Z_2, Z_3) \to (Z_1, Z_2, Z_3 + 1)$	$\gamma(Z_2 - Z_3)$

Table 3.2: SEIR events and transition rates.

The use of a CTMC model results in the duration of the latency period being exponentially distributed with mean $1/\sigma$. The end of the latency period signifies the onset of an individuals infectiousness. As with the SIR observation process, we assume the concurrent onset of observable symptoms and infectiousness and so the observation process is defined as

$$P(y_t | (Z_1, Z_2, Z_3)) = \begin{cases} 1 & \text{if } y_t = Z_2, \\ 0 & \text{else.} \end{cases} \tag{3.6}$$

The SIR and SEIR models assume all infectious individuals are identical in transmissibility and ability to be observed by a clinician. During the course of an influenza outbreak, not all infectious individuals are uniformly infective and not all will be observed by a clinician [48, 77]. This may be due to a number of reasons such as a reduced viral load resulting in mild or no symptoms or restricted availability to report illness. As such, a non-trivial proportion of infectious individuals are *asymptomatic*, displaying mild or no symptoms and will not be clinically observable, but are infectious to others. The symptomatic versus asymptomatic process can be included within an epidemic model via the SEIsIaR model, introduced in the next section.

3.1.3 SEIsIaR Model

The SEIsIaR model accounts for the proportion of cases that are symptomatic, $p_s \in [0, 1]$, by demarcating symptomatic and asymptomatic individuals in separate compartments, I_S and I_A respectively. When an individual's latency period expires and they leave the exposed class, they become symptomatic (with probability p_s) or are otherwise asymptomatic (with probability $1 - p_s$).

The introduction of symptomatic and asymptomatic infectious individuals raises the question of how to model severity of cases. For simplicity, we have assumed two levels of illness, those ill enough to be observed and those who are not, but how do these two case types behave with respect to disease transmission. That is, can we assume symptomatic and asymptomatic individuals have equal average length infectious periods and are equally likely to infect susceptible individuals? This is an open question with a severe lack of evidence available [60]. We could expect symptomatic individuals to transmit influenza at a higher rate due to increased symptom-based transmission from coughs and sneezes. Conversely, symptomatic individuals are more likely to attempt some form of self-isolation or increased hygiene given their awareness of their infection and consciously reduce their transmission rate. Hence, the difference in transmission is a complicated issue. Within the literature, this decision is open to debate, dependent on disease and simplifying model assumptions. There are a variety of approaches taken, for example: equal infectious periods [26], unique (or zero) asymptomatic infectiousness [48], unique infectious period and asymptomatic infectiousness [17] etc. Here, we allow the least restricting assumption of unique infectious periods and infectiousness.

A diagram of the possible compartments an individual can be in is given in Figure 3.3.

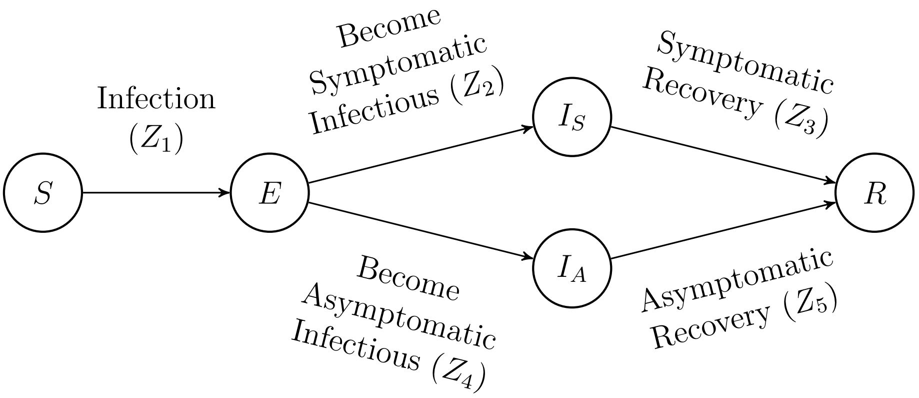


Figure 3.3: Compartment diagram of SEIsIaR system with DoA events.

Let $Z_1(t)$ be the number of *infection* events, $Z_2(t)$ the number of *become symptomatic infectious* events, $Z_3(t)$ the number of *symptomatic recovery* events, $Z_4(t)$ the number of *become asymptomatic infectious* events and $Z_5(t)$ the number of *asymptomatic recovery* events, by time t respectively. The population numbers are given by

$$\begin{aligned} S(t) &= N - Z_1(t), \\ E(t) &= Z_1(t) - Z_2(t) - Z_4(t), \\ I_S(t) &= Z_2(t) - Z_3(t), \\ I_A(t) &= Z_4(t) - Z_5(t), \\ R(t) &= Z_3(t) + Z_5(t). \end{aligned} \tag{3.7}$$

The SEIsIaR process is expressed as a five-variable CTMC $\{(Z_1(t), Z_2(t), Z_3(t), Z_4(t), Z_5(t))\}_{t\geq 0}$ with state space

$$\begin{aligned} \mathcal{S} =& \{(Z_1, Z_2, Z_3, Z_4, Z_5) | Z_1, Z_2, Z_3, Z_4, Z_5 \in \mathbb{N}, \\ & 0 \leq Z_3 + Z_5 \leq Z_2 + Z_4 \leq Z_1 \leq N, Z_4 \leq Z_2, Z_5 \leq Z_3\}. \end{aligned}$$

The transition rates from a state $(Z_1, Z_2, Z_3, Z_4, Z_5)$ are displayed in Table 3.3.

Transition Type	State Change	Transition Rate
Infection	$Z_1 \to Z_1 + 1$	$\frac{\beta_s(N-Z_1)(Z_2-Z_3)}{N-1} + \frac{\beta_a(N-Z_1)(Z_4-Z_5)}{N-1}$
Become Symptomatic Infectious	$Z_2 \to Z_2 + 1$	$p_s\sigma(Z_1 - Z_2 - Z_4)$
Symptomatic Recovery	$Z_3 \to Z_4 + 1$	$\gamma_s(Z_2 - Z_3)$
Become Asymptomatic Infectious	$Z_4 \to Z_4 + 1$	$(1 - p_s)\sigma(Z_1 - Z_2 - Z_4)$
Asymptomatic Recovery	$Z_5 \to Z_5 + 1$	$\gamma_a(Z_4 - Z_5)$

Table 3.3: SEIsIaR events and transition rates. Note, State Change lists the variable of the process that undergoes change by a transition; all other remain unchanged.

The symptomatic and asymptomatic infectious periods are exponentially distributed, with means $1/\gamma_s$ and $1/\gamma_a$ respectively. Symptomatic and asymptomatic individuals transmit according to unique transmission parameters, β_s and β_a. As with the SEIR model, we assume the concurrent onset of symptoms with infectiousness but only for symptomatic individuals. Therefore, we assume perfect observation of only symptomatic individuals and the observation process is defined as

$$P(y_t|(Z_1, Z_2, Z_3, Z_4, Z_5)) = \begin{cases} 1 & \text{if } y_t = Z_2, \\ 0 & \text{else.} \end{cases} \tag{3.8}$$

In the SEIsIaR model, we have accounted for two levels of case severity and infectiousness of individuals but the observation process still relies on the assumption that symptoms and infectiousness occur concurrently. As influenza infectiousness is related to the viral load within an individual, there exists a period of time where the viral load is high enough to allow disease transmission but not to induce symptoms [67, 69]. This time is known as the pre-symptomatic infectious period and can be replicated by an additional compartment within the SEIpIsIaR model, introduced in the next section.

3.1.4 SEIpIsIaR Model

The SEIpIsIaR model accounts for the *pre-symptomatic infectious period*, the time between an individual becoming infectious and demonstrating symptoms [69], within the compartment I_P. Here, we only allow a pre-symptomatic period for the symptomatic cases as asymptomatic cases will not develop observable symptoms.

Similar to the introduction of asymptomatic infections in the SEIsIaR model, introducing a pre-symptomatic period requires assumptions about the duration of this period and the relative infectiousness of a pre-symptomatic individual. The pre-symptomatic period is epidemiologically different in definition to the symptomatic-asymptomatic dynamics and hence deserves unique consideration. Clinical studies have shown the existence of pre-symptomatic infectious transmission for influenza in ferrets [67], but there is little evidence about its effect on epidemic/pandemic disease transmission [60]. Hence, there exists significant motivation to account for the period within the model to further understand its role in influenza epidemics.

A diagram of the possible compartments an individual can be in is given in Figure 3.4.

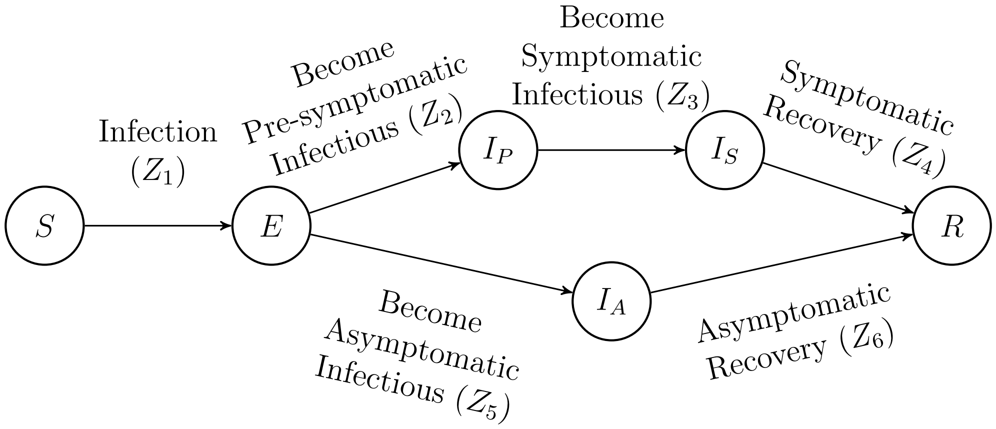


Figure 3.4: Compartment diagram of SEIpIsIaR system with DoA events.

Let $Z_1(t)$ be the number of *infection* events, $Z_2(t)$ the number of *become pre-symptomatic infectious* events, $Z_3(t)$ the number of *become symptomatic infectious* events, $Z_4(t)$ the number of *symptomatic recovery* events, $Z_5(t)$ the number of *become asymptomatic infectious* events and $Z_6(t)$ the number of *asymptomatic recovery* events, by time t respectively. The population numbers are given by

$$\begin{aligned} S(t) &= N - Z_1(t), \\ E(t) &= Z_1(t) - Z_2(t) - Z_5(t), \\ I_P(t) &= Z_2(t) - Z_3(t), \\ I_S(t) &= Z_3(t) - Z_4(t), \\ I_A(t) &= Z_5(t) - Z_6(t), \\ R(t) &= Z_4(t) + Z_6(t). \end{aligned} \tag{3.9}$$

The SEIpIsIaR process is expressed as a six-variable CTMC $\{(Z_1(t), Z_2(t), Z_3(t)), Z_4(t), Z_5(t), Z_6(t))\}_{t\geq 0}$ with state space

$$\begin{aligned}\mathcal{S} = &\{(Z_1, Z_2, Z_3, Z_4, Z_5, Z_6)|Z_1, Z_2, Z_3, Z_4, Z_5, Z_6 \in \mathbb{N},\\ &0 \leq Z_4 + Z_6 \leq Z_2 + Z_5 \leq Z_1 \leq N, Z_4 \leq Z_3 \leq Z_2, Z_6 \leq Z_5\}.\end{aligned}$$

The transition rates from a state $(Z_1, Z_2, Z_3, Z_4, Z_5, Z_6)$ are displayed in Table 3.4.

Transition Type	State Change	Transition Rate
Infection	$Z_1 \to Z_1 + 1$	$\dfrac{\beta_p(N-Z_1)(Z_2-Z_3)}{N-1} + \dfrac{\beta_s(N-Z_1)(Z_3-Z_4)}{N-1} + \dfrac{\beta_a(N-Z_1)(Z_5-Z_6)}{N-1}$
Become Pre-Symptomatic Infectious	$Z_2 \to Z_2 + 1$	$p_s\sigma(Z_1 - Z_2 - Z_5)$
Become Symptomatic Infectious	$Z_3 \to Z_3 + 1$	$\alpha(Z_2 - Z_3)$
Symptomatic Recovery	$Z_4 \to Z_4 + 1$	$\gamma_s(Z_3 - Z_4)$
Become Asymptomatic Infectious	$Z_5 \to Z_5 + 1$	$(1-p_s)\sigma(Z_1 - Z_2 - Z_5)$
Asymptomatic Recovery	$Z_6 \to Z_6 + 1$	$\gamma_a(Z_5 - Z_6)$

Table 3.4: SEIpIsIaR events and transition rates. Note, State Change lists the variable of the process that undergoes change by a transition; all other remain unchanged.

The pre-symptomatic period is exponentially distributed with mean $1/\alpha$. Pre-symptomatic individuals transmit the disease according to unique transmission parameter, β_p. The end of the pre-symptomatic period signifies the onset of an individual displaying symptoms. Hence, we assume the onset of symptoms upon entry to the symptomatic infectious compartment and so the observation process is defined as

$$P(y_t|(Z_1, Z_2, Z_3, Z_4, Z_5, Z_5, Z_6)) = \begin{cases} 1 & \text{if } y_t = Z_3, \\ 0 & \text{else.} \end{cases} \tag{3.10}$$

The SEIpIsIaR model will be the foundation of our influenza investigation but other factors within the model must be scrutinised. The initial distribution of the population amongst the model compartments requires consideration. This is related to two questions about the population prior to the outbreak: the *introduction of infection* and *prior immunity*. The introduction of infection refers to the number and infection status of the individuals who introduce a disease into the population. See Chapter 4 for the discussion of how this is accounted for within the ship data. For the purpose of this chapter, we assume one infected individual is introduced to the population at a known time.

Prior immunity refers to a proportion of the population that has developed an immunity to the disease before the onset of an outbreak [26, 49]. Hence, these individuals are mixing within the population but cannot contract or transmit the disease. Specifically for pandemic influenza, prior immunity can be high due to the development of a lasting immune response from previous attacks of influenza strains [77]. If a significant proportion of the population is immune, it can greatly reduce the number of cases caused by an epidemic and so should be included within the model. Here, we only consider lasting immunity over the course of the outbreak as the duration of the ship outbreaks are too short to consider waning immunity [49]. We can account for prior immunity by changing the initial number of individuals within the *recovered* compartment. If a single infected individual enters an entirely susceptible population at time 0, then the model's initial state is

$$(Z_1(0), Z_2(0), Z_3(0), Z_4(0), Z_5(0), Z_6(0)) = (1, 0, 0, 0, 0, 0),$$

which translates to compartment numbers

$$(S(0), E(0), I_P(0), I_S(0), I_A(0), R(0)) = (N-1, 1, 0, 0, 0, 0, 0).$$

To model for prior immunity within the remaining population we introduce the parameter, p_{sus}, the probability a random individual within the population is susceptible to infection. The number of immune individuals within the remaining susceptible population, n_{im}, is randomly sampled from a binomial$(N-1, 1-p_{sus})$ distribution. Hence, we assume all immune individuals are unobserved and the initial state is

$$(Z_1(0), Z_2(0), Z_3(0), Z_4(0), Z_5(0), Z_6(0)) = (1+n_{im}, 0, 0, 0, n_{im}, n_{im}),$$

which translates to compartment numbers

$$(S(0), E(0), I_P(0), I_S(0), I_A(0), R(0)) = (N-1-n_{im}, 1, 0, 0, 0, 0, n_{im}).$$

Note, within the SMC algorithm each particle is assigned an initial state with a randomly sampled $n_{im} \sim$ binomial$(N-1, 1-p_{sus})$.

The form of *contact-transmission structure*, frequency-dependent versus density-dependent also should be considered. Both contact structures are discussed in Chapter 2. Density-dependent can be more informative by the inclusion of the known area in which the outbreak takes place, but the choice is inconsequential unless considering outbreaks across populations and/or areas of distinct sizes. This requires consideration in Chapter 4 but is irrelevant here. Hence, for the following simulation studies frequency-dependent transmission is used.

3.2 Inference Method Validation

For the above models, we validate our inference methods by conducting experimental studies on simulated data. In this section, simulated datasets from each model will be used to compare the efficiency and accuracy of the Metropolis-Hastings and PMMH algorithms, as described in Chapter 2. To recap, the Metropolis-Hastings algorithm and PMMH algorithm are identical except for the method of likelihood calculation. The Metropolis-Hastings algorithm uses an "exact" likelihood; calculated by solving the master equation using the implicit Euler method with global precision τ. The PMMH algorithm uses an SMC likelihood estimate with number of particles N_p. Denote the respective likelihood methods NumInt(τ) and SMC(N_p). Hence, the comparison of these algorithms is a comparison of the behaviours of the likelihood calculation methods; compared by efficiency, accuracy and effect on the mixing of the MH chain.

Note, to replicate the ship data all simulated data will be generated with similar population sizes and parameter values consistent with previous studies of 1918 pandemic influenza [23, 54]. Only significant outbreaks greater than 10 cases will be considered and the data will be collected at daily resolution. As above, we assume a known initial state of the model.

3.2.1 SIR Model Parameter Inference

To be able to compare NumInt(τ) and SMC(N_p) both must be feasible methods of likelihood calculation for use with the SIR model. SMC(N_p) is straightforward to implement by using simulations of the SIR model CTMC given in Section 3.1.1 and the observation process given in Equation (3.3). As simulations are fast to generate by a stochastic simulation algorithm and the observation process is trivial to compute, SMC(N_p) is efficient for a large number of particles.

Although, it is possible for SMC(N_p) implemented in this manner to suffer from *likelihood failure*, defined as prematurely returning a zero likelihood estimate of a positive likelihood. Here, this occurs if every particle simulation is inconsistent with the observed data. That is, for the SIR model the SMC(N_p) algorithm will return a zero likelihood estimate if at any time $t = 1, ..., T$, every particle $\{(Z_1, Z_2)_t^{(i)} | i = 1, ..., N_p\}$ is such that the number of infections within the simulation does not equal the observed case data, $Z_1 \neq y_t$. Hence, the likelihood estimate

$$\hat{P}(y_t | \mathbf{y}_{1:t-1}) = \frac{1}{N_p} \sum_{i=1}^{N_p} P(y_t | (Z_1, Z_2)_t^{(i)}) = 0.$$

This may reflect a real error with the model or data collection that results in an observed data point with zero probability of occurring, but this is extremely unlikely. In most cases, we are able to determine that there exists a positive probability of an observed data point occurring, but the data is too inconsistent with model such that it has an extremely low probability of occurring. These *rare event* probabilities are difficult to estimate from Monte Carlo methods. The frequency of likelihood failure can be reduced by increasing the number of particles but this is not always practical. Typical rare-event simulation methods such as importance sampling are discussed in Chapter 2, but are not applicable here for computational reasons. Further methods of minimising likelihood failure are explored further in Section 3.3. If likelihood failure consistently occurs within the MH chain, it can have a negative effective on the mixing of the chain.

NumInt(τ) is far more computationally expensive method as it involves using forward substitution to repeatedly solve systems of linear equations. The number of linear equations is determined by the size of the CTMC's Q-matrix, an $|\mathcal{S}|$ by $|\mathcal{S}|$ matrix. Hence, the size of the state space can determine the efficiency or even viability of the algorithm. (Note, state space truncation is possible for increased efficiency but is not used here, see Chapter 5.) The state space size for the SIR model is given by

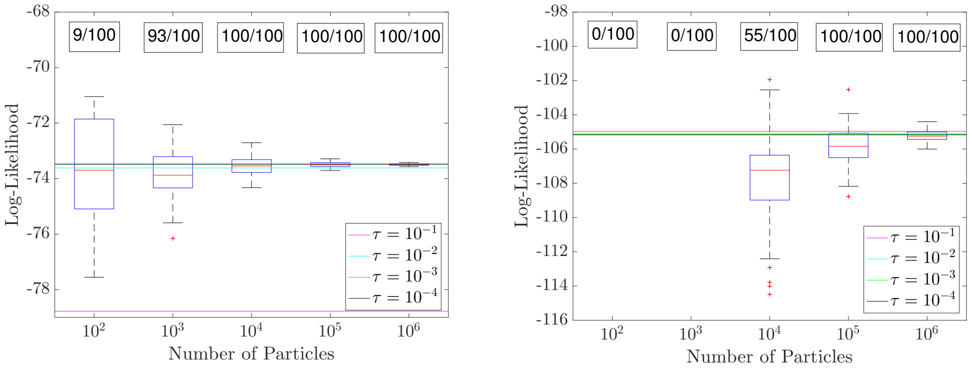
$$|\mathcal{S}| = \sum_{Z_1=0}^{N} \sum_{Z_2=0}^{Z_1} 1 = \frac{(N+1)(N+2)}{2}.$$

As the size of the state space is of order $\mathcal{O}(N^2)$, the population size will determine if NumInt(τ) is a viable method. Take for example, the runtime in calculating the NumInt(τ) likelihood for 10 sets of SIR model simulated data with increasing population sizes, N. The results are shown in Table 3.5.

Population Size N	Runtime (s)	Population Size N	Runtime (s)
100	0.168	600	6.988
200	0.736	700	9.901
300	1.553	800	13.317
400	2.886	900	19.629
500	4.705	1000	25.143

Table 3.5: Runtime of NumInt($\tau = 10^{-2}$) for SIR model with varying population size. Likelihoods calculated using true parameters. Datasets: observed the first 25 days of outbreak at daily resolution, $R_0 = 2$, $\frac{1}{\gamma} = 3$. Run in `Matlab` on an iMac (2013) with 2.7 GHz Intel Core i5 processor.

We observe an exponential growth in the runtime of NumInt(τ) as N increases. While the NumInt(τ) is a feasible solution for the SIR model, it is far more computational expensive for larger population sizes than the SMC(N_p) estimate. Obviously, the decision to use NumInt(τ) or SMC(N_p) is going to be dependent on the choices of τ and N_p. The precision τ must be high enough for acceptable error to the true likelihood, while the number of particles N_p must be large enough to reduce the variance of the likelihood estimate and the possibility of likelihood failure. Decreasing τ or increasing N_p comes at a computational cost, so a minimum required accuracy should be investigated. While this is problem specific, we have SIR model simulated data indicative of the ship data to validate the comparison. Here, we compare the NumInt(τ) and SMC(N_p) methods by varying τ and N_p. The algorithms are used to calculate likelihood of a SIR model simulated dataset with a set of "good-fit" ($R_0 = 2, 1/\gamma = 3$) and "poor-fit" ($R_0 = 3, 1/\gamma = 3$) parameters. The box plots of 100 SMC(N_p) estimates and the NumInt(τ) solution are shown in Figure 3.5 and comparative algorithm runtimes are given in Table 3.6.



(a) Good-fit parameters.

(b) Poor-fit parameters.

Figure 3.5: SIR model log-likelihood comparison for NumInt(τ) and SMC(N_p) by τ and N_p. NumInt(τ) solution by decreasing τ given in legend. Box plot of 100 SMC(N_p) calculations for each N_p with number of positive likelihood estimates out of 100 given in box. Dataset: $N = 1000$, observed the first 25 days of outbreak at daily resolution, $R_0 = 2$, $\frac{1}{\gamma} = 3$.

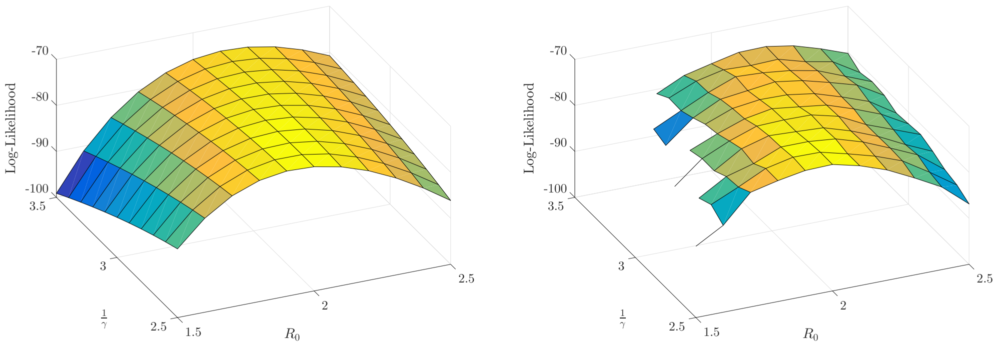
Precision τ	Runtime (s)
10^{-1}	2.512
10^{-2}	25.33
10^{-3}	213.9
10^{-4}	1858

Number of Particles N_p	Runtime (s)
10^2	0.007
10^3	0.064
10^4	0.652
10^5	6.655
10^6	72.54

Table 3.6: NumInt(τ) runtime by τ versus SMC(N_p) average runtime by N_p. Runtime taken from "good-fit" likelihood calculations given in Figure 3.5. Run in `Matlab` on an iMac (2013) with 2.7 GHz Intel Core i5 processor.

From Figure 3.5, the NumInt(τ) solution is poor and underestimates the likelihood for $\tau > 10^{-2}$ and Table 3.6 shows the runtime is inversely proportional to τ. Hence, a standard precision of $\tau = 10^{-2}$ will be used for best accuracy at acceptable runtime. While the SMC(N_p) is known to be unbiased for any N_p, Figure 3.5 demonstrates an increased probability of likelihood failure for $N_p < 10^4$. Within this chapter, the number of particles $N_p = 10^4$ will be standard within SMC(N_p) as it produces a desirable middle ground between runtime and variance reduction.

We can demonstrate the difference in likelihoods by comparing the *likelihood distribution shape*. That is, we calculate the likelihood at grid points over a subsection of the parameter space. Here, we cover a two-dimensional grid across $R_0 \in [1.5, 2.5]$ and $1/\gamma \in [2.5, 3.5]$ with 121 points at 0.1 intervals. The likelihood is calculated at each point with the NumInt($\tau = 10^{-2}$) and SMC($N_p = 10^4$) algorithms. The likelihood distributions are shown as three-dimensional surfaces in Figure 3.6. Note, missing points in the SMC($N_p = 10^4$) likelihood distribution are zero likelihood estimates. Here, we see SMC($N_p = 10^4$) suffers likelihood failure for the lower likelihood values, and the variability of estimator is obvious in the unevenness of the surface compared to NumInt($\tau = 10^{-2}$), although the overall likelihood shape is consistent between methods. Hence, both methods perform similarly in calculating the likelihood across the support of the likelihood distribution, except for SMC(N_p) difficulties in calculating low probability likelihoods.



(a) NumInt($\tau = 10^{-2}$)

(b) SMC($N_p = 10^4$)

Figure 3.6: SIR model log-likelihood comparison for NumInt(τ) and SMC(N_p) likelihood distribution shapes. Likelihood calculated in 121 even spaced grid points over $R_0 \in [1.5, 2.5]$ and $\frac{1}{\gamma} \in [2.5, 3.5]$. Missing points denote zero likelihood. Dataset: $N = 1000$, observed the first 25 days of outbreak at daily resolution, $R_0 = 2$, $\frac{1}{\gamma} = 3$.

A study on simulated data is conducted to compare the effectiveness of the Metropolis-Hastings and PMMH algorithms for the SIR model. For the SIR model, we conduct inference on the model parameters transformations

$$\theta = \left\{ R_0 = \frac{\beta}{\gamma}, \; \frac{1}{\gamma} \right\}.$$

We target these parameter transformations due to their physical definitions, which allows for natural interpretation and intuitive assignment of prior distributions. Both parameters are assigned uninformative uniform$(0.1, 8)$ prior distributions. The Metropolis-Hastings and PMMH algorithms are run on 100 SIR model simulated datasets for 10,000 iterations with 1,000 steps as burn-in time. Note, the iteration count is chosen to meet a specified estimator variance of the R_0 posterior sample mean, $\widehat{\text{Var}}(\bar{R}_0) < 0.01$. R_0 is chosen as the indicator parameter due to its physical interpretation. To compare the accuracy of the respective algorithms in identifying the posterior distribution, for each dataset the posterior sample mean bias, sample variance, posterior mean 95% CI and HPD interval widths are collated. These summary statistics are presented as they are indicative of the shape of the posterior distribution, as well as the accuracy of the point estimates. The comparative posterior results are shown in Figure 3.7.

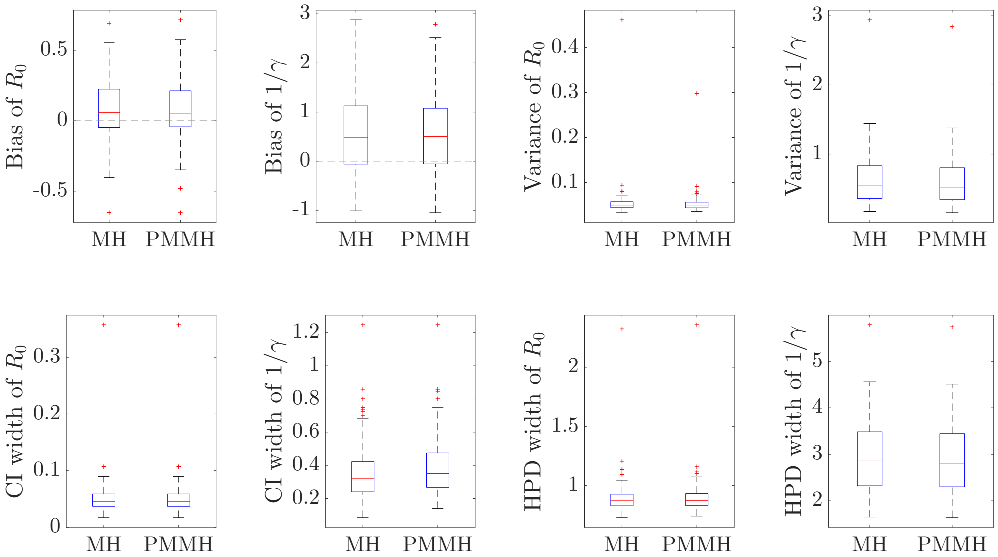


Figure 3.7: Box plot diagram comparing Metropolis-Hastings and PMMH posterior results for 100 SIR model simulated datasets. Posterior sample mean bias, sample variance, mean 95% CI interval width and HPD interval width from each run are displayed in box plots by parameter. Datasets: $N = 250$, observed the length of outbreak at daily resolution, $R_0 = 2$, $\frac{1}{\gamma} = 3$.

From Figure 3.7, it is clear to see that the Metropolis-Hastings and PMMH posterior results are near identical. The mean estimates of R_0 for both methods are similarly centred around the true values while both methods tend to slightly overestimate $1/\gamma$. Likewise, the sample variance, CI and HPD interval widths across parameters are similar for both methods. Hence, the Metropolis-Hastings and PMMH algorithms perform near identically in identifying the posterior distribution for the SIR model. From Table 3.6, the runtime of using the SMC(N_p) estimate instead of the NumInt(τ) solution (with standard precision) is roughly 40 times faster. Hence, the PMMH algorithm is preferential in parameter inference for the SIR model.

3.2.2 SEIR Model Parameter Inference

Likewise, to be able to compare NumInt(τ) and SMC(N_p) for use with the SEIR model, they must be feasible methods of likelihood calculation. The performance of the SMC(N_p) algorithm for the SEIR model is consistent with the SIR model except for a slight increase in variability and computation cost due to the increased dimension of the process. NumInt(τ) faces a far greater decrease in efficiency to the exponentially increased state space size. The state space size for the SEIR model is given by

$$|\mathcal{S}| = \sum_{Z_1=0}^{N} \sum_{Z_2=0}^{Z_1} \sum_{Z_3=0}^{Z_2} 1 = \frac{(N+1)(N+2)(N+3)}{6}.$$

As the size of the state space is of order $\mathcal{O}(N^3)$ the population size will severely restrict the cases where NumInt(τ) is a viable method. As with the SIR model, we take the runtime in calculating the likelihood for SEIR model simulated datasets with increasing population sizes. The runtime results of NumInt(τ) are shown in Table 3.7.

Population Size N	Runtime (s)
100	1.264
200	9.983
300	38.29
400	94.09
500	209.2
600	> 600

Table 3.7: Runtime of NumInt($\tau = 10^{-2}$) for SEIR model with varying population size. Likelihoods calculated using true parameters. Datasets: observed the first 25 days of outbreak at daily resolution, $R_0 = 2$, $\frac{1}{\sigma} = 1$, $\frac{1}{\gamma} = 3$. Run in `Matlab` on an iMac (2013) with 2.7 GHz Intel Core i5 processor.

Due to long computation time of NumInt(τ), it is not a practical method for the SEIR model with population sizes greater than 200. A comparison of the Metropolis-Hastings and PMMH algorithms for the SEIR model is recommended to validate the use of PMMH for more complex models. A study is conducted on simulated data with a reduced population size of $N = 100$. For the SEIR model, we conduct inference on the model parameters transformations

$$\theta = \left\{ R_0 = \frac{\beta}{\gamma}, \ \frac{1}{\sigma}, \ \frac{1}{\gamma} \right\}.$$

All parameters are assigned uniform$(0.1, 8)$ prior distributions. The Metropolis-Hastings and PMMH algorithms are run on 100 SEIR model simulated datasets for 10,000 iterations with 1,000 steps as burn-in. Note, the iteration count is chosen such that $\widehat{\text{Var}}(\bar{R}_0) \approx 0.01$. The comparative posterior results are shown in Figure 3.8.

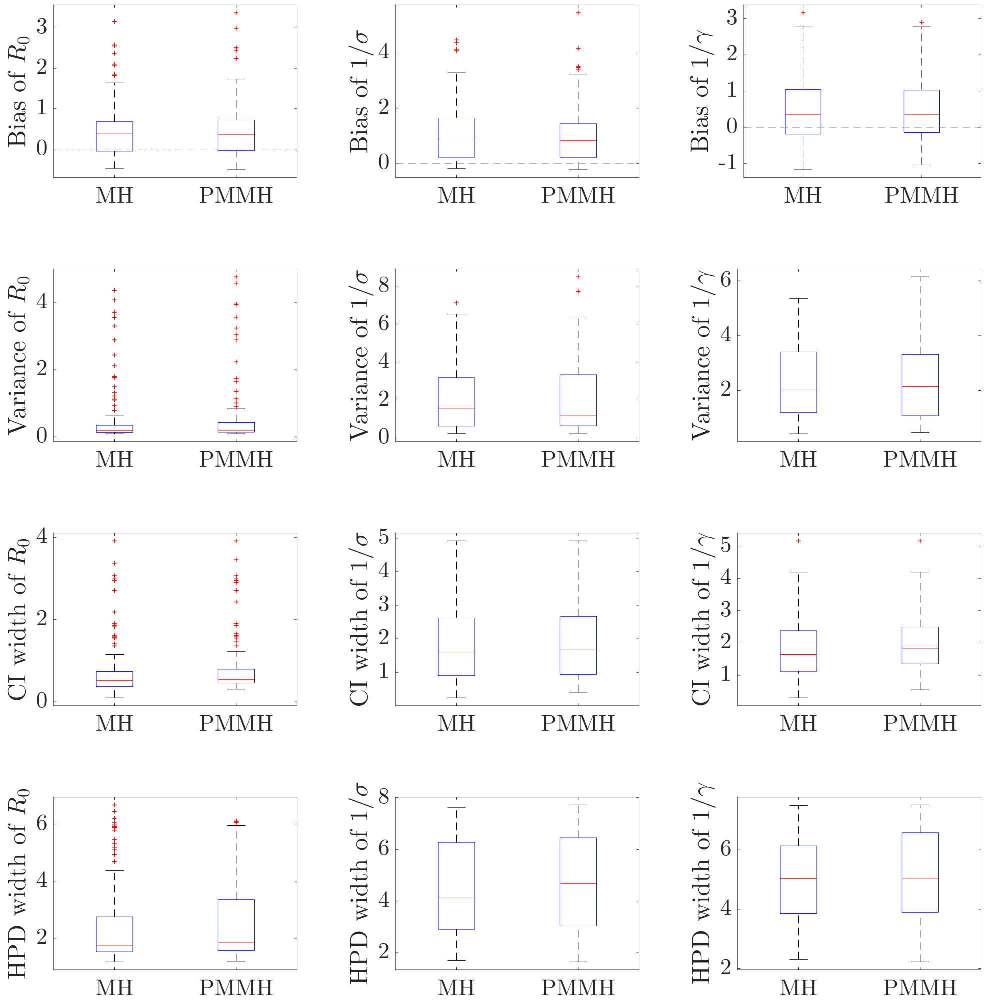


Figure 3.8: Box plot diagram comparing Metropolis-Hastings and PMMH posterior results for 100 SEIR model simulated datasets. Posterior sample mean bias, sample variance, mean 95% CI interval width and HPD interval width from each run are displayed in each box plot by parameter. Datasets: $N = 100$, observed the length of outbreak at daily resolution, $R_0 = 2$, $\frac{1}{\sigma} = 1$, $\frac{1}{\gamma} = 2$.

The Metropolis-Hastings and PMMH posterior distributions are near identical across R_0, $1/\sigma$ and $1/\gamma$. The sample mean distributions are similarly centred around the true values (with slight positive bias) and the sample variance, CI and HPD interval widths estimates are essentially equal. Hence, the Metropolis-Hastings and PMMH algorithms perform near identically in identifying the posterior for the SEIR model. Although as in Table 3.7, the runtime of the NumInt(τ) method is impractical for use within the Metropolis-Hastings algorithm for population sizes greater than 200. This restriction makes it an unsuitable method for the ship data and so the PMMH algorithm is preferential in identifying parameters for the SEIR model.

3.2.3 SEIsIaR Model Parameter Inference

The use of NumInt(τ) is infeasible for SEIsIaR likelihood calculations due to the increased size of the state space. The state space size for the SEIsIaR model is given by

$$\begin{aligned} |\mathcal{S}| &= \sum_{Z_1=0}^{N} \sum_{Z_2=0}^{Z_1} \sum_{Z_3=0}^{Z_2} \sum_{Z_4=0}^{Z_1-Z_2} \sum_{Z_5=0}^{Z_4} 1, \\ &\sim \mathcal{O}(N^5). \end{aligned} \tag{3.11}$$

As the size of the state space is of order $\mathcal{O}(N^5)$, the cases where NumInt(τ) is a viable solution is restricted to very small populations. Hence, NumInt(τ) is not a feasible method for use with the SEIsIaR model and is eliminated from further consideration. The performance of the SMC(N_p) algorithm for the SEIsIaR model is consistent with the SIR and SEIR model with a similar increased penalty for the increased dimensions of the process.

A study is conducted on simulated data to test the ability of the PMMH algorithm to identify the parameters of the SEIsIaR model. For the SEIsIaR model, we conduct inference on the model parameters transformations

$$\theta = \left\{ R_0 = \frac{p_s \beta_s}{\gamma_s} + \frac{(1-p_s)\beta_a}{\gamma_a},\ \frac{1}{\sigma},\ \frac{1}{\gamma_s},\ \frac{1}{\gamma_a},\ p_s,\ \kappa_S = \frac{(R_0)_S}{R_0} \right\}.$$

Denote $(R_0)_S$ and $(R_0)_A$ to be the R_0 contributions of symptomatic and asymptomatic individuals respectively, such that

$$\begin{aligned} (R_0)_S &= \frac{p_s \beta_s}{\gamma_s}, \\ (R_0)_A &= \frac{(1-p_s)\beta_a}{\gamma_a}. \end{aligned}$$

Hence, κ_S and $\kappa_A = (1 - \kappa_S)$ denote the *proportion of transmission* occurring from symptomatic and asymptomatic individuals. All parameters are assigned uniform$(0.1, 8)$ prior distributions, except p_s and κ_S which are assigned a uniform$(0.01, 0.99)$ prior distribution. The κ_S prior reflects our uncertainty in the respective proportion of transmission from symptomatic and asymptomatic individuals. The PMMH algorithm is run on 50 SEIsIaR model simulated datasets for 80,000 iterations with 5,000 steps as burn-in. Note, the iteration count is chosen such that $\widehat{\text{Var}}(\bar{R}_0) \approx 0.05$. Here we run two studies to highlight some key behaviours that result from the introduction of the symptomatic-asymptomatic dynamic. The first study focuses on data produced with low $R_0 = 1.2$, and the second with data produced with high $R_0 = 2$. The bias of posterior results for the low and high R_0 data are displayed in Figure 3.9.

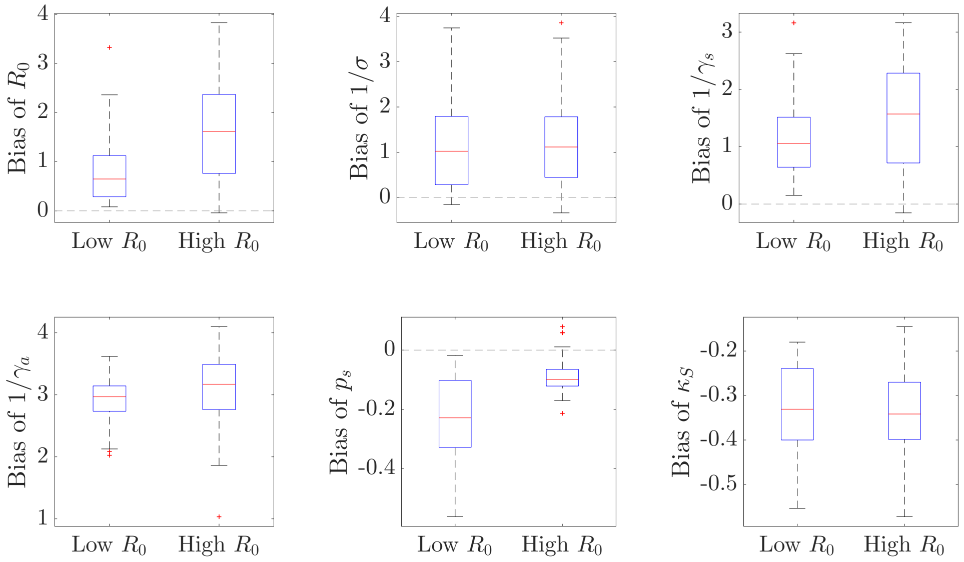
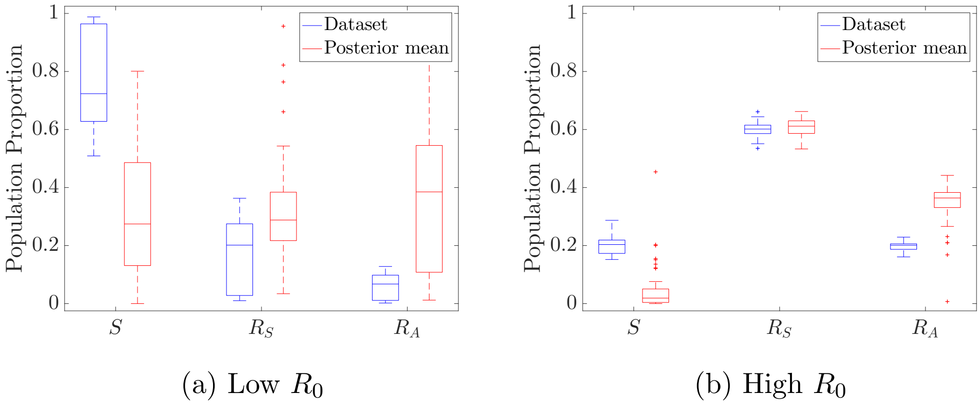


Figure 3.9: Box plot diagram of PMMH algorithm posterior for 50 SEIsIaR model simulated datasets generated with low $R_0 = 1.2$ values and high $R_0 = 2$ values. Posterior sample mean bias are displayed in each box plot by parameter. Datasets: $N = 1000$, observed the length of outbreak at daily resolution, $\frac{1}{\sigma} = 1$, $\frac{1}{\gamma_s} = 2$, $\frac{1}{\gamma_a} = 1$, $p_s = 0.75$, $\kappa_S = 0.91$.

In both studies, PMMH displays an ability to identify $1/\sigma$ and $1/\gamma_s$ consistent with the SEIR model. We note a clear inability to identify $1/\gamma_a$ due to the lack of information able to be extracted from symptomatic case data. We also observe a slight underestimation of κ_S and conversely overestimated κ_A. The behaviour of R_0 and p_s identifiability deserves further scrutiny.

For the low R_0 data, PMMH does an acceptable job at identifying R_0 and displays a degree of uncertainty about p_s with some negative bias. Note, only considering data of outbreaks > 10 cases could explain some of the R_0 positive bias [52]. Comparatively for the high R_0 data, R_0 is overestimated to a large degree and p_s shows significant negative bias. Hence, dependent on the size of the outbreak as dictated by R_0, the model inference is biased towards overestimating transmission and underestimating the proportion of symptomatic cases. The overestimated R_0 results in almost all of the population becoming infected and the underestimated p_s is fitted to the *clinical attack rate* (proportion of population observed to have been infected) rather than the true symptomatic - asymptomatic probability.

We can demonstrate this by studying posterior simulations compared to the dataset. Denote the *end-state* of a simulation to be the distribution of population amongst compartments at the end of outbreak. In full, the population proportion that is susceptible, symptomatic recovered or asymptomatic recovered post outbreak. The true end-state is taken from the 50 SEIsIaR model simulated datasets used for inference in Figure 3.9. For each corresponding posterior distribution, 100 parameter sets are sampled to generate *posterior simulation* outbreaks. The *mean posterior simulation end-state* is calculated from each set of posterior simulations. We compare the true end-state distribution and the mean posterior end-state distribution, in Figure 3.10. Note, all simulated outbreaks with less than 10 cases are removed from consideration, as with the true datasets.



(a) Low R_0

(b) High R_0

Figure 3.10: Box plot diagram of end-state distribution comparison of dataset and posterior simulations. Datasets and posterior results taken from inference in Figure 3.9.

In support of the above theory, the posterior simulations demonstrate a much higher *total attack rate* (attack rate of observed and unobserved infections) than the data due to a large increase in the number of asymptomatic infections. In the low R_0 datasets, approximately 70% of the population escape infection, where posterior simulations estimate only 30% escape infection. In the high R_0 datasets, 20% of the population escape infection, where the posterior simulations estimate 2% escape infection. Hence, the overestimated R_0/underestimated p_s bias causes the posterior distribution to overestimate asymptomatic infections and transmission. This bias should be considered when conducting inference with asymptomatic infections from a single outbreak.

3.2.4 SEIpIsIaR Model Parameter Inference

The state space size for the SEIpIsIaR model is given by

$$\begin{aligned} |\mathcal{S}| &= \sum_{Z_1=0}^{N} \sum_{Z_2=0}^{Z_1} \sum_{Z_3=0}^{Z_2} \sum_{Z_4=0}^{Z_3} \sum_{Z_5=0}^{Z_1-Z_2} \sum_{Z_6=0}^{Z_5} 1, \\ &\sim \mathcal{O}(N^6). \end{aligned} \tag{3.12}$$

Again, the cases where NumInt(τ) is a viable solution is restricted to very small populations; therefore NumInt(τ) is eliminated from further consideration.

A study is conducted on simulated data to test the ability of the PMMH algorithm to identify the parameters of the SEIpIsIaR model. For the SEIpIsIaR model, we conduct inference on the model parameter transformations

$$\begin{aligned} \theta = \Bigg\{ R_0 = p_s \left(\frac{\beta_p}{\alpha} + \frac{\beta_s}{\gamma_s} \right) + \frac{(1-p_s)\beta_a}{\gamma_a}, \ \frac{1}{\sigma}, \ \frac{1}{\alpha}, \ \frac{1}{\gamma_s}, \ \frac{1}{\gamma_a}, \ p_s, \\ \kappa_S = \frac{(R_0)_S}{R_0}, \ \kappa_P = \frac{(R_0)_P}{R_0} \Bigg\}. \end{aligned}$$

Denote $(R_0)_P$ to be the R_0 contribution of pre-symptomatic individuals, such that

$$\begin{aligned} (R_0)_P &= \frac{p_s \beta_p}{\alpha}, \\ (R_0)_S &= \frac{p_s \beta_s}{\gamma_s}, \\ (R_0)_A &= \frac{(1-p_s)\beta_a}{\gamma_a}. \end{aligned}$$

Hence, κ_S, κ_P and $\kappa_A = (1 - \kappa_S - \kappa_P)$ denote the *proportion of transmission* occurring from symptomatic, pre-symptomatic and asymptomatic individuals. All parameters are assigned uniform$(0.1, 8)$ prior distributions except p_s, κ_S and κ_P which are assigned a uniform$(0.01, 0.99)$ prior distribution (such that $\kappa_S + \kappa_P < 1$). Again, the κ_S and κ_P prior distributions reflects our uncertainty in the respective proportion of transmission. The PMMH algorithm is run on 50 SEIpIsIaR model simulated datasets for 80,000 iterations with 5,000 steps as burn-in. Note, the iteration count is chosen such that $\widehat{\text{Var}}(\bar{R}_0) \approx 0.1$. We run two studies on data produced with low $R_0 = 1.2$ and high $R_0 = 2$. The bias of posterior results are displayed in Figure 3.11.

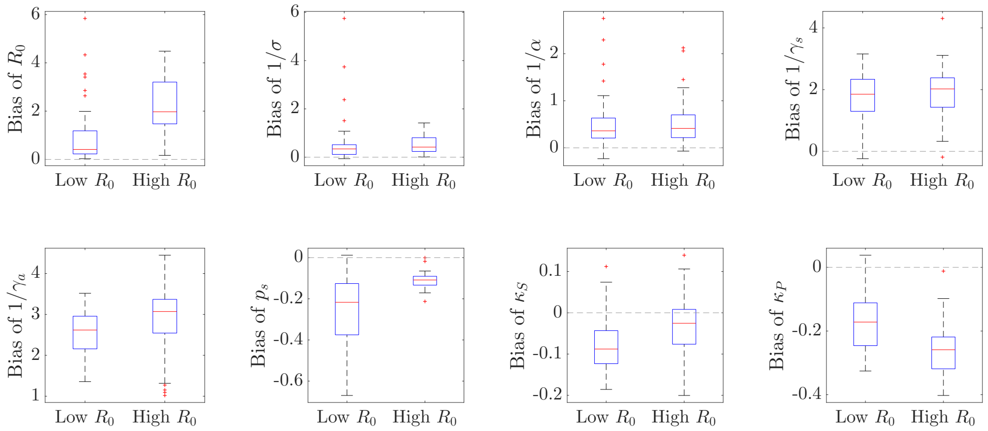
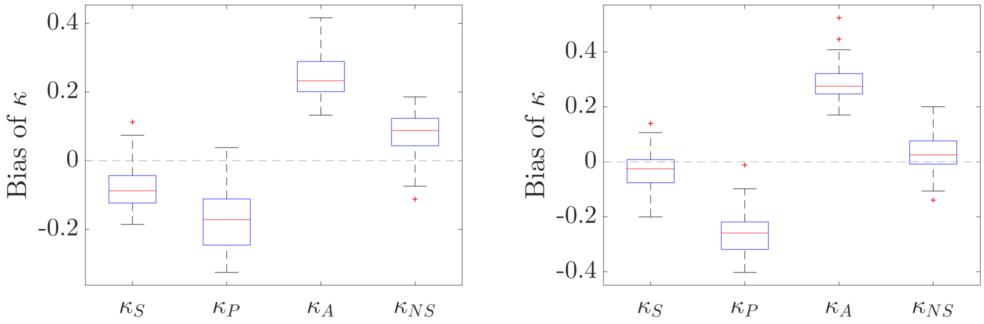


Figure 3.11: Box plot diagram of PMMH posterior results for 50 SEIpIsIaR model simulated datasets generated with low $R_0 = 1.2$ values and high $R_0 = 2$ values. Posterior sample mean bias are displayed in each box plot by parameter. Datasets: $N = 1000$, observed the length of outbreak at daily resolution, $\frac{1}{\sigma} = 0.5$, $\frac{1}{\alpha} = 0.5$, $\frac{1}{\gamma_s} = 2$, $\frac{1}{\gamma_a} = 1$, $p_s = 0.75$, $\kappa_S = 0.31$, $\kappa_P = 0.63$.

The SEIpIsIaR model shows a similar ability to the SEIsIaR model to identify R_0, $1/\sigma$, $1/\gamma_s$, p_s (with the same R_0/p_s bias) and an inability to identify $1/\gamma_a$. Conversely to the asymptomatic dynamics, we are able to accurately identify $1/\alpha$, due to it's relation to the incubation period. Although, in particular we are interested in the ability to identify the proportion of transmission contributed by symptomatic, κ_S, pre-symptomatic κ_P, and asymptomatic individuals, κ_A. We include in this analysis the combined contributions of *non-symptomatic* (pre-symptomatic or asymptomatic) individuals, $\kappa_{NS} = (1 - \kappa_S)$. The posterior sample mean bias of κ_S, κ_P, κ_A and κ_{NS} are shown in Figure 3.12. Here, we observe significant overestimation of κ_A and underestimation of κ_P in both cases. However, we observe an ability to accurately identify κ_S and κ_{NS}. Hence, while we cannot identify the correct transmission contribution from "hidden" pre-symptomatic and asymptomatic individuals, we have shown the ability to identify the respective transmission from symptomatic and non-symptomatic individuals.



(a) Low R_0

(b) High R_0

Figure 3.12: Box plot diagram of proportion of transmission posterior sample mean bias. Posterior results extracted from Figure 3.11.

The introduction of prior immunity into the SEIpIsIaR model complicates parameter inference as there are now two processes that allow individuals to not be observable infections. That is, it is difficult to determine if the proportion of the population that are not symptomatic infections are: immune, asymptomatic individuals or if the disease died out before infecting them. We conduct a study on simulated data to test the ability to identify prior immunity within the SEIpIsIaR model. For the SEIpIsIaR model with prior immunity, we conduct inference on the model parameter transformations

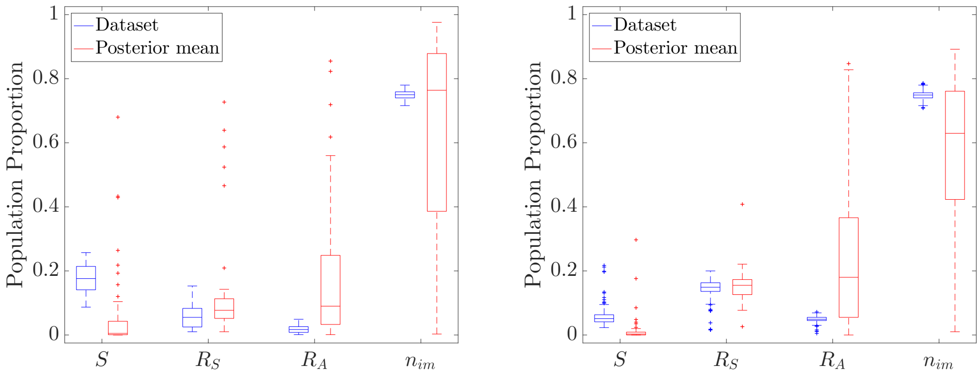
$$\begin{aligned}\theta = \Bigg\{ R_0 = p_{sus}\left(p_s\left(\frac{\beta_p}{\alpha}+\frac{\beta_s}{\gamma_s}\right)+\frac{(1-p_s)\beta_a}{\gamma_a}\right),\ \frac{1}{\sigma},\ \frac{1}{\alpha},\\ \frac{1}{\gamma_s},\ \frac{1}{\gamma_a},\ p_s,\ \kappa_S=\frac{(R_0)_S}{R_0},\ \kappa_P=\frac{(R_0)_P}{R_0},\ p_{sus}\Bigg\},\end{aligned} \tag{3.13}$$

where

$$\begin{aligned}(R_0)_P &= p_{sus}\left(\frac{p_s\beta_p}{\alpha}\right),\\ (R_0)_S &= p_{sus}\left(\frac{p_s\beta_s}{\gamma_s}\right),\\ (R_0)_A &= p_{sus}\left(\frac{(1-p_s)\beta_a}{\gamma_a}\right).\end{aligned}$$

All parameters are assigned the same prior distributions as for the SEIpIsIaR model, with p_{sus} being assigned a uniform$(0.01, 1)$ prior distribution. The PMMH algorithm is run on 50 SEIpIsIaR model with prior immunity simulated datasets for 150,000 iterations with 5,000 steps as burn-in. Note, the iteration count is chosen such that $\widehat{\text{Var}}(\bar{R}_0) \approx 0.1$. Again, we run two studies on data produced with low $R_0 = 1.2$ and high $R_0 = 2$. The bias of posterior results are displayed in Figure 3.14.

The posterior results in Figure 3.14 demonstrate the difficulty in identifying prior immunity from a single dataset. Again, the model demonstrates the same R_0/p_s bias as seen in the SEIsIaR model, and we see a degree of uncertainty about p_{sus}. The difficulty of identifying these dynamics from a single dataset can be demonstrated by a simulation end-state comparison. We compare the true end-state distribution and the posterior mean end-state distribution in Figure 3.13. Here, we see the same greatly overestimated total attack rate caused by a larger number of asymptomatic infections as a result of the R_0/p_s interaction. More importantly, we note great variability in the number of immune individuals; again, an interaction with the unknown asymptomatic infections. This result demonstrates the inability to adequately identify p_{sus} from a single outbreak.



(a) Low R_0

(b) High R_0

Figure 3.13: Box plot diagram of end-state distribution comparison of dataset and posterior simulations. Datasets and posterior results taken from inference in Figure 3.14.

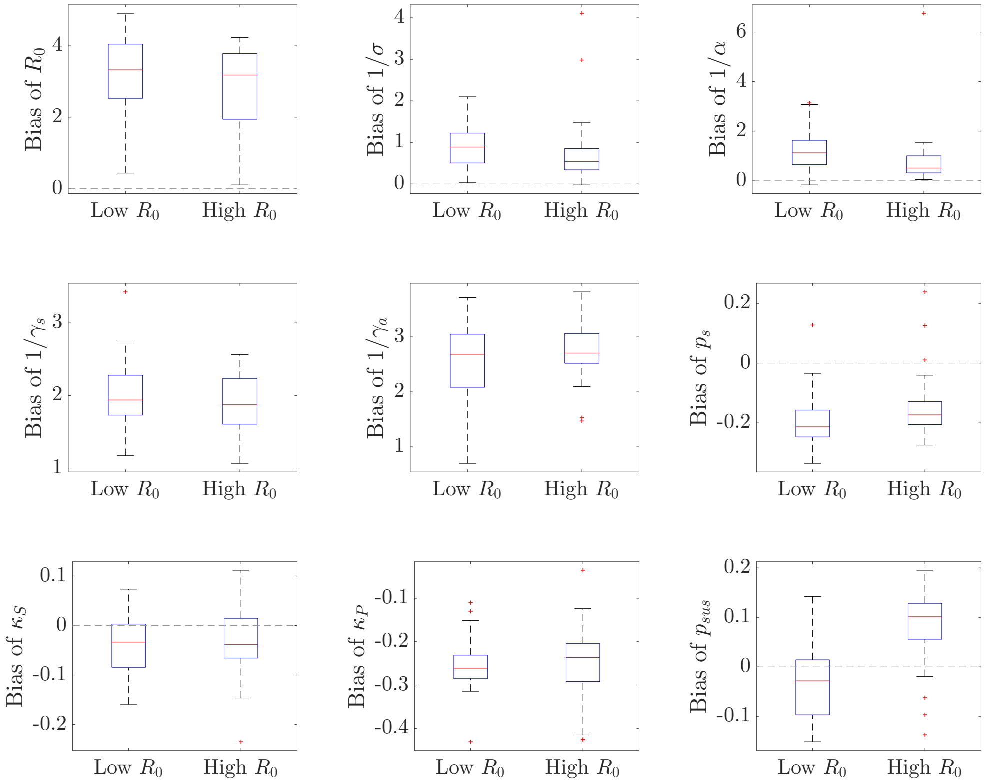


Figure 3.14: Box plot diagram of PMMH posterior results for 50 SEIpIsIaR model with prior immunity simulated datasets generated with low $R_0 = 1.2$ values and high $R_0 = 2$ values. Posterior sample mean bias are displayed in each box plot by parameter. Datasets: $N = 1000$, observed the length of outbreak at daily resolution, $\frac{1}{\sigma} = 0.5$, $\frac{1}{\alpha} = 0.5$, $\frac{1}{\gamma_s} = 2$, $\frac{1}{\gamma_a} = 1$, $p_s = 0.75$, $\kappa_S = 0.31$, $\kappa_P = 0.63$, $p_{sus} = 0.25$.

This section highlights the issue of identifiability within the SEIpIsIaR model with prior immunity. The model contains too many parameters to accurately identify from a single outbreak with symptomatic cases observations; a symptom of the *curse of dimensionality.* The only solution is to use more data for inference. If multiple outbreaks share disease dynamics, we can be better informed about said dynamics by combining the information from each outbreak. This is a common practice to characterise epidemic and pandemic level disease outbreaks [49, 54]. Hence, we next investigate if we are able to identify these dynamics by conducting inference on multiple outbreak datasets through parallel inference.

3.2.5 SEIpIsIaR Model Parallel Inference

We conduct a study of parallel inference on multiple datasets to further assert if the parameters of the SEIpIsIaR model with prior immunity are identifiable. Here, *parallel inference* refers to the process of using a PMMH algorithm to target the joint posterior of K independent datasets $\{D^{(1)}, ..., D^{(K)}\}$, under the assumption that they are realisations of a shared model. That is, we assume each dataset $\{D^{(1)}, ..., D^{(K)}\}$ is a realisation of a model under the same (or a subset of common) parameters. The posterior distribution,

$$P(\theta|\{D^{(1)}, ..., D^{(K)}\}) \propto P(\{D^{(1)}, ..., D^{(K)}\}|\theta)P(\theta),$$

can be evaluated by expressing the joint likelihood as

$$L(\theta) := P(\{D^{(1)}, ..., D^{(K)}\}|\theta) = \prod_{k=1}^{K} P(D^{(k)}|\theta).$$

As each $P(D^{(k)}|\theta)$ can be estimated by a SMC(N_p) likelihood, $\hat{P}(D^{(k)}|\theta)$, we can estimate the joint likelihood as

$$\hat{L}(\theta) = \prod_{k=1}^{K} \hat{P}(D^{(k)}|\theta).$$

Using multiple datasets within the inference scheme can result in a more informative likelihood function, which can improve our ability to identify parameters compared to using a single dataset. This practice of parallel inference is relevant as we will use multiple datasets for inference in epidemiologically characterising the 1918 pandemic ship data (see Chapter 4).

Fitting the model to multiple datasets imbues the posterior distribution with more information about the true parameter values. The more datasets we include, the more confidence we can have in the accuracy of the posterior distribution. Ideally the more datasets inferred upon the better, but the upscale of data is not without difficulty. With each dataset, we greatly increase the runtime of the PMMH algorithm through multiple SMC likelihood calculations per iteration. Also, we increase the probability of SMC likelihood failure as it requires only one estimate of $\hat{P}(D^{(k)}|\theta)$ to suffer to return an overall zero-likelihood, $\hat{L}(\theta)$. As such, for parallel inference we increase the standard number of particles $N_p = 25,000$ to reduce this occurrence.

Here, we give an example of the potential benefits of parallel inference for comparison with the single dataset inference conducted in Section 3.2.4. We take 125 SEIpIsIaR model with prior immunity simulated datasets, randomly sorted into 25 collections of 5 datasets for inference. We assume the same parameter set produced each collection of 5 datasets. Hence, inference is conducted on the same parameter transforms as presented in Equation (3.13) with the same prior distributions. Like before, we run two studies on data produced with low $R_0 = 1.2$ and high $R_0 = 2$. The PMMH algorithm is run on each set of 5 datasets for 60,000 iterations with 5,000 steps as burn-in. Note, the iteration count is chosen such that $\widehat{\text{Var}}(\bar{R}_0) \approx 0.1$. The bias of posterior results are displayed in Figure 3.15.

The use of parallel inference with only 5 datasets in each collection greatly improves the accuracy of the posterior distribution and reduces the bias of the posterior estimates. These results are only from 25 runs so some variation can be explained by a smaller sample size, but they do indicate the parameter identifiability benefits from parallel inference. We can also verify the increased accuracy of the posterior distribution by the end-state comparison. We compare the true end-state distribution and the posterior mean end-state distribution in Figure 3.16. Here, we see the posterior mean end-state distributions are far more consistent with the data distributions than previously seen in single dataset inference. Although there still exists some slight bias in overestimating the total attack rate, this is an encouraging result for the benefits of parallel inference. Hence, there is great motivation to conduct inference on multiple datasets, given we can assume they are modelled by the same process and common parameters. Although, this assumption should be scrutinized for any practical situations.

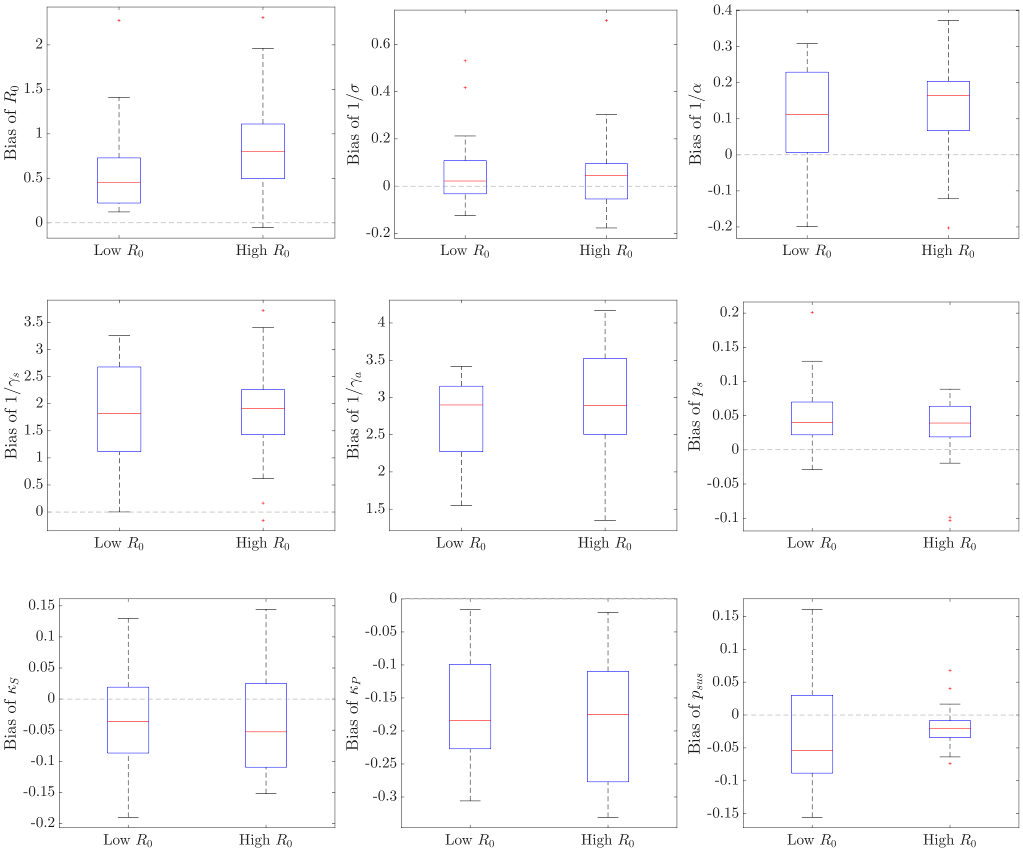
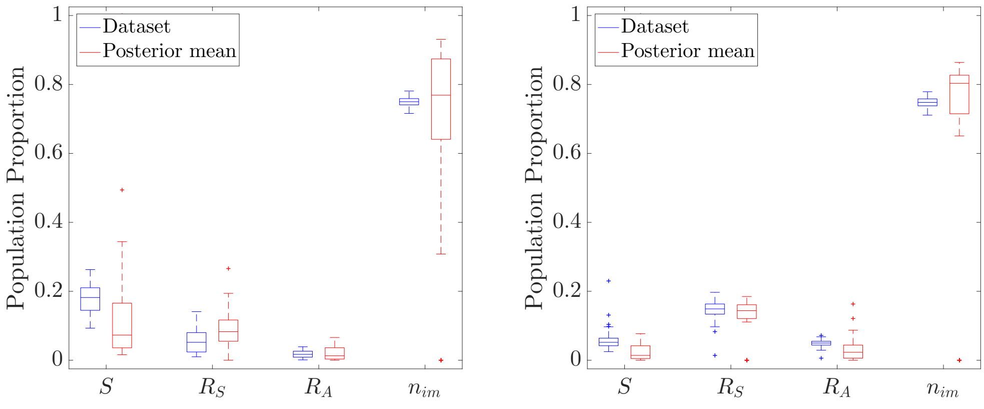


Figure 3.15: Box plot diagram of PMMH parallel inference posterior results for 25 sets of 5 SEIpIsIaR model with prior immunity simulated datasets generated with low $R_0 = 1.2$ values and high $R_0 = 2$ values. Posterior sample mean bias are displayed in each box plot by parameter. Datasets: $N = 1000$, observed the length of outbreak at daily resolution, $\frac{1}{\sigma} = 0.5$, $\frac{1}{\alpha} = 0.5$, $\frac{1}{\gamma_s} = 2$, $\frac{1}{\gamma_a} = 1$, $p_s = 0.75$, $\kappa_S = 0.31$, $\kappa_P = 0.63$, $p_{sus} = 0.25$.



(a) Low R_0 (b) High R_0

Figure 3.16: Box plot diagram of end-state distribution comparison of dataset and parallel inference posterior simulations. Datasets and posterior results taken from inference in Figure 3.15.

In this thesis, parallel inference is used to epidemiologically characterise the 1918 pandemic influenza strain from the ship data (as detailed in Chapter 4). However, we cannot safely assume common R_0 values between ships (see Chapter 4 for an in-depth discussion). As such, we are interested in the effect of allowing unique R_0 values between ships upon inference. We repeat the previous experiment, except a unique R_0 value randomly sampled from a uniform$(1, 3)$ distribution is used to generate each dataset. We take 125 SEIpIsIaR model with prior immunity simulated datasets, randomly sorted into 25 collections of 5 datasets. We assume unique R_0 values (and an otherwise common parameter set) produced each dataset and conduct inference on the following parameter transformations,

$$\theta = \left\{ R_0^{(1)}, ..., \ R_0^{(5)}, \ \frac{1}{\sigma}, \ \frac{1}{\alpha}, \frac{1}{\gamma_s}, \ \frac{1}{\gamma_a}, \ p_s, \ \kappa_S, \ \kappa_P, \ p_{sus} \right\}.$$

We use the same prior distributions as with the common R_0 inference above. The PMMH algorithm is run on each set of 5 datasets for 60,000 iterations with 5,000 steps as burn-in. Note, the iteration count is chosen such that $\widehat{\text{Var}}(\bar{R}_0) \approx 0.1$. The bias of unique R_0 posterior results (alongside the previous common R_0 results) are displayed in Figure 3.17.

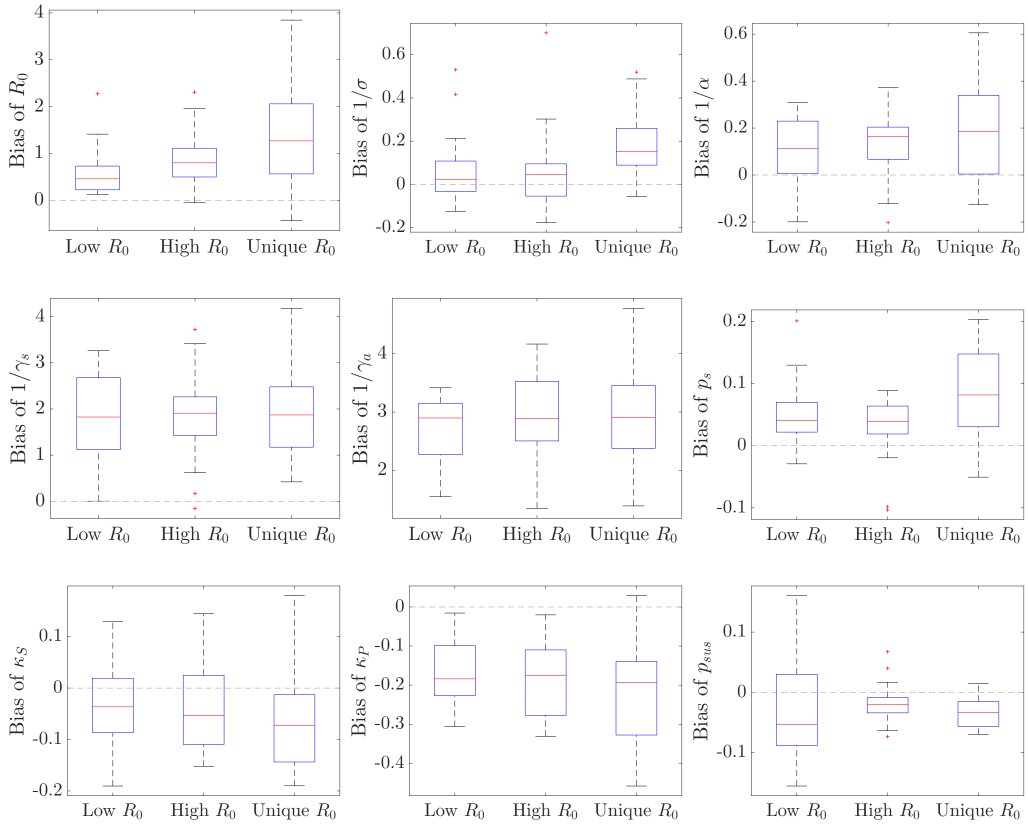


Figure 3.17: Box plot diagram of PMMH parallel inference posterior results for 25 sets of 5 SEIpIsIaR model with prior immunity simulated datasets generated with low $R_0 = 1.2$, high $R_0 = 2$ and unique R_0 values $\sim$ uniform(1,3). Posterior sample mean bias are displayed in each box plot by parameter. Datasets: $N = 1000$, observed the length of outbreak at daily resolution, $\frac{1}{\sigma} = 0.5$, $\frac{1}{\alpha} = 0.5$, $\frac{1}{\gamma_s} = 2$, $\frac{1}{\gamma_a} = 1$, $p_s = 0.75$, $\kappa_S = 0.31$, $\kappa_P = 0.63$, $p_{sus} = 0.25$.

The use of unique R_0 values between datasets has slightly degraded our ability to identify parameters, particularly R_0, when compared to the common R_0 inference. Although, we still observe marked improvements in bias reduction compared to single dataset inference. Hence, parallel inference for the SEIpIsIaR model with prior immunity can greatly improve parameter identifiability. We have also shown that allowing for unique R_0 values between outbreaks doesn't have a severe detrimental effect on identifying common parameters.

The key features of this model is that it accounts for prior immunity and pre-symptomatic and asymptomatic transmission. These are features we wish to identify for the 1918 pandemic, and therefore we are interested in the accuracy of our inference methods in targeting them. Given we find these features are identifiable in simulations studies, we can have confidence in our estimates for the 1918 pandemic. Hence, we are particularly interested in the our ability to identify the p_{sus} parameter, which defines the prior immunity within the population. Previously, we found that inferring p_{sus} from a single dataset can be difficult, but parallel inference on multiple data sets greatly increases the accuracy of the posterior estimates. As seen in Figure 3.17, we can typically identify the true p_{sus} value within 5-10% of the true value, but tend to underestimate slightly. We support this by the end-state comparison. We compare the true end-state distribution and the posterior mean end-state distribution in Figure 3.18. Here, we see good agreement in the proportion of immune individuals between the dataset and the posterior simulations. Therefore, we maintain confidence in our ability to identify prior immunity within the SEIpIsIaR model while using parallel inference.

The ability to identify pre-symptomatic and asymptomatic transmission is a more complicated issue. Previously, the proportion of transmission arising from pre-symptomatic (κ_P) and asymptomatic (κ_A) individuals were unable to be identified when conducting inference on a single dataset, but the symptomatic (κ_S) and non-symptomatic (κ_{NS}) contributions were. Using the inference in Figure 3.17, we test our ability to identify these ratios. The posterior sample mean bias of κ_S, κ_P, κ_A and κ_{NS} are shown in Figure 3.19. Again, we see an inability to adequately identify κ_P and κ_A, with corresponding underestimation of the pre-symptomatic contribution and overestimation of the asymptomatic contribution. Although, we still observe an ability to accurately identify the κ_S and κ_{NS} contributions to within 5-10% of the true value. Hence, while we may not be able to determine the differences between the pre-symptomatic and asymptomatic transmission, we are able to identify the proportion of transmission arising from symptomatic and non-symptomatic individuals.

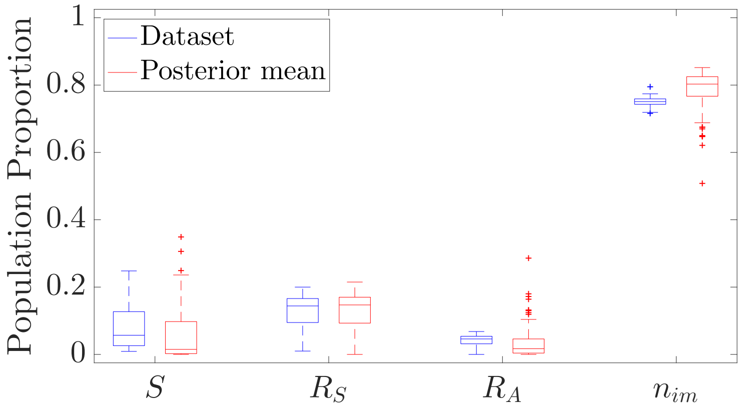


Figure 3.18: Box plot diagram of end-state distribution comparison of dataset and parallel inference posterior simulations. Datasets and posterior results taken from inference in Figure 3.17.

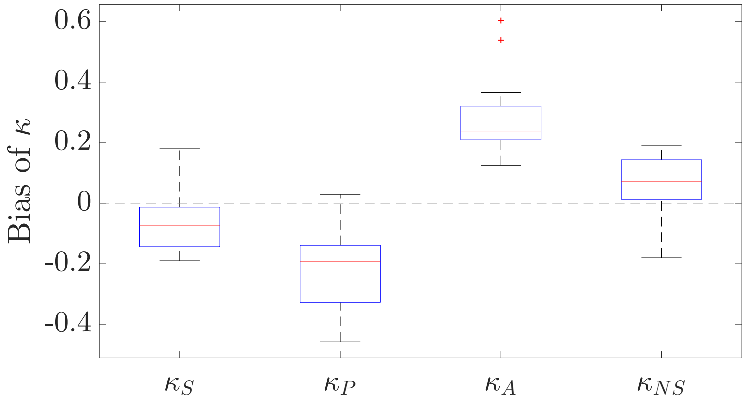


Figure 3.19: Box plot diagram of proportion of transmission posterior sample mean bias. Posterior results extracted from Figure 3.17.

The overestimation of the asymptomatic transmission contribution appears related to the *expected number of infections* that arise from an asymptomatic infectious individual. Define the expected number of infections to arise from a single symptomatic, pre-symptomatic or asymptomatic individuals as

$$\begin{aligned} E[I_S \text{ infections}] &= p_{sus}\left(\frac{\beta_s}{\gamma_s}\right), \\ E[I_P \text{ infections}] &= p_{sus}\left(\frac{\beta_p}{\alpha}\right), \\ E[I_A \text{ infections}] &= p_{sus}\left(\frac{\beta_a}{\gamma_a}\right). \end{aligned}$$

The posterior sample mean bias of the expected number of infections are shown in Figure 3.20. Here, we see an ability to accurately identify the respective expected number of infections from symptomatic and pre-symptomatic individuals, but a severe inability to identify the expected number of infections for asymptomatic individuals. As such, the $E[I_A$ infections] estimates tend towards the prior distribution mean, which overestimates the proportion of transmission from asymptomatic individuals, κ_A. This could cause the underestimation of κ_P as the pre-symptomatic proportion of transmission is lowered to fit $\kappa_{NS} = \kappa_P + \kappa_A$, and the overestimation of p_s as the number of asymptomatic infections is reduced to account for their increased expected infections. Although, note we correctly estimate the ordering of $\kappa_P > \kappa_A$; this indicates identifying κ_P and κ_A may be possible with a larger number of datasets. This dynamic should be noted in future model inference, and is discussed further in Chapter 5.

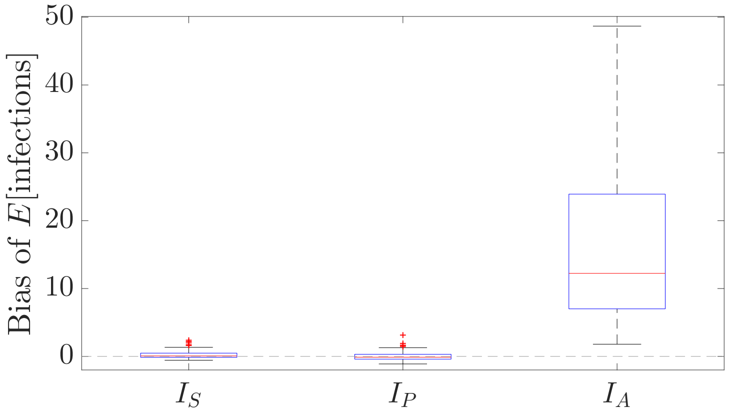


Figure 3.20: Box plot diagram of expected number of infections posterior sample mean bias. Posterior results extracted from Figure 3.17.

In summary, we have developed the SEIpIsIaR model to reflect the disease dynamics of influenza; this includes modelling previously little understood phenomena such as the pre-symptomatic period, asymptomatic infections and their respective contributions to transmission. We have also included a mechanism to account for the presence of prior immunity within the population, an important consideration for pandemic influenza. We have demonstrated a novel method of inference in the PMMH algorithm that makes it possible to conduct inference on the SEIpIsIaR model with prior immunity and identify the key parameters that dictate the model dynamics. In doing so, we have identified possible biases inherently within the likelihood function that must be considered, and ways to reduce their effects by conducting inference on multiple datasets. We have shown that by utilizing parallel inference we are able to better identify the parameters within the model and characterise the epidemiological dynamics dictating the simulated outbreaks. In particular, we have shown an ability to accurately identify key features of pandemic influenza including prior immunity and non-symptomatic transmission; these results validate the use of the model and inference method for further study of real data. These processes will be required to characterise the epidemiology of the 1918 pandemic influenza strain from multiple ship outbreaks within Chapter 4.

As an aside, the following section proposes an alternative method of likelihood calculation, built upon the SMC algorithm, for further consideration. SMC particle degeneracy, likelihood failure and the negative effect they can have on the mixing of PMMH algorithm has been well documented within this thesis. We propose a *Hybrid* algorithm that combines SMC and exact likelihood methodologies to combat these issues. This algorithm was not applicable to our analysis here for computational reasons, but may have benefits for other models/studies.

3.3 Hybrid Algorithm

As discussed in Chapter 2 and Section 3.2, SMC particle degeneracy and likelihood failure can have a significant negative effect on the mixing of the MH chain. To recap, likelihood failure occurs if all particles simulations end in states that return a non-positive estimate of a positive likelihood; i.e.,

$$\hat{P}(y_{t+1}|\mathbf{y}_{1:t}) = \frac{1}{N_p}\sum_{i=1}^{N_p} P(y_{t+1}|x_{t+1}^{*(i)}) = 0.$$

Particle degeneracy occurs if the number of particles that contribute a positive probability of observing y_{t+1} is too low, hence introducing too much error into the Monte Carlo estimate of the likelihood and the re-sampled particle distribution,

$$\{x_{t+1}^{(1)}, ..., x_{t+1}^{(N_p)}\} \sim p(x_{t+1}|\mathbf{y}_{1:t+1}). \tag{3.14}$$

These issues are caused by *rare event* observations that are inconsistent with the model, and so have a very low probability of occurring. These low probability observation issues could theoretically be solved by increasing the number of particles. Although, this is often undesirable due to the increased runtime of the simulations and the unknown number of particles required.

We propose a likelihood calculation algorithm to counteract these issues, denoted the *Hybrid algorithm*, which is essentially a combination of the SMC and "exact" (using the implicit Euler method to solve the master equation) likelihood calculation methods presented in Chapter 2. In short, the Hybrid algorithm decides between using an SMC or exact approach on an observation-by-observation basis. The algorithm progresses using SMC, but if an observation's likelihood estimate suffers from particle degeneracy/likelihood failure then the likelihood is recalculated by the exact method. The Hybrid(N_p, τ) algorithm is given in Algorithm 5.

The Hybrid(N_p, τ) algorithm works as follows. As in SMC, consider the set of particles at time t,

$$\{x_{t}^{(1)}, ..., x_{t}^{(N_p)}\} \sim p(x_{t}|\mathbf{y}_{1:t}). \tag{3.15}$$

These particles are forward simulated to generate

$$\{x_{t+1}^{*(1)}, ..., x_{t+1}^{*(N_p)}\} \sim p(x_{t+1}|\mathbf{y}_{1:t})$$

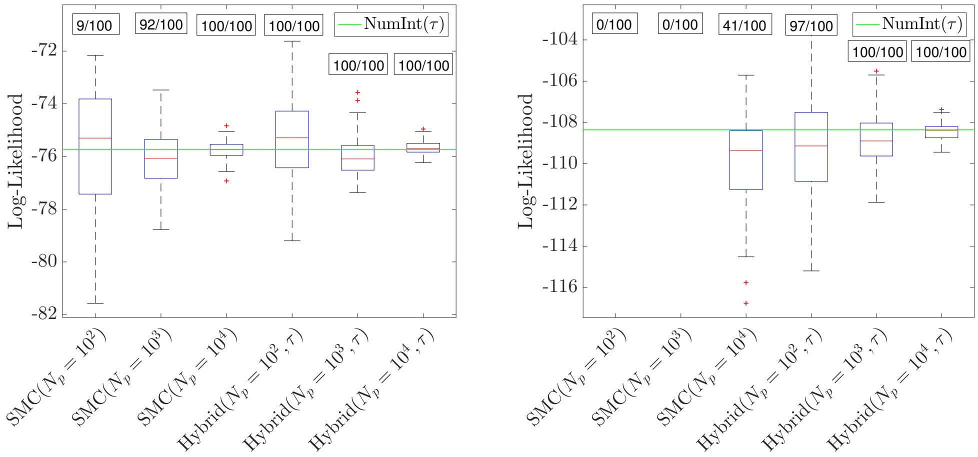
and used calculate a likelihood estimate, $\hat{P}(y_{t+1}|\mathbf{y}_{1:t})$. The Hybrid algorithm will recalculate the likelihood estimate using exact methods, if there is evidence of particle degeneracy or likelihood failure. Hence, we introduce the *probability threshold*, δ, such that if a likelihood estimate $\hat{P}(y_{t+1}|\mathbf{y}_{1:t}) < \delta$, the Hybrid$(N_p, \tau)$ algorithm will recalculate the likelihood of y_{t+1} using exact likelihood methodology.

```
Data: Data $\mathbf{y}_{1:T}$, initial distribution $p(x_0)$, number of particles $N_p$,
      precision $\tau$, probability threshold $\delta$.
Set $t = 0$;
Sample $\{x_0^{(1)}, ..., x_0^{(N_P)}\} \sim p(x_0)$ ;
for $i = 1, ..., N_p$ do
    Simulate $x_1^{*(i)} \sim p(x_1|x_0^{(i)})$ ;
    Assign weight $W_1^{(i)} = P(y_1|x_1^{*(i)})$ ;
end
Calculate likelihood $\hat{P}(y_1) = \frac{1}{N_p}\sum_{i=1}^{N_p} W_1^{(i)}$;
if $\hat{P}(y_1) \geq \delta$ then
    Re-sample $\{x_1^{(1)}, ..., x_1^{(N_P)}\}$ with weights $P(x_1 = x_1^{*(j)}) = \frac{W_1^j}{\sum_k W_1^k}$ ;
else
    Calculate $p(x_1|x_0)$ from $p(x_0)$ using implicit Eulers method;
    Calculate likelihood $\hat{P}(y_1) = \sum_{x_0 \in \mathcal{S}} P(y_1|x_0)P(x_0)$;
    Sample $\{x_1^{(1)}, ..., x_1^{(N_P)}\} \sim p(x_1|y_1)$;
end
for $t = 2, ..., T$ do
    for $i = 1, ..., N_p$ do
        Simulate $x_t^{*(i)} \sim p(x_t|x_{t-1}^{(i)})$ ;
        Assign weight $W_t^{(i)} = P(y_t|x_t^{*(i)})$ ;
    end
    Calculate likelihood $\hat{P}(y_t|\mathbf{y}_{1:t-1}) = \frac{1}{N_p}\sum_{i=1}^{N_p} W_t^{(i)}$;
    if $\hat{P}(y_t|\mathbf{y}_{1:t-1}) \geq \delta$ then
        Re-sample $\{x_t^{(1)}, ..., x_t^{(N_P)}\}$ with weights
        $P(x_{t+1} = x_{t+1}^{*(j)}) = \frac{W_{t+1}^j}{\sum_k W_{t+1}^k}$;
    else
        Normalise $\tilde{p}(x_{t-1}|\mathbf{y}_{1:t-1})$ from $\{x_{t-1}^{(1)}, ..., x_{t-1}^{(N_P)}\}$;
        Calculate $p(x_t|\mathbf{y}_{1:t-1})$ from $\tilde{p}(x_{t-1}|\mathbf{y}_{1:t-1})$ using implicit Eulers
        method;
        Calculate likelihood $\hat{P}(y_t|\mathbf{y}_{1:t-1}) = \sum_{x_t \in S} P(y_t|x_t)P(x_t|\mathbf{y}_{1:t-1})$;
        Sample $\{x_t^{(1)}, ..., x_t^{(N_P)}\} \sim p(x_t|y_{1:t})$;
    end
end
Marginal likelihood given by $\hat{P}(\mathbf{y}_{1:T}) = \hat{P}(y_1)\prod_{t=2}^{T}\hat{P}(y_t|\mathbf{y}_{1:t-1})$;
Result: Marginal likelihood estimate $\hat{P}(\mathbf{y}_{1:T})$.
```

Algorithm 5: Hybrid algorithm.

This is done by approximating the Markov process distribution by the particles at time t. That is, by normalising across the Markov process states of particles $\{x_t^{(1)}, ..., x_t^{(N_p)}\}$, it is possible to approximate the probability mass function, $\tilde{p}(x_t|\mathbf{y}_{1:t})$. This is a valid approximation as the particle distribution has been to shown to converge in distribution to said probability mass function as $N_p \to \infty$ (see Chapter 2). The particle-estimated $\tilde{p}(x_t|\mathbf{y}_{1:t})$ is used as the initial distribution to calculate $\tilde{p}(x_{t+1}|\mathbf{y}_{1:t})$ using the implicit Euler method. The distribution is conditioned on y_{t+1}, $\tilde{p}(x_{t+1}|\mathbf{y}_{1:t+1})$, which is then used to calculate the likelihood $\tilde{P}(y_{t+1}|\mathbf{y}_{1:t})$. Hence, we obtain a *pseudo-exact* likelihood of the observation, $\tilde{P}(y_{t+1}|\mathbf{y}_{1:t})$, and the probability mass function, $\tilde{p}(x_{t+1}|\mathbf{y}_{1:t+1})$. The term pseudo-exact is used as accuracy of the exact methods is dependent on the estimated initial distribution $\tilde{p}(x_t|\mathbf{y}_{1:t})$. The algorithm samples from $\tilde{p}(x_{t+1}|\mathbf{y}_{1:t+1})$ to generate a set of particles $\{x_{t+1}^{(1)}, ..., x_{t+1}^{(N_p)}\}$ and continues the SMC approach. Hence, given that N_p is chosen sufficiently high to accurately estimate the probability mass function and the implicit Euler precision τ is chosen sufficiently small, the Hybrid(N_p, τ) algorithm will produce an unbiased estimate of the likelihood. The Hybrid(N_p, τ) algorithm reduces the frequency of particle degeneracy/likelihood failure and the estimator variance, when compared to the SMC(N_p) algorithm. Note, it is still possible for the Hybrid algorithm to suffer likelihood failure, but this is far less frequent than in SMC(N_p). This occurs when every particle reaches an absorbing state such that it cannot enter a state that contributes a positive probability of observation.

Here, we compare the SMC(N_p) and Hybrid(N_p, τ) likelihood methods on a SIR model simulated dataset, to highlight the benefits of the Hybrid algorithm. The algorithms are compared in calculating the likelihood of a dataset with the true "good-fit" ($R_0 = 2, 1/\gamma = 3$) parameters and a chosen "poor-fit" ($R_0 = 3, 1/\gamma = 3$) parameter set. The box plots of 100 SMC(N_p) and Hybrid(N_p, τ) estimates by varying N_p, alongside the NumInt(τ) solution are shown in Figure 3.21 and comparative algorithm runtimes are given in Table 3.10. Note, probability threshold $\delta = 10^{-2}$ will be used as standard for the Hybrid algorithm and $\tau = 10^{-3}$ will be standard for Hybrid and NumInt algorithms.



(a) Good-fit parameters.

(b) Poor-fit parameters.

Figure 3.21: SIR model log-likelihood comparison for SMC(N_p) and Hybrid(N_p, τ) by N_p. Box plot of 100 SMC(N_p) and Hybrid(N_p, τ) calculations for each N_p with number of positive likelihood estimates out of 100 given in box. NumInt(τ) solution show for comparison. Dataset: $N = 1000$, observed the first 25 days of outbreak at daily resolution, $R_0 = 2$, $\frac{1}{\gamma} = 3$.

Algorithm	Runtime (s)
NumInt(τ)	212
SMC($N_p = 10^2$)	< 0.01
SMC($N_p = 10^3$)	0.06
SMC($N_p = 10^4$)	0.62
Hybrid($N_p = 10^2, \tau$)	39.7
Hybrid($N_p = 10^3, \tau$)	14.7
Hybrid($N_p = 10^4, \tau$)	15.8

Table 3.8: Good-fit parameters.

Algorithm	Runtime (s)
NumInt(τ)	220
SMC($N_p = 10^2$)	N/A
SMC($N_p = 10^3$)	N/A
SMC($N_p = 10^4$)	0.77
Hybrid($N_p = 10^2, \tau$)	157
Hybrid($N_p = 10^3, \tau$)	92.8
Hybrid($N_p = 10^4, \tau$)	96.6

Table 3.9: Poor-fit parameters.

Table 3.10: NumInt(τ) runtime, SMC(N_p) and Hybrid(N_p, τ) average runtime by number of particles N_p. Runtime taken from likelihood calculations given in Figure 3.21. Run in `Matlab` on an iMac (2013) with 2.7 GHz Intel Core i5 processor.

For the good-fit parameters, we see the SMC algorithm suffers significant likelihood failure for $N_p = 10^2$ and $N_p = 10^3$, where as the Hybrid algorithm always returns a positive likelihood estimate. The variance of the Hybrid estimate is roughly half that of the SMC estimate. Hence, the use of the Hybrid algorithm always returns a positive likelihood estimate, and the variance of the estimator is greatly reduced. Obviously, this comes at significant computational cost, but the runtime of the Hybrid algorithm is still roughly 12 times faster than the NumInt(τ) runtime. For the poor-fit parameters, we see a similar behaviour, as the SMC estimate fails to return a likelihood estimate for any runs with less than $N_p = 10^4$ and less than half of the runs are completed when $N_p = 10^4$. In comparison, the Hybrid algorithm nearly always returns a positive likelihood estimate. The variance of the Hybrid algorithm estimate is approximately 44 times smaller than the SMC estimate. In this case, the average Hybrid algorithm runtime is approximately half that of the NumInt(τ) solution. Again, this highlights the benefits of the Hybrid algorithm in that it almost always ensures a positive likelihood estimate of a positive likelihood, and greatly reduces the variance of the estimator when compared to the SMC algorithm with the same number of particles.

The limitation of the Hybrid algorithm is that it requires the exact likelihood methods, hence the implicit Euler method, to be comparatively efficient to SMC. That is, if the implicit Euler method is infeasible or an order of magnitude larger in runtime than SMC, the benefits of the Hybrid algorithm are mostly lost by the huge increase in runtime; even if exact methods are only used for a small number of time points. As discussed in Chapter 2, the efficiency of the implicit Euler method is proportional to the size of the CTMC Q matrix, an $|\mathcal{S} \times \mathcal{S}|$ matrix. The SEIpIsIaR model has a state space, and hence Q matrix, that are too large for the implicit Euler method to be used effectively, which prohibits the use of the Hybrid algorithm for our purposes. However, the Hybrid algorithm could be used effectively for simpler models such as the SIR/SEIR models, or other models outside the field. The further applications of the Hybrid algorithm are discussed in Chapter 5.

Chapter 4

1918 Influenza Pandemic Inference

This chapter details the investigation conducted on the 1918 influenza pandemic ship data, collated from Cumpston [20]. Here, we outline the ship outbreaks and develop the *ship epidemic* model, built upon the SEIpIsIaR model with prior immunity (see Chapter 3). The ship epidemic model is employed to characterise the epidemiological dynamics of the 1918 pandemic. Parallel inference is used to investigate possible epidemiological differences between the second and third waves of the pandemic.

4.1 Ship Data Description

The 1918 pandemic ship data details 15 influenza outbreaks aboard ships travelling to Australia between October 1918 and April 1919. The documentation of the data contains (where applicable): daily resolution case counts, port arrival/departure dates, use of inoculations/quarantine measures and corresponding dates, and all landings of healthy and infected passengers. The following is a general description of the ship outbreaks. For a detailed summary of each ship, see the Appendix. For a description of how the data is used, see Section 4.2.

The ships can be divided into categories of naval troopships and passenger/trading vessels. The naval troopships were generally the largest vessels, tasked with returning ANZAC troops to Australia from locations such as England, South Africa and Egypt after the end of World War I. The passenger and trading vessels were usually smaller in size, and ferried passengers and cargo to Australia from countries such as New Zealand, Canada and Singapore. The ships' arrival in these *source-of-infection* countries coincided with active influenza epidemics. Consequently, these vessels embarked a troop or passenger carrying the influenza virus unknown to the ship command. The ships could then travel for days at sea before observing the first influenza case. Within the contained environment of the ship, the infection quickly spreads to other individuals and becomes an *on-board epidemic*.

The populations aboard these ships were especially at risk to an epidemic as they demonstrated consistent mixing of healthy and infected individuals, due to the contained environment and lack of effective control measures. Cumpston [20] consistently highlights the difficulty in isolating infected individuals from the healthy population, stating in regards to various ships:

> *'The construction of the vessel was such that it cannot be said that there was any isolation between various sections of the ships company'*,
>
> *'Senior Medical Officer stated that initial isolation was attempted but the hospital accommodation being inadequate, a portion of the troop deck had to be used for cases, and through this men had to pass to get to their quarters'*,
>
> *'Effective isolation was not possible'*.

Any other attempted control measures such as zinc-sulphate inhalation (nasopharynx disinfectant) were found to have little impact in stopping the spread of infection [20]. Hence, influenza was effectively allowed to transmit throughout the population unimpeded.

The only reprieve in slowing the spread of infection was removing infected individuals from the vessel. While travelling to Australia, the ships would call into ports to land influenza cases and resupply. Detailed records of all landed individuals, including their observed infection status, were kept, and strict containment procedures were undertaken to prevent the boarding of further infection.

In October 1918, prompted by severe influenza epidemics in New Zealand and South Africa, mandatory quarantine procedures were implemented upon entry into Australian ports via *quarantine stations* [20]. On arrival, ships were required to report all possible influenza cases, which had been documented by the on-board ship surgeons. (Note, influenza was difficult to identify so cases were diagnosed by influenza-like-illnesses.) The mandatory reporting of all influenza cases provided the daily influenza case counts aboard these vessels as recorded in Cumpston [20]. All vessels that had observed influenza cases during the voyage were ordered into quarantine. The duration of stay was determined by the time since last active influenza case. In quarantine, infected passengers would be landed to shore for observation. For the remaining passengers, new cases were removed at onset of symptoms assessed by daily thermometer parades. Quarantine officials attempted to slow the spread of further infections via zinc-sulphate inhalation and inoculations where available. The ships were detained in quarantine until new cases ceased for 7 days and it was determined safe for passengers to enter Australia. Hence, Cumpston [20] provides the daily influenza case counts for the full epidemic time line, and records of all quarantine measures implemented; a level of detail atypical of 1918 influenza case data.

The on-board epidemics are natural *pseudo experiments* of influenza transmission, demonstrating uncontrolled transmission within a contained environment. As Cumpston [20] states:

> *'An outbreak of infectious disease on a ship at sea offers the most favourable naturally occurring conditions for the study of the natural history of that disease. There can be, if the epidemic is at all extensive, no effective isolation; there can be no new factor introduced from outside and therefore the epidemic can pursue its unadulterated course, and there is reasonable accuracy in the records'.*

Hence, the collection of ship data gives an unparalleled picture of the evolution of a contained outbreak of the 1918 pandemic influenza strain. There have been numerous mathematical studies of the 1918 pandemic, but the clarity of the ship epidemic data allows an unprecedented investigation. Previous studies typically use data from citywide case/mortality records [11, 17, 23, 54], household studies [26], or institutional records such as schools or military bases [19, 48, 49, 70], to varying degrees of success. Epidemic modelling, such as in these studies, is complicated by the unknown mixing of the population and uncontrollable factors, such as externally-introduced transmission and unreliable case observation. Where these investigations falter is that they require far more complex models, deterministic approximations and unjustified assumptions to account for these unknown population dynamics. For example, for a citywide influenza epidemic there are ample external factors that need to be taken into account such as: immigration, disease introduction, case recognition, natural birth/death and complex spatial mixing dynamics that are difficult to identify and replicate. The contained nature of the ship outbreaks nullifies the need for many of these complex models/assumptions; thus greatly improving our ability to characterise the epidemiology of 1918 pandemic influenza.

Here, we aim to epidemiologically characterise the 1918 pandemic influenza strain from the ship data. In doing so, the waves of the 1918 pandemic require consideration due to possible changes in epidemiology between waves. Hence, the ship data is classified *by wave* of the pandemic, to allow parallel inference across all ships belonging to the same wave. This allows the epidemiological characterisation of the 1918 pandemic, and comparison across pandemic waves.

As discussed in Chapter 1, the 1918 influenza pandemic progressed in three distinctive waves over a 12 month period from 1918 to 1919. The exact dates of the waves are unknown and location dependent; the commonly accepted global wave time line follows the northern hemisphere seasons of approximate summer, autumn and winter waves [74]. Hence, we allocate the ships to waves based on the date of first observed infection. All ships with first cases observed between September 1918 - December 1918 are Wave 2 and between January 1919 - April 1919 are Wave 3. That is, the *(1) Niagara, (2) Mataram, (3) Devon, (4) Marathon, (5) Atua, (6) Manuka, (7) Medic, (8) Boonah,* and *(9) Nestor* outbreaks correspond to Wave 2 of the 1918 pandemic. The *(10) Ceramic, (11) Lancashire, (12) Kenilworth Castle, (13) Orca, (14) Kashmir* and *(15) Euripides* outbreaks correspond to Wave 3 of the 1918 pandemic.

See Table 4.1 for a summary of all 15 of the 1918 pandemic ship datasets and see the Appendix for a detailed description of each ship. Note, the ship area (m^2) is approximated as the ship beam (width at waterline) multiplied by the length at waterline.

	Ship Name	Wave	Source of Infection	First Case	Last Case	Case Total	Pop. Size	Ship Area (m^2)
(1)	Niagara	2	Vancouver (Canada)	27/9/18	17/10/18	156	567	3200
(2)	Mataram	2	Singapore (Singapore)	10/10/18	Unknown	61	199	1326
(3)	Devon	2	Suez (Egypt)	13/10/18	7/11/18	95	1096	2592
(4)	Marathon	2	Devonport (England)	20/10/18	2/11/18	89	1041	2618
(5)	Atua	2	Auckland (New Zealand)	3/11/18	22/11/18	91	163	1261
(6)	Manuka	2	Wellington (New Zealand)	Unknown	19/11/18	39	203	1568
(7)	Medic	2	Wellington (New Zealand)	10/11/18	20/12/18	313	989	3306
(8)	Boonah	2	Durban (South Africa)	28/11/18	7/1/19	470	1095	2484
(9)	Nestor	2	London (England)	11/12/18	25/1/19	69	1903	3300
(10)	Ceramic	3	Devonport (England)	25/1/19	10/3/19	194	2361	4200
(11)	Lancashire	3	Devonport (England)	7/2/19	14/3/19	53	1643	2600
(12)	Kenilworth Castle	3	Liverpool (England)	15/2/19	10/3/19	24	505	3306
(13)	Orca	3	Liverpool (England)	20/2/19	28/3/19	48	1698	3480
(14)	Kashmir	3	Southhampton (England)	12/3/19	23/4/19	97	1500	2628
(15)	Euripides	3	London (England)	17/3/19	15/4/19	53	1323	3340

Table 4.1: Ship data summary.

4.2 Ship Epidemic Model

An equivalent representation of the SEIpIsIaR model with prior immunity (see Chapter 3) will be fitted to the ship data, with extensions to account for the possible reduction in transmission arising from quarantine measures and the landing of healthy and infected individuals. We refer to this model as the *ship epidemic model* for simplicity. A diagram of the possible compartments an individual can be in is given in Figure 4.1.

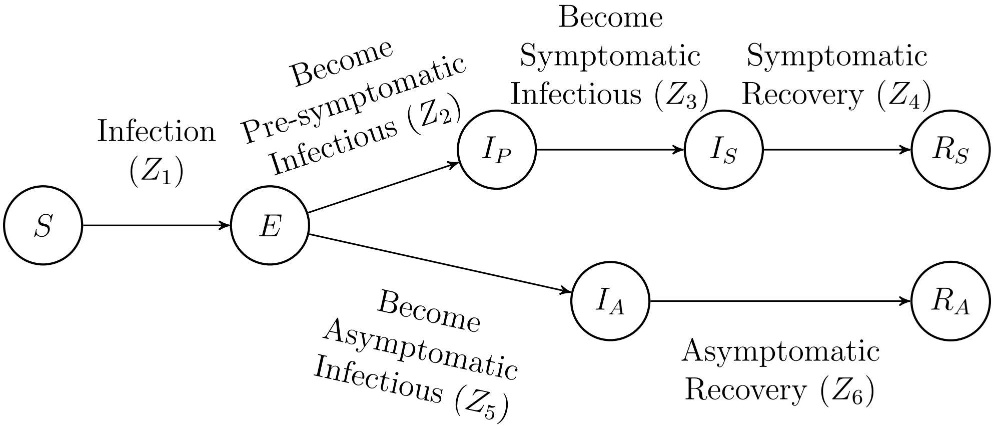

Become Symptomatic Infectious (Z_3)
Symptomatic Recovery (Z_4)
Become Pre-symptomatic Infectious (Z_2)
Infection (Z_1)
I_P
I_S
R_S
S
E
I_A
R_A
Become Asymptomatic Infectious (Z_5)
Asymptomatic Recovery (Z_6)

Figure 4.1: Compartment diagram of ship epidemic system with DoA events.

Let $Z_1(t)$ be the number of *infection* events, $Z_2(t)$ the number of *become pre-symptomatic infectious* events, $Z_3(t)$ the number of *become symptomatic infectious* events, $Z_4(t)$ the number of *symptomatic recovery* events, $Z_5(t)$ the number of *become asymptomatic infectious* events and $Z_6(t)$ the number of *asymptomatic recovery* events, by time t respectively. The population numbers are given by

$$\begin{aligned}
S(t) &= N - Z_1(t), \\
E(t) &= Z_1(t) - Z_2(t) - Z_5(t), \\
I_P(t) &= Z_2(t) - Z_3(t), \\
I_S(t) &= Z_3(t) - Z_4(t), \\
I_A(t) &= Z_5(t) - Z_6(t), \\
R_S(t) &= Z_4(t), \\
R_A(t) &= Z_6(t).
\end{aligned} \tag{4.1}$$

The ship epidemic process is expressed as a six-variable CTMC $\{(Z_1(t), Z_2(t), Z_3(t), Z_4(t), Z_5(t), Z_6(t))\}_{t\geq 0}$ with state space

$$\begin{aligned}\mathcal{S} =&\{(Z_1, Z_2, Z_3, Z_4, Z_5, Z_6)|Z_1, Z_2, Z_3, Z_4, Z_5, Z_6 \in \mathbb{N},\\ &0 \leq Z_4 + Z_6 \leq Z_2 + Z_5 \leq Z_1 \leq N, Z_4 \leq Z_3 \leq Z_2, Z_6 \leq Z_5\}.\end{aligned}$$

The transition rates from a state $(Z_1, Z_2, Z_3, Z_4, Z_5, Z_6)$ are displayed in Table 4.2. Note, the transition rates are scaled for *density-dependent* transmission using the approximate ship area, denoted A.

Transition Type	State Change	Transition Rate
Infection	$Z_1 \to Z_1 + 1$	$\frac{\beta_p(N-Z_1)(Z_2-Z_3)}{A} + \frac{\beta_s(N-Z_1)(Z_3-Z_4)}{A} + \frac{\beta_a(N-Z_1)(Z_5-Z_6)}{A}$
Become Pre-symptomatic Infectious	$Z_2 \to Z_2 + 1$	$p_s\sigma(Z_1 - Z_2 - Z_5)$
Become Symptomatic Infectious	$Z_3 \to Z_3 + 1$	$\alpha(Z_2 - Z_3)$
Symptomatic Recovery	$Z_4 \to Z_4 + 1$	$\gamma_s(Z_3 - Z_4)$
Become Asymptomatic Infectious	$Z_5 \to Z_5 + 1$	$(1 - p_s)\sigma(Z_1 - Z_2 - Z_5)$
Asymptomatic Recovery	$Z_6 \to Z_6 + 1$	$\gamma_a(Z_5 - Z_6)$

Table 4.2: Ship epidemic model events and transition rates. Note, State Change lists the variable of the process that undergoes change by a transition; all other remain unchanged.

As discussed in Chapter 3, *prior immunity* refers to the proportion of the population immune to a disease before the onset of an outbreak. The presence of prior immunity has been shown to have played an essential role in the 1918 pandemic, and that levels of immunity increased between pandemic waves [49]. Therefore, the ships' population likely exhibit prior immunity and it should be accounted for within the ship epidemic model. As in Chapter 3, prior immunity can be modelled (under the assumption that immunity lasts the duration of the outbreak) by introducing the parameter, p_{sus}, the probability a random individual within the population is susceptible to infection. Assume an initial susceptible population of size N with a single exposed individual. The number of immune individuals within the population, n_{im}, is randomly sampled from a binomial$(N-1, 1-p_{sus})$ distribution. Hence, the initial state with a single exposed individual is

$$(Z_1(0), Z_2(0), Z_3(0), Z_4(0), Z_5(0), Z_6(0)) = (1 + n_{im}, 0, 0, 0, n_{im}, n_{im}),$$

which translates to compartment numbers

$$(S(0), E(0), I_P(0), I_S(0), I_A(0), R(0)) = (N - 1 - n_{im}, 1, 0, 0, 0, 0, n_{im}).$$

We extend the model to account for the possible changes to transmission in quarantine and the landing of healthy and infected individuals. Once a ship enters quarantine, we account for the possible reduction in transmission arising from quarantine measures by introducing a *transmission reduction* parameter, $\lambda \in [0, 1]$. When a ship enters quarantine in Australia, all transmission parameters are scaled by λ. That is, under quarantine symptomatic individuals transmit at rate $\lambda\beta_s$, pre-symptomatic individuals at rate $\lambda\beta_p$, and asymptomatic individuals at rate $\lambda\beta_a$. The change in transmission could be caused by a number of quarantine measures such as reduction in contacts by restricted passenger movement or reduction in infectiousness from inoculations and nasopharynx disinfectant. Therefore, when a ship reaches Australia and enters quarantine, the model switches to the *quarantine transmission* transition rates where the infection event occurs according to Table 4.3. Note, the *Marathon* and *Nestor* have uniquely defined λ given specific population conditions. The *Marathon* used formaldehyde inhalation as a control measure, as opposed to the quarantine zinc-sulphate inhalation, and the *Nestor* has records of pre-embarking troop inoculation, which alters the effect of quarantine inoculations. See Appendix for further details.

Transition Type	State Change	Transition Rate
Infection	$Z_1 \to Z_1 + 1$	$\frac{\lambda\beta_p(N - Z_1)(Z_2 - Z_3)}{A} + \frac{\lambda\beta_s(N - Z_1)(Z_3 - Z_4)}{A} + \frac{\lambda\beta_a(N - Z_1)(Z_5 - Z_6)}{A}$

Table 4.3: Quarantine transmission infection transition rate. Note, State Change lists the variable of the process that undergoes change by a transition; all other remain unchanged.

While in an international port or Australian quarantine, the ships may land passengers from the ship to shore. Importantly, records of the dates and number of landed passengers alongside their infection status are documented in Cumpston [20]. All individuals landed are denoted as either *healthy*, not observed to have contracted influenza, or *infected*, observed to have contracted influenza. Using this information, we model the changes in population by removing individuals from the corresponding compartments. That is, if a healthy individual is removed from the ship we randomly sample an individual to remove from compartments corresponding to passengers who are not infected or have unobserved infections, that is any of the S, E, I_P, I_A or R_A compartments. For an infected individual, we randomly sample an individual who has had an observed symptomatic infection, from either the

I_S or R_S compartments. The compartment diagram of the ship epidemic system with possible removals is given in Figure 4.2. Note, for simplicity we use the terms *healthy* and *infected* as denoted in Cumpston [20], but these terms are observation-dependent and can be properly interpreted as "apparently healthy", and "symptomatic infection (past or present)".

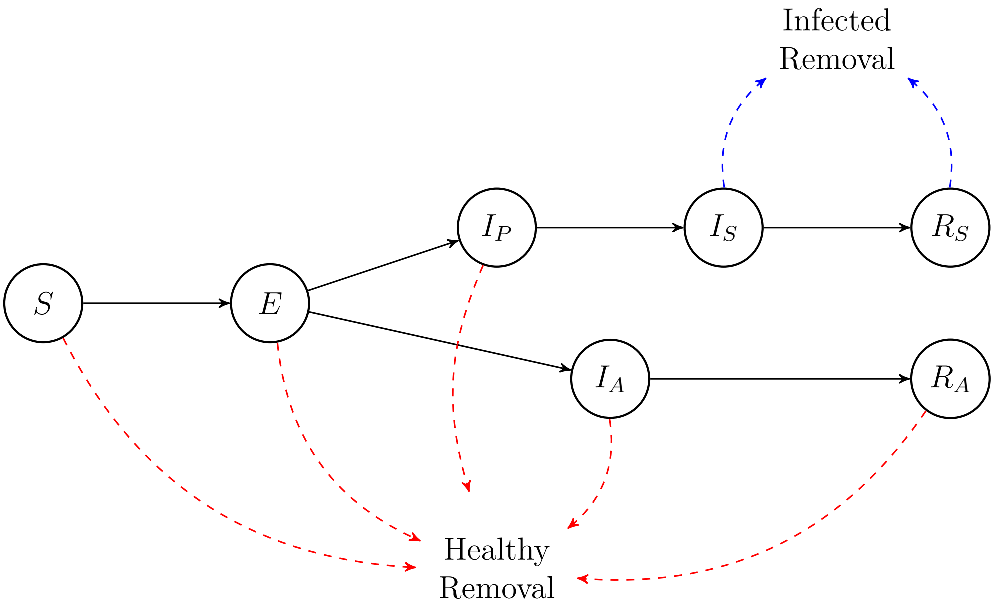


Figure 4.2: Compartment diagram of ship epidemic system with possible removals. Blue signifies possible infected individuals removals. Red signifies possible healthy individuals removals.

The following assumptions are made within the model. Firstly, we assume a closed population with no births, natural deaths, unregistered migrations or external transmission. This is valid due to the contained environment of the ship and the ship surgeon records presented in Cumpston [20]. The population is restricted to the ship for the duration of the outbreak, removing external transmission. Records show no births or natural deaths occur on the vessel, and all passenger landings are modelled appropriately, as already defined.

Secondly, we assume a homogeneously-mixing population. Clearly, as in most epidemic modelling pursuits, the assumption of a homogeneous population is an approximation to the heterogeneous reality. Here we argue this approximation is much closer to the true dynamics than commonly accepted in mathematical studies. This homogeneous assumption is complex as it is dependent on the structure of the ship, how the crew interact and how the ship officials attempt to isolate cases. Hence, we propose an argument to validate the approximation: that all healthy crew/passengers mix roughly uniformly and infected individuals continue interaction with the healthy population.

As discussed in Section 4.1, effective isolation of any infected individuals was near impossible aboard these ships due to the structure of the vessels and the number of cases on board. Hence, we can assume infected individuals continue to have consistent interaction with the remaining population. Furthermore, we note that the crew/passengers aboard the ships were attacked in roughly equal proportion, as Cumpston [20] states,

> *'Where there is such a large number of troops that differentiation between troops and crew is hardly practicable, the troops and crew became attacked in more equal proportions.'*

The proportional attack rates suggest that there is uniform-mixing between the healthy passengers and crew due to the close contact required for influenza transmission [77]. Therefore, we allow the assumption of homogeneous mixing.

Thirdly, we assume perfect observation of all symptomatic individuals upon immediate development of symptoms. We take this to be valid due to the ship surgeon mandate to provide detailed records of all potential influenza cases upon entry to Australia, as provided in Cumpston [20]. Hence, we observe an individual upon entry to the symptomatic infectious compartment and define the observation process as

$$P(y_t|(Z_1, Z_2, Z_3, Z_4, Z_5, Z_5, Z_6)) = \begin{cases} 1 & \text{if } y_t = Z_3, \\ 0 & \text{else.} \end{cases} \tag{4.2}$$

Lastly, we assume a single exposed individual is the only infected individual to board the ship prior to the outbreak (the remaining population is either susceptible or immune; see above) and they embark during stay at the source-of-infection city. While it is possible for more than one infected individual to have come aboard, we make this assumption for simplicity. Here, we allow the exposed individual to embark and begin transmitting at any time during the stay in port to account for the possibility of multiple infections prior to the on-board outbreak. Therefore, approximating the effect of multiple infected individuals embarking before departure. This process is described below.

We outline the development of a typical outbreak and how extensions of the model are used. While in port at the source-of-infection city, a single exposed individual enters the otherwise susceptible population with n_{im} immune individuals. An unknown (but bounded) time later the first case is observed at the end of Day 1 of the outbreak. To determine the population compartment numbers at the start of the outbreak, we condition on the time of first observed case given the process started with a single exposed individual. That is, let T_{arrival}, $T_{\text{departure}}$ and T_{first} be the date of arrival to the source-of-infection city, the date of departure and the date of first observed case. The model initiates in state

$$(Z_1(0), Z_2(0), Z_3(0), Z_4(0), Z_5(0), Z_6(0)) = (1 + n_{im}, 0, 0, 0, n_{im}, n_{im}),$$

and the process conditions on observing the first case, y_1, only within the bounded time interval $t \in [T_{\text{first}} - T_{\text{arrival}}, T_{\text{first}} - T_{\text{departure}}]$ such that,

$$\{(Z_1(t), Z_2(t), Z_3(t), Z_4(t), Z_5(t), Z_6(t)) | Z_3(t) = y_1\}.$$

For example, the *Boonah* ship brought aboard the influenza virus during its stay in Durban between November 16th and 24th, and the first case was observed 5 days after leaving port on November 29th. Hence, the model initiates with a single exposed individual and conditions on observing the first observed case at least 5 days later but no longer than 13 days. Therefore, effectively inferring the population compartment numbers at the time of the first observed case.

The ship continues on its journey and the epidemic progresses according to the standard transition rates, with new cases recorded at daily resolution. Once the ship has arrived into Australia, the process switches to the quarantine transmission transition rates as given in Table 4.3. Upon arrival, a number of recorded healthy or infected individuals are landed into the quarantine station. These individuals are uniformly sampled for removal by their infection status, sampled from the corresponding compartments as depicted in Figure 4.2. After all observed cases are landed, the ship enters the period of *cases removed at onset*. In this stage all cases are removed immediately upon development of symptoms. That is, all individuals are removed immediately upon entry to the symptomatic infectious compartment. Over the next few days, the number of daily new cases reduces as the epidemic dies out, until the population aboard the ship are cleared of quarantine and allowed free passage.

Below, we detail how the ship epidemic model is used within the SMC algorithm. Let $D^{(k)}$ denote the complete dataset corresponding to ship outbreak k. Each dataset $D^{(k)}$ will contain the following information, where applicable to the ship outbreak. Let $\mathbf{y}_{1:T} = \{y_1, ..., y_T\}$ denote the data set of the *cumulative number of cases observed* by the end of day $t = 1, ..., T$ for the duration of the outbreak. Let $\mathbf{h}_{1:T} = \{h_1, ..., h_T\}$ and $\mathbf{c}_{1:T} = \{c_1, ..., c_T\}$ be the number of healthy and infected individuals respectively landed on day $t = 1, ..., T$. Let T_{arrival}, $T_{\text{departure}}$, T_{first}, $T_{\text{quarantine}}$ and T_{onset} be the day of arrival and departure from the source-of-infection city, the day of first case, the day of arrival to quarantine in Australia and the day of cases being landed at onset of symptoms, respectively. The SMC(N_p) algorithm for the ship epidemic model is shown in Algorithm 6. Note, particle history in the SMC algorithm is omitted as it is not required here.

```
Data: Ship dataset $D^{(k)}$, number of particles $N_p$.
for $i = 1, ..., N_p$ do
    Sample $n_{im} \sim \text{binomial}(N-1, 1-p_{sus})$;
    Initialize particle with single exposed individual,
    $x_0^{(i)} = (Z_1, Z_2, Z_3, Z_4, Z_5, Z_6)_0^{(i)} = (1+n_{im}, 0, 0, 0, n_{im}, n_{im})$;
    Simulate using standard transmission model $x_1^{*(i)} \sim p(x_1|x_0^{(i)})$ over
    time $[0, T_{\text{first}} - T_{\text{departure}}]$;
    Sample $T_{\text{sample}} \sim \{s \in [0, T_{\text{first}} - T_{\text{departure}}] | y_t = (Z_3)_s^{(i)}\}$;
    Assign weight $W_1^{(i)} = P(y_1|x_1^{*(i)})$ where

    $$P(y_1|x_1^{*(i)}) = \begin{cases} 1 & \text{if } y_t = (Z_3)_1^{(i)} \text{ \& } T_{\text{sample}} \in [T_{\text{first}} - T_{\text{arrival}}, T_{\text{first}} - T_{\text{departure}}], \\ 0 & \text{else;} \end{cases}$$

    Sample and remove $h_1$ individuals from the $S$, $E$, $I_P$, $I_A$ and $R_A$
    compartments;
    Sample and remove $c_1$ individuals from the $I_S$ and $R_S$
    compartments;
end
Calculate likelihood $\hat{P}(y_1) = \frac{1}{N_p} \sum_{i=1}^{N_p} W_1^{(i)}$;
Re-sample $\{x_1^{(1)}, ..., x_1^{(N_P)}\}$ with weights $P(x_1 = x_1^{*(j)}) = \frac{W_1^j}{\sum_k W_1^k}$ ;
for  $t = 2, ..., T$ do
    Select appropriate transmission model based on $T_{\text{quarantine}}$, $T_{\text{onset}}$;
    for $i = 1, ..., N_p$ do
        Simulate $x_t^{*(i)} \sim p(x_t|x_{t-1}^{(i)})$;
        Assign weight $W_t^{(i)} = P(y_t|x_t^{*(i)})$ where

        $$P(y_t|x_t^{*(i)}) = \begin{cases} 1 & \text{if } y_t = (Z_3)_t^{(i)}, \\ 0 & \text{else;} \end{cases}$$

        Sample and remove $h_t$ individuals from the $S$, $E$, $I_P$, $I_A$ and
        $R_A$ compartments;
        Sample and remove $c_t$ individuals from the $I_S$ and $R_S$
        compartments;
    end
    Calculate likelihood $\hat{P}(y_t|\mathbf{y}_{1:t-1}) = \frac{1}{N_p} \sum_{i=1}^{N_p} W_t^{(i)}$;
    Re-sample $\{x_t^{(1)}, ..., x_t^{(N_P)}\}$ with weights $P(x_{t+1} = x_{t+1}^{*(j)}) = \frac{W_{t+1}^j}{\sum_k W_{t+1}^k}$ ;
end
Calculate marginal likelihood $\hat{P}(\mathbf{y}_{1:T}) = \hat{P}(y_1) \prod_{t=2}^{T} \hat{P}(y_t|\mathbf{y}_{1:t-1})$ ;
Result: Marginal likelihood estimate $\hat{P}(\mathbf{y}_{1:T})$.
```

Algorithm 6: Ship epidemic model SMC(N_p) algorithm.

4.3 Parallel Inference

In this section, we epidemiologically characterise the 1918 influenza pandemic ship data by conducting parameter inference on the ship epidemic model. Here, we use parallel inference to identify the parameters of the ship epidemic model from multiple ship outbreaks, classified by pandemic wave. As detailed in Chapter 3, parallel inference refers to conducting inference on multiple datasets under the assumption they are realisations of the same model with a common set (or subset) of parameters.

We assume each outbreak is a realisation of the ship epidemic model and there is a subset of parameters that are common across outbreaks. We assume the following disease-dependent parameters to be common across outbreaks. These include: the lengths of latent ($1/\sigma$), pre-symptomatic ($1/\alpha$), symptomatic infectious ($1/\gamma_s$) and asymptomatic infectious ($1/\gamma_a$) periods, the probability of symptomatic infection (p_s), and the proportion of transmission from symptomatic, pre-symptomatic and asymptomatic individuals (κ_S, κ_P, κ_A). There are also parameters which are situation dependent and require unique consideration. We assume a unique R_0 for each ship, as the β term that determines R_0 is directly related to the mixing and contact rates aboard each ship, which we cannot assume to be common. We assume the transmission reduction parameter λ is common across outbreaks, except in the case of unique quarantine measures. As each ship is subject to the same quarantine measures upon entry to Australia, we assume the relative reduction factor is constant. Hence, we assume λ is shared across all ships upon entry to Australia. There are two special cases in the *Marathon* and *Nestor* which require unique λ values due to noted different quarantine measures implemented than those in Australian quarantine.

We assign the model parameters the following uninformative prior distributions. They are chosen to allow the data to be dominant in determining the posterior distribution, and therefore best capture the on-board dynamics. All relatively unknown parameters are assigned deliberatively uninformative prior distributions. R_0 is assigned a uniform$(0.1, 20)$ prior distribution, λ and p_{sus} are assigned uniform$(0.01, 1)$ prior distributions, κ_S and κ_P are assigned uniform$(0.01, 0.99)$ prior distribution (where $\kappa_S + \kappa_P < 1$). The epidemiological period lengths such as $1/\sigma$ and $1/\gamma$ have been studied previously and we use these past studies to inform our prior distributions [37, 49, 54, 77]. We express the prior knowledge as truncated normal and gamma distributions centred around past mean values and truncated to realistic values. We use relatively uninformative prior variances to allow the likelihood to best inform the posterior distribution. We assign $1/\sigma$ and $1/\alpha$ gamma(2,1) prior distributions and $1/\gamma_s$ and $1/\gamma_a$ gamma(2,2) prior distributions, all truncated over $(0, 8)$. We assigned p_s a normal(0.66,0.33^2) prior distribution truncated over $(0.01, 0.99)$. The prior distribution probability density functions are shown in Figure 4.3.

The posterior distribution samples are generated as follows. The PMMH algorithm is run using a SMC($N_p = 25,000$) estimate of the likelihood, as detailed in Algorithm 6. The number of particles N_p has increased from $10,000$ as in Chapter 3 for increased accuracy within parallel inference. We use an MCMC scheme as inspired by the *adaptive Metropolis algorithm*, to improve mixing of the high dimensional MH chain [5, 32]. Exploratory *burn-in chains* are run from randomly sampled initial values to assess convergence and inform a multivariate normal (MVN) proposal distribution. Here, burn-in chains are started from 10 initial values, randomly sampled from the prior distribution; these chains are run until roughly converged and stopped after 20,000 iterations. From the post-convergence samples, 20 randomly sampled initial values are taken to start 20 new chains. The new chains use a MVN proposal distribution, where the covariance matrix is chosen as the scaled sample covariance of the post-convergence samples. As MVN truncation is computationally difficult at high dimensions [27], we augment the likelihood function such that for any parameter sets outside the support of the prior distribution, the likelihood is zero. The 20 chains are given an additional 50,000 iterations of burn-in, and then any further iterations are taken as samples from the posterior distribution. Each chain is run until the posterior sample mean satisfies $\widehat{\mathrm{Var}}(\bar{\theta}) < 0.01$.

Here, we study the ship outbreaks in subsets defined by pandemic wave. As discussed in Section 4.1, we have segregated each ship outbreak corresponding to the second and third waves of the 1918 pandemic. We use these data subsets for parallel inference using the above scheme to classify the epidemiological characteristics of the 1918 pandemic influenza strain, and any epidemiological differences between the second and third wave.

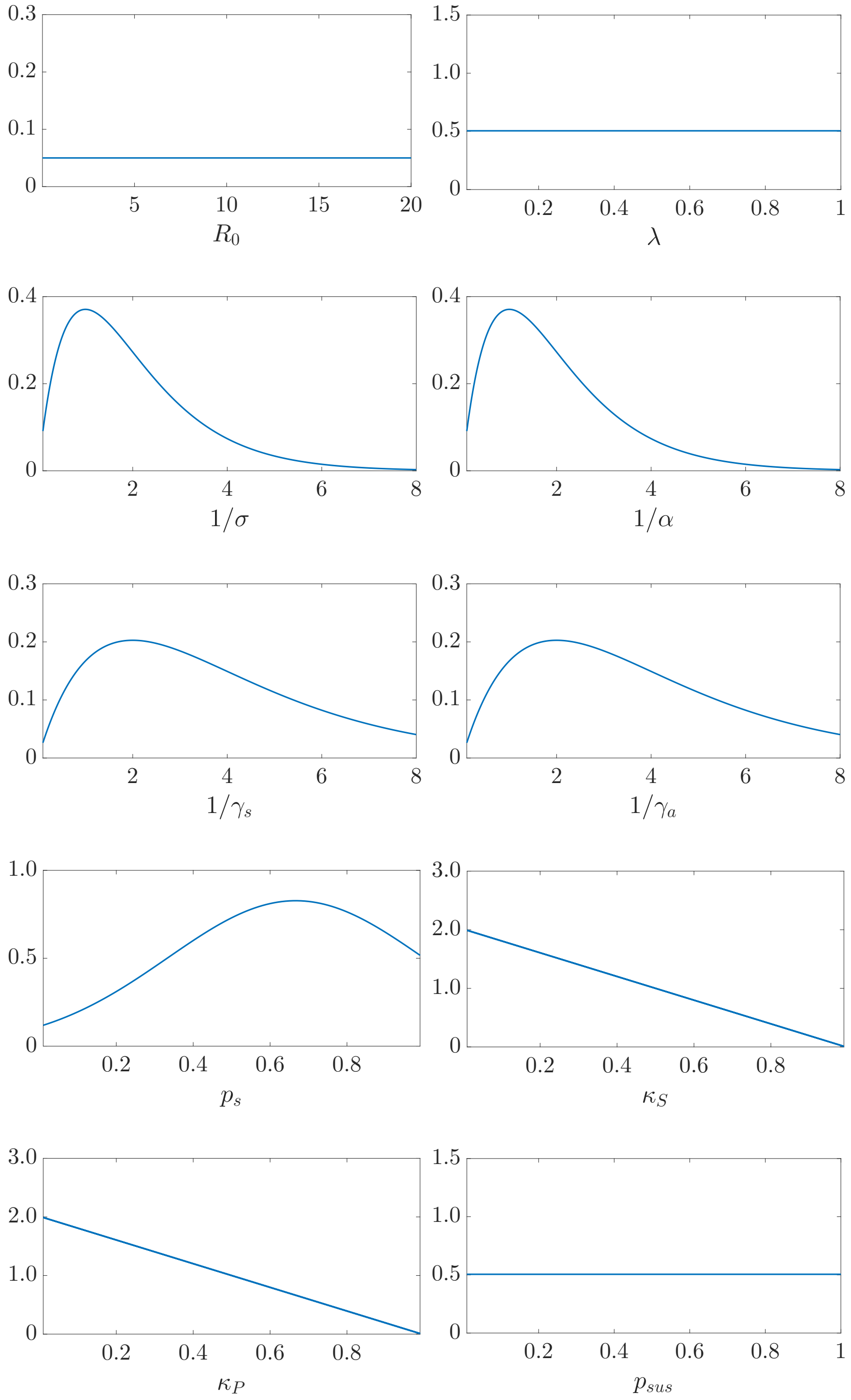


Figure 4.3: Ship epidemic model prior marginal probability density functions.

4.3.1 Wave 2

In this section we conduct parallel inference on all Wave 2 ships: *(1) Niagara, (2) Mataram, (3) Devon, (4) Marathon, (5) Atua, (6) Manuka, (7) Medic, (8) Boonah,* and *(9) Nestor.* Hence, we target the posterior distribution,

$$p(\theta|\{D^{(1)}, ..., D^{(9)}\}) \propto p(\{D^{(1)}, ..., D^{(9)}\}|\theta)p(\theta).$$

We conduct inference on the following parameter transformations

$$\theta = \left\{R_0^{(1)}, ..., \ R_0^{(9)}, \ \lambda, \ \lambda^{(4)}, \ \lambda^{(9)}, \ \frac{1}{\sigma}, \ \frac{1}{\alpha}, \ \frac{1}{\gamma_s}, \ \frac{1}{\gamma_a}, \ p_s, \ \kappa_S, \ \kappa_P, \ p_{sus}\right\}.$$

Here, $R_0^{(1)}, ..., \ R_0^{(9)}$ denote the unique R_0 values for each ship, λ denotes the standard transmission reduction parameter where $\lambda^{(4)}$ and $\lambda^{(9)}$ denote the unique parameters for the *Marathon* and *Nestor* ships. We use the prior distributions and MCMC scheme denoted above to generate 3,500,000 posterior samples for analysis. The mean and 95% HPD intervals of the posterior samples are given in Table 4.4. The marginal kernel density estimates of the posterior samples are shown in Figure 4.4. The trace plots of the MH chain are shown in Figure 4.5. (Note, the kernel density estimates presented within chapter are calculated using the algorithm presented in Botev *et al.* [12].)

Parameter	Mean (95% HPD)
$R_0^{(1)}$	2.01 (0.61, 4.73)
$R_0^{(2)}$	7.43 (1.57, 14.3)
$R_0^{(3)}$	1.33 (0.59, 2.21)
$R_0^{(4)}$	4.75 (0.66, 8.57)
$R_0^{(5)}$	9.54 (1.80, 18.3)
$R_0^{(6)}$	6.25 (1.30, 11.3)
$R_0^{(7)}$	4.23 (1.45, 7.44)
$R_0^{(8)}$	3.79 (1.70, 6.64)
$R_0^{(9)}$	2.78 (0.96, 4.50)
λ	0.42 (0.01, 0.86)

Parameter	Mean (95% HPD)
$\lambda^{(4)}$	0.12 (0.01, 0.40)
$\lambda^{(9)}$	0.72 (0.31, 1.00)
$\frac{1}{\sigma}$	0.96 (0.10, 2.16)
$\frac{1}{\alpha}$	0.83 (0.10, 2.22)
$\frac{1}{\gamma_s}$	4.86 (2.13, 7.89)
$\frac{1}{\gamma_a}$	3.53 (0.18, 7.01)
p_s	0.97 (0.95, 0.99)
κ_S	0.14 (0.01, 0.27)
κ_P	0.25 (0.08, 0.45)
p_{sus}	0.55 (0.44, 0.66)

Table 4.4: Wave 2 ships posterior point estimates. Mean and 95% HPD intervals of posterior samples.

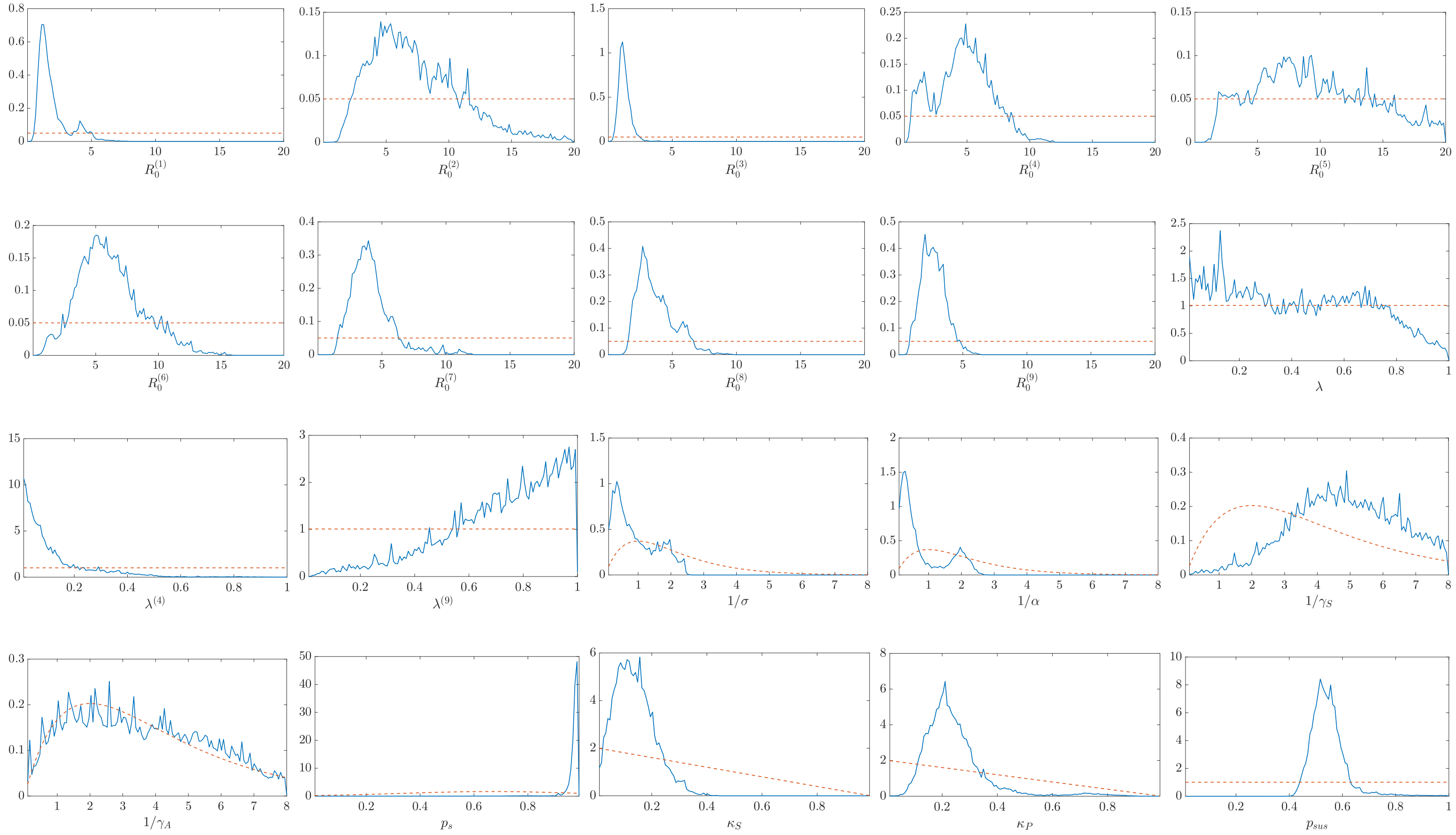


Figure 4.4: Wave 2 ships posterior marginal kernel density estimates. Red line denotes prior probability density function.

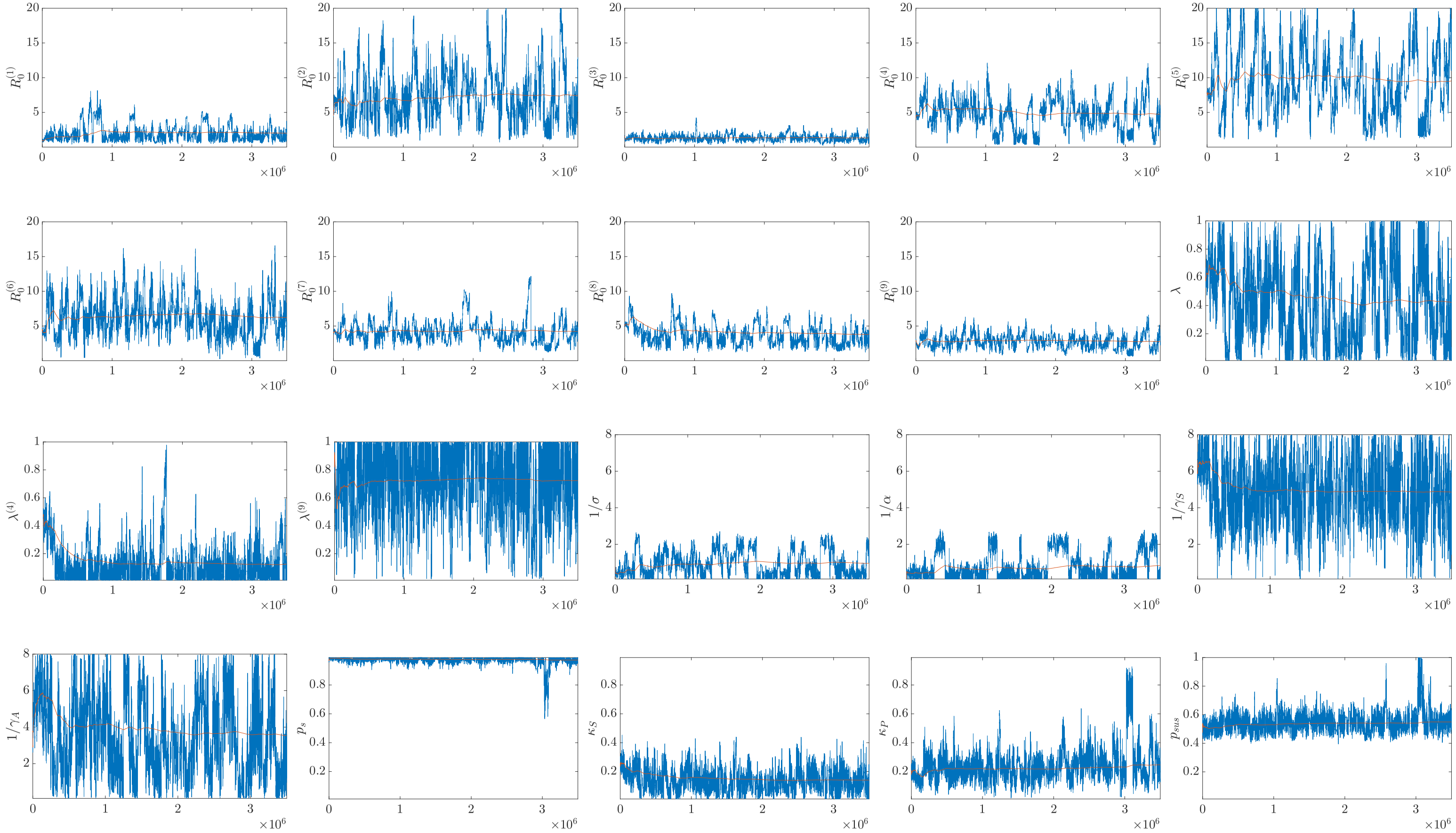


Figure 4.5: Wave 2 ships posterior trace plots by parameter. Red line denotes the cumulative mean.

The posterior point estimates highlight some interesting results. We estimate all ships had an $R_0 > 1$, as expected by the threshold theorem, and 6 out of the 9 ships had an $R_0 > 3$. We note some marginal R_0 distributions are heavily right-tailed with 97.5% percentiles as high as 18. This indicates that once above a certain threshold for each ship, R_0 is less influential in fitting to the data; possibly due to interaction with other parameters such as prior immunity. These R_0 values are high for typical pandemic influenza, but are consistent with previous studies of the 1918 pandemic; particularly, outbreaks within contained populations, as expected due to higher levels of mixing compared to an open population [25, 49, 54]. The high R_0 values help explain the high clinical attack rate observed aboard the Wave 2 ships.

The λ estimate indicates an effective reduction in transmission rate of 58% after the ships had entered Australian quarantine. The *Nestor*, which had pre-inoculated troops on board, had a lesser reduction of approximately 28%. Although, there is a large degree of uncertainty in these estimates. The ship's entry into quarantine often coincided with the natural decline of the epidemic, and so the exact responsibilities for the reduction in transmission are unclear. In comparison, the effective reduction in transmission aboard the *Marathon*, which used formaldehyde inhalation mid-outbreak at sea, was approximately 88%. Hence, formaldehyde inhalation appears to have had a far greater effect on reducing transmission aboard the *Marathon* than the Australian quarantine measures.

The epidemiological periods are estimated to be an average 23 hour latent period and 20 hour pre-symptomatic period; resulting in an average incubation period of 43 hours. The average infectious periods are estimated to be: symptomatic infectious period of 5 days, and asymptomatic infectious period of 3.5 days. These values are in line with previous studies [48, 54, 77]. The estimated latent and pre-symptomatic periods suggest that symptomatic individuals were infectious for approximately half of the incubation period. This is a novel estimate of the duration of the pre-symptomatic period, which is often ignored in epidemic modelling.

The generation time can be estimated through simulation, in which a primary exposed individual is introduced into the population and the time until their first transmission is recorded. Denote the generation time aboard ship i, $T_g^{(i)}$. Here, we estimate $T_g^{(i)}$ from 10,000 posterior simulations, where each simulation uses a randomly sampled parameter set from the posterior samples; the estimated mean and 95% HPD intervals are shown in Table 4.5. The combined T_g kernel density estimate is shown in Figure 4.6. Here, we see on average a generation time of 53 hours, which is slightly shorter than previous estimates [48]. The short generation time can explain the fast epidemic growth in number of observed cases aboard the ships; a result of the high R_0 values and short latency period.

Ship	Mean (95% HPD)
$T_g^{(1)}$	2.51 (0.00, 8.36)
$T_g^{(2)}$	1.86 (0.00, 5.98)
$T_g^{(3)}$	2.79 (0.01, 9.46)
$T_g^{(4)}$	2.06 (0.00, 6.67)
$T_g^{(5)}$	1.73 (0.00, 5.60)
$T_g^{(6)}$	1.93 (0.01, 6.27)
$T_g^{(7)}$	2.19 (0.01, 7.31)
$T_g^{(8)}$	2.24 (0.00, 7.42)
$T_g^{(9)}$	2.46 (0.01, 8.02)

Table 4.5: Wave 2 ships posterior T_g estimates by ship. Mean and 95% HPD intervals estimated from 10,000 simulations.

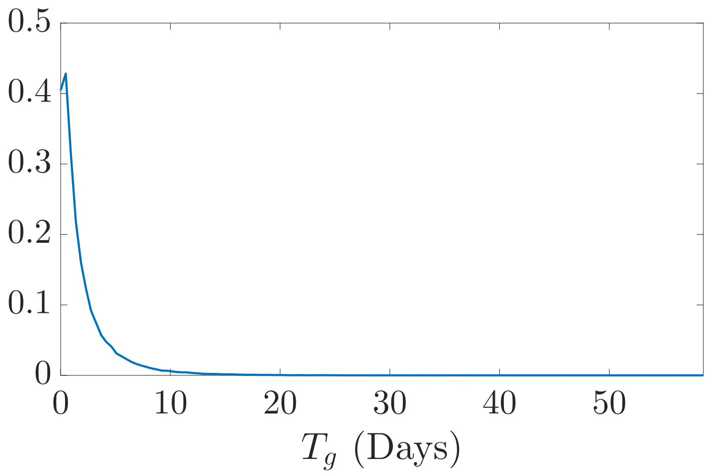


Figure 4.6: Wave 2 ships posterior T_g kernel density estimate.

We also find approximately 97% of all infections are symptomatic. This p_s value is somewhat high compared to previous studies, where estimates vary between 30 - 90% [48]. The high symptomatic proportion could be explained by the well known high virulence of the 1918 pandemic and high case recognition rates by the on-board surgeon. We estimate 45% of the population was immune prior to the Wave 2 outbreaks. Given previous estimates of the immune proportion prior to the first wave of the 1918 pandemic of 30-50%, this estimate is reasonable [49]. Hence, this result supports the conjecture that a significant proportion of the population was immune prior to the onset of the second wave of the 1918 pandemic; whether gained from the first wave or cross immunity of a past influenza strain, it is unclear. Interestingly, the combination of almost-one p_s estimate and a lower p_{sus} that appears to fit to the data's largest clinical attack rate (*Atua*; 56% clinical attack rate), suggests that almost all susceptible individuals become infected and we observed almost all of them as clinical illnesses.

We can demonstrate this elevated total attack rate by using simulation. Here, we have run 10,000 posterior simulations of the ship epidemic model, where each simulation uses a randomly sampled parameter set from the posterior samples. The simulated cumulative case counts are shown against the observed ship data in Figure 4.7. The comparison of data clinical attack rate to posterior simulation clinical and total attack rates is shown in Table 4.6. Encouragingly, we see consistent agreement with the ship data, which indicates good model fit. We note the *Devon* simulations deviate significantly from the observed data, but attribute this to potential error in the data; see the Appendix for further details. We see almost zero difference between the simulation clinical and total attack rates, signifying almost all infections are observed. We also note the larger outbreaks such as the *Atua*, *Medic* and *Boonah* tend toward complete infection of the susceptible population. The smaller populations are far more variable and have a higher probability of epidemic die-out.

	Ship Name	Data Attack Rate	Clinical Attack Rate Mean (95% HPD)	Total Attack Rate Mean (95% HPD)
(1)	Niagara	0.28	0.37 (0.05, 0.56)	0.39 (0.05, 0.65)
(2)	Mataram	0.31	0.36 (0.02, 0.59)	0.37 (0.02, 0.61)
(3)	Devon	0.09	0.02 (0.00, 0.18)	0.03 (0.00, 0.18)
(4)	Marathon	0.10	0.09 (0.00, 0.49)	0.09 (0.00, 0.50)
(5)	Atua	0.56	0.55 (0.43, 0.66)	0.57 (0.43, 0.68)
(6)	Manuka	0.19	0.34 (0.15, 0.53)	0.35 (0.16, 0.56)
(7)	Medic	0.30	0.46 (0.31, 0.58)	0.48 (0.32, 0.60)
(8)	Boonah	0.42	0.47 (0.37, 0.56)	0.49 (0.38, 0.59)
(9)	Nestor	0.04	0.07 (0.00, 0.50)	0.07 (0.00, 0.51)

Table 4.6: Wave 2 ship data versus posterior simulation attack rates. Mean and 95% HPD intervals estimated from 10,000 simulations. Posterior simulations show in Figure 4.7.

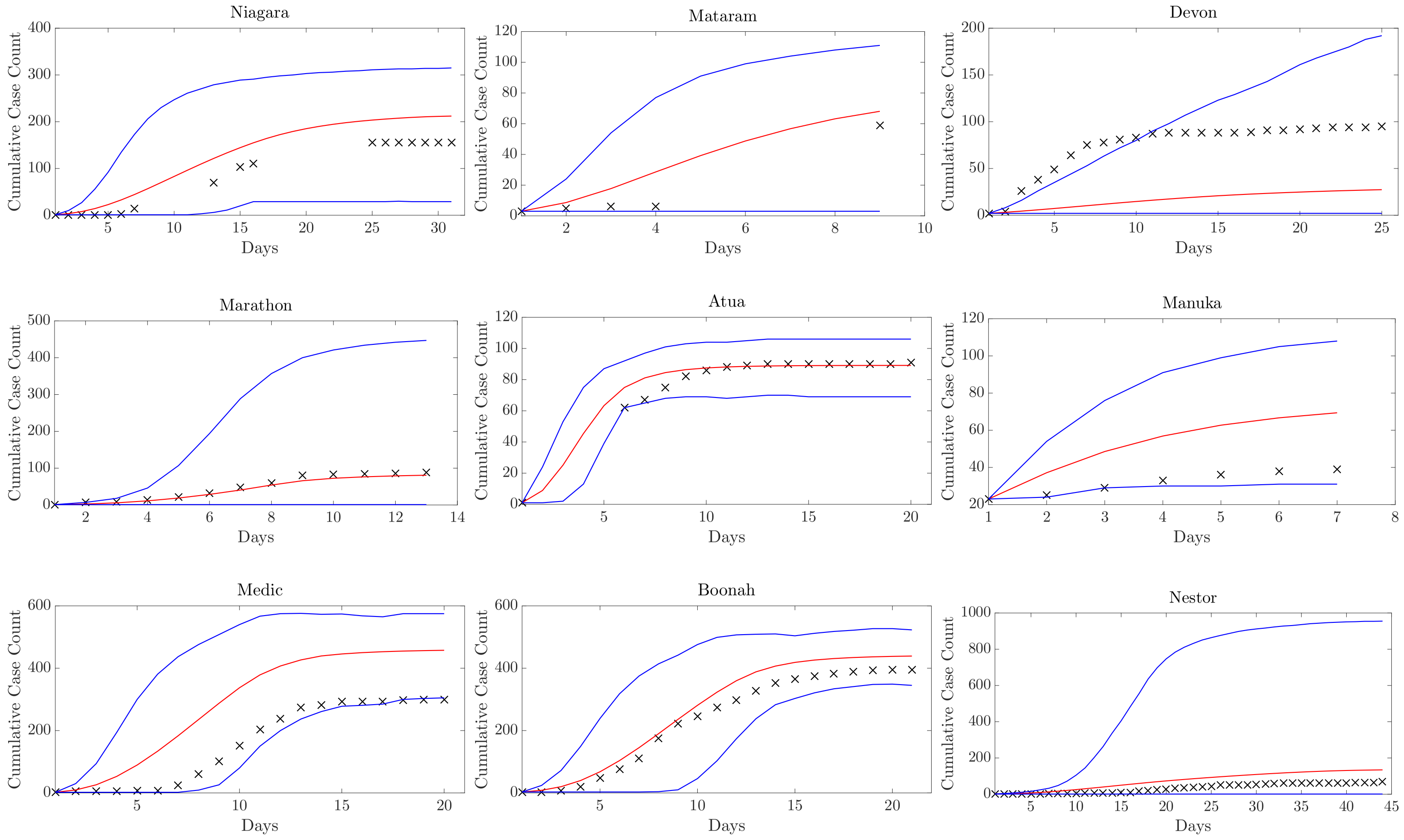


Figure 4.7: Wave 2 ship data versus posterior simulations. 10,000 ship epidemic model simulated datasets are generated using posterior sampled parameter sets. Black crosses denotes the observed cases. Red denotes the posterior simulation mean. Blue denotes the 95% HPD interval.

The other key features of the 1918 pandemic we wished to identify were the proportion of transmission arising from symptomatic (κ_S), pre-symptomatic (κ_P), asymptomatic (κ_A) and non-symptomatic (κ_{NS}) individuals. The κ posterior mean and 95% HPD intervals are shown in Table 4.7. We plot the κ kernel density estimates in Figure 4.8. We estimate the symptomatic, pre-symptomatic and asymptomatic mean transmission proportions as 0.14, 0.25 and 0.61 respectively. Hence, we estimate the non-symptomatic mean transmission proportion as 0.86. Symptomatic individuals have the lowest contribution to the spread of the disease and non-symptomatic individuals account for roughly 85% of transmission. This result illustrates an important factor in the 1918 pandemic, in that the majority of transmission arose from asymptomatic infections or prior to the onset of symptoms.

Parameter	Mean (95% HPD)
κ_S	0.14 (0.01, 0.27)
κ_P	0.25 (0.08, 0.45)
κ_A	0.61 (0.39, 0.80)
κ_{NS}	0.86 (0.73, 0.99)

Table 4.7: Wave 2 ships posterior κ estimates. Mean and 95% HPD intervals of posterior samples.

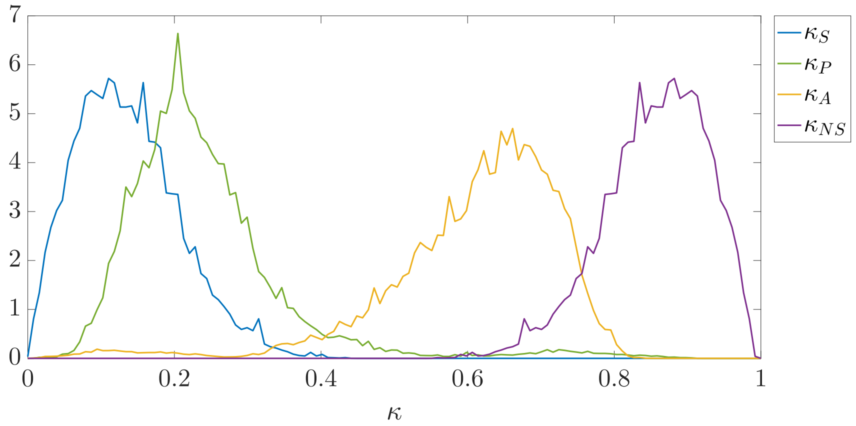


Figure 4.8: Wave 2 ship posterior κ kernel density estimates.

We must note the expected number of infections that arise from a symptomatic, pre-symptomatic or asymptomatic individual. The E[infections] posterior mean and 95% HPD intervals are shown by ship in Table 4.8. The combined E[infections] kernel density estimates are shown in Figure 4.9. Here, we see a average expected infections from symptomatic, presymptomatic and asymptomatic individuals of 0.64, 1.09, and 144, respectively. We observe difficulty in identifying $E[I_A$ infections], which allows the possibility of asymptomatic individuals to be *super-spreader* individuals that can cause a huge number of infections. Although, due to the almost-one p_s values, there is on average only five asymptomatic individuals in the duration of the epidemics.

	Ship Name	$E[I_S$ infections] Mean (95% HPD)	$E[I_P$ infections] Mean (95% HPD)	$E[I_A$ infections] Mean (95% HPD)
(1)	Niagara	0.26 (0.02, 0.60)	0.50 (0.07, 1.24)	68.1 (0.14, 214)
(2)	Mataram	1.02 (0.04, 2.32)	1.73 (0.36, 3.36)	240 (0.09, 559)
(3)	Devon	0.18 (0.02, 0.35)	0.33 (0.08, 0.63)	42.7 (0.22, 87.2)
(4)	Marathon	0.68 (0.01, 1.47)	1.10 (0.09, 1.78)	159 (0.24, 343)
(5)	Atua	1.30 (0.14, 3.12)	2.26 (0.36, 4.63)	299 (0.81, 697)
(6)	Manuka	0.85 (0.07, 1.84)	1.42 (0.49, 2.38)	202 (0.12, 429)
(7)	Medic	0.58 (0.05, 1.13)	1.00 (0.31, 1.91)	138 (0.46, 289)
(8)	Boonah	0.54 (0.03, 1.19)	0.91 (0.25, 1.68)	126 (0.19, 300)
(9)	Nestor	0.38 (0.04, 0.70)	0.65 (0.25, 1.06)	92.2 (0.1, 198)

Table 4.8: Wave 2 ships posterior E[infections] estimates by ship. Mean and 95% HPD intervals of posterior samples.

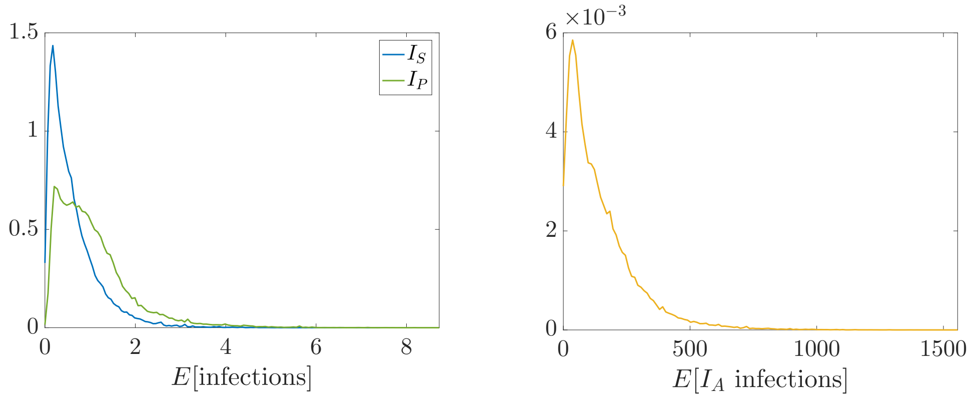


Figure 4.9: Wave 2 ships posterior E[infections] kernel density estimates.

In summary, we have characterised the epidemiological dynamics occurring aboard 9 ships during the second wave of the 1918 pandemic. We found the following statistics of note. R_0 values aboard these ships are in the range of 1.3 - 9.5. The quarantine measures implemented upon entry to Australian quarantine stations resulted in a 58% reduction in transmission, where as the formaldehyde inhalation conducted on the *Marathon* resulted in a 88% reduction. Approximately 97% of all infections were symptomatic and these individuals were infectious on average 20 hours before development of symptoms; that is, for approximately half the incubation period. Only 55% of the population was susceptible prior to onset of outbreak in the second wave. Non-symptomatic individuals (asymptomatic or prior to onset of symptoms) were responsible for approximately 85% of transmission. In particular, there were only a handful of asymptomatic individuals per outbreak, but they had the potential to be extremely infectious to others.

4.3.2 Wave 3

In this section we conduct parallel inference on all Wave 3 ships: *(10) Ceramic, (11) Lancashire, (12) Kenilworth Castle, (13) Orca, (14) Kashmir* and *(15) Euripides.* Hence, we target the posterior distribution,

$$p(\theta|\{D^{(10)}, ..., D^{(15)}\}) \propto p(\{D^{(10)}, ..., D^{(15)}\}|\theta)p(\theta).$$

We conduct inference on the following parameter transformations

$$\theta = \left\{ R_0^{(10)}, ..., \ R_0^{(15)}, \ \lambda, \ \frac{1}{\sigma}, \ \frac{1}{\alpha}, \ \frac{1}{\gamma_s}, \ \frac{1}{\gamma_a}, \ p_s, \ \kappa_S, \ \kappa_P, \ p_{sus} \right\}.$$

Here, $R_0^{(10)}, ..., \ R_0^{(15)}$ denote the unique R_0 values for each ship. We use the prior distributions and MCMC scheme denoted above to generate 3,250,000 posterior samples for analysis. The mean and 95% HPD intervals of the posterior samples are given in Table 4.9. The marginal kernel density estimates of the posterior samples are shown in Figure 4.10. The trace plot of the MH chain are shown in Figure 4.11.

Parameter	Mean (95% CI)
$R_0^{(10)}$	2.90 (0.65, 6.85)
$R_0^{(11)}$	3.47 (0.22, 9.41)
$R_0^{(12)}$	5.28 (0.74, 15.6)
$R_0^{(13)}$	3.51 (0.67, 8.60)
$R_0^{(14)}$	4.29 (1.13, 13.2)
$R_0^{(15)}$	4.35 (0.84, 12.0)
λ	0.67 (0.24, 1.00)

Parameter	Mean (95% CI)
$\frac{1}{\sigma}$	0.99 (0.10, 3.64)
$\frac{1}{\alpha}$	0.74 (0.10, 2.02)
$\frac{1}{\gamma_s}$	5.46 (2.19, 8.00)
$\frac{1}{\gamma_a}$	5.09 (1.37, 8.00)
p_s	0.94 (0.74, 0.99)
κ_S	0.23 (0.01, 0.50)
κ_P	0.24 (0.04, 0.49)
p_{sus}	0.10 (0.06, 0.17)

Table 4.9: Wave 3 ships posterior point estimates. Mean and 95% HPD intervals of posterior samples.

For the Wave 3 ships, we estimate all ships had an $R_0 > 1$, and 5 out of 6 ships have an $R_0 > 3$. Again, we note some marginal R_0 distributions are heavily right-tailed with 97.5% percentiles as high as 15. The Wave 3 R_0 values are on average smaller than the second wave estimates. The reduced R_0 could explain the lower attack rate seen in the third wave of the 1918 pandemic. Also, the λ values indicate an effective reduction in transmission of about 45% once entered Australian quarantine, which is slightly less effective reduction than for the second wave, but again we note the uncertainty in λ.

The epidemiological periods are estimated to be an average 24 hour latent period and 18 hour pre-symptomatic period; resulting in an average incubation period of 42 hours. Hence, infected individuals develop symptoms on average 2 hours faster than in Wave 2, but the overall incubation period is consistent with the second wave. Individuals are still infectious for roughly half of the time between infection and the development of symptoms. The average infectious periods are estimated as: symptomatic infectious period of 5.5 days and asymptomatic infectious period of 5 days, which is consistent with the second wave.

The generation time aboard the Wave 3 ships can be estimated through simulation. Here, we estimate $T_g^{(i)}$ from 10,000 posterior simulations; the estimated mean and 95% HPD intervals are shown in Table 4.10. The combined T_g kernel density estimate is shown in Figure 4.12. We see on average

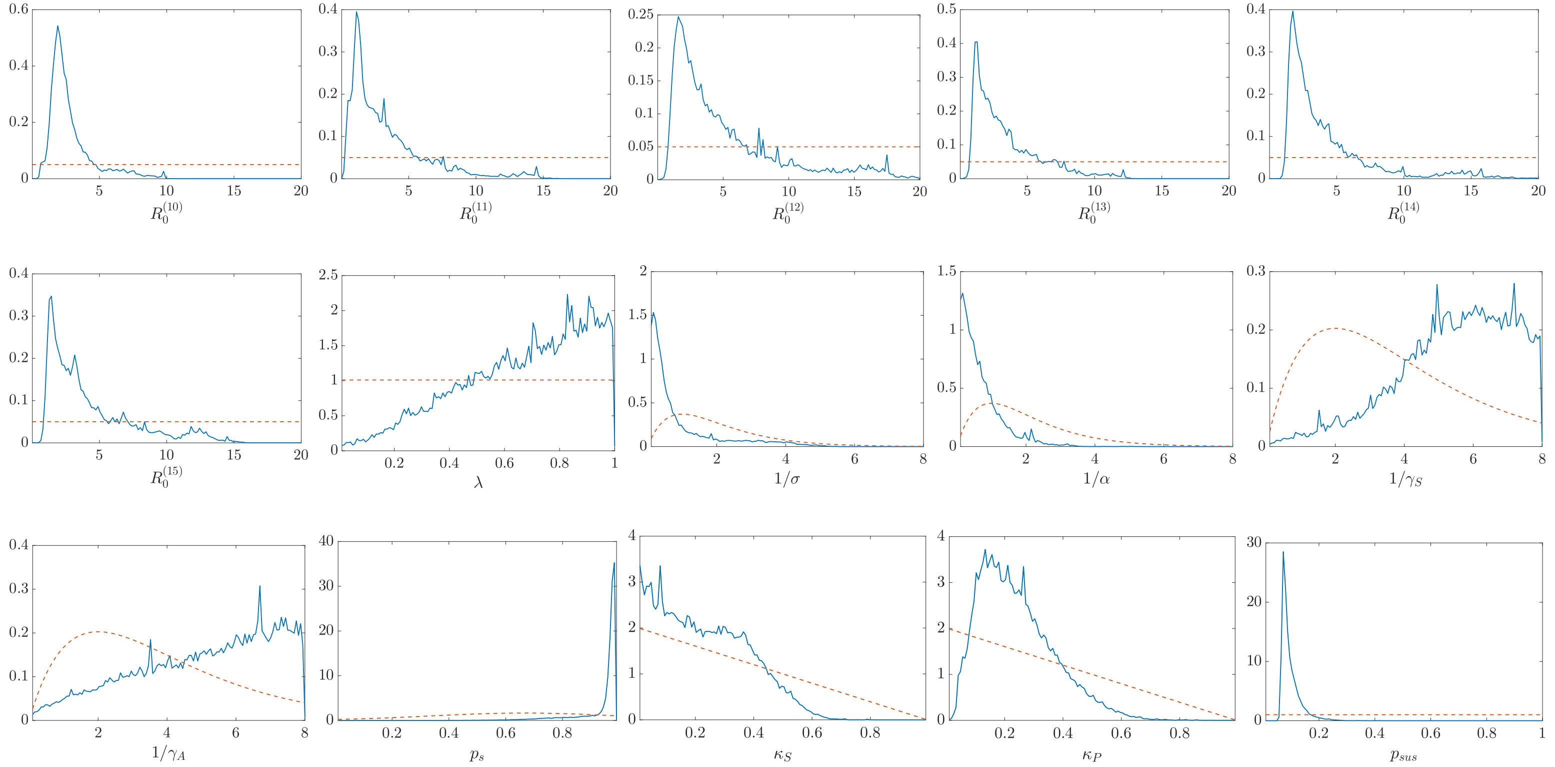


Figure 4.10: Wave 3 ships posterior marginal kernel density estimates. Red line denotes prior probability density function.

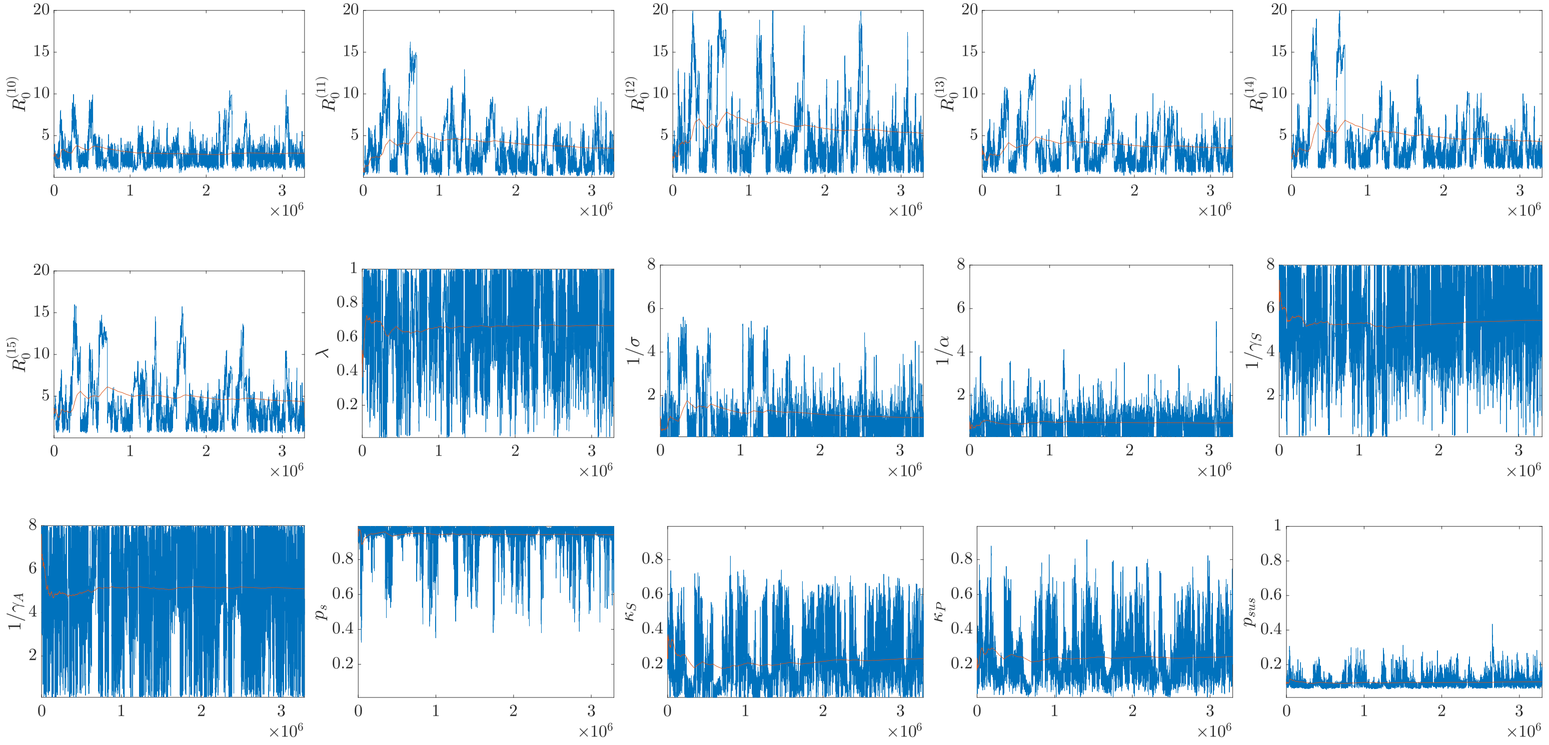


Figure 4.11: Wave 3 ships posterior trace plots by parameter. Red line denotes the cumulative mean.

a generation time of 58 hours, which is approximately 5 hours longer than the Wave 2 estimate. The longer generation time could be a result of the reduced R_0 compared to Wave 2, and hence slowed the epidemic growth in number of observed cases aboard the ships.

Ship	Mean (95% HPD)	Ship	Mean (95% HPD)
$T_g^{(10)}$	2.44 (0.00, 8.64)	$T_g^{(13)}$	2.63 (0.01, 9.00)
$T_g^{(11)}$	2.62 (0.01, 8.97)	$T_g^{(14)}$	2.46 (0.01, 8.48)
$T_g^{(12)}$	2.33 (0.00, 8.05)	$T_g^{(15)}$	2.44 (0.00, 8.26)

Table 4.10: Wave 3 ships posterior T_g estimates by ship. Mean and 95% HPD intervals estimated from 10,000 simulations.

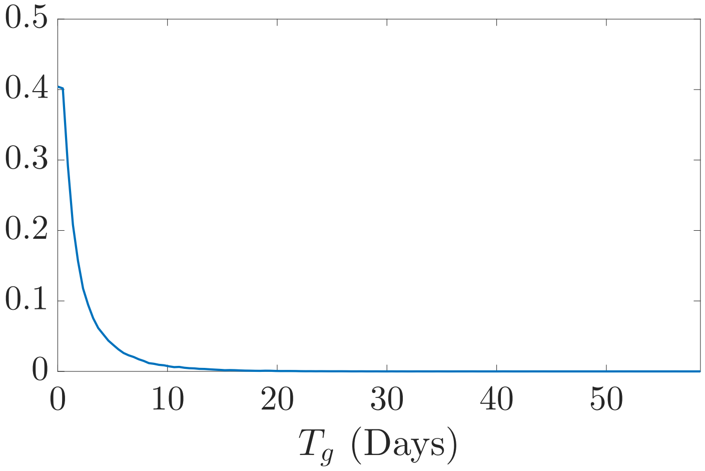


Figure 4.12: Wave 3 posterior T_g kernel density estimate.

We estimate approximately 95% of all infections are symptomatic, similar to Wave 2 estimates. Again, this could be explained by the better case recognition by on-board surgeons. In terms of prior immunity, we observed a significant increase compared to Wave 2. For the third wave ships, we estimate 90% of the population was immune to infection prior to the Wave 3 outbreaks; an increase from the estimated immune Wave 2 populations by roughly 50%. Again, this estimate is consistent with previous prior immunity wave analysis of the 1918 pandemic conducted in Mathews *et al.* [49]. Hence, we can support the conjecture that the increase in prior immunity from Wave 2 to Wave 3 of the 1918 pandemic was significant and aided in a far reduced attack rate observed in the third wave.

The combination of the high p_s estimates and low p_{sus} value that appears to fit to the data's largest clinical attack rate (*Ceramic*; 7% clinical attack rate), supports the conjecture that all susceptible individuals contract the disease, and we observed nearly all of them. Although, this elevated total attack rate appears less significant than in the second wave. Here, we have run 10,000 posterior simulations of the ship epidemic model. The simulated cumulative case counts are shown against the observed ship data in Figure 4.13. The comparison of data clinical attack rate to posterior simulation clinical and total attack rates is shown in Table 4.11. Again, we see consistent agreement with the ship data and almost zero difference between the simulation and clinical attack rates. The larger outbreaks (*Ceramic* and *Kenilworth Castle*) tend toward complete infection of the susceptible population, and the smaller outbreaks show significant probability of epidemic die-out.

	Ship Name	Data Attack Rate	Clinical Attack Rate Mean (95% HPD)	Total Attack Rate Mean (95% HPD)
(10)	Ceramic	0.07	0.08 (0.04, 0.12)	0.09 (0.04, 0.13)
(11)	Lancashire	0.03	0.05 (0.00, 0.10)	0.05 (0.00, 0.11)
(12)	Kenilworth Castle	0.05	0.08 (0.05, 0.12)	0.09 (0.05, 0.13)
(13)	Orca	0.03	0.05 (0.00, 0.10)	0.05 (0.00, 0.10)
(14)	Euripides	0.06	0.07 (0.01, 0.11)	0.08 (0.01, 0.13)
(15)	Kashmir	0.04	0.03 (0.00, 0.09)	0.03 (0.00, 0.10)

Table 4.11: Wave 3 ship data versus posterior simulation attack rates. Mean and 95% HPD intervals estimated from 10,000 simulations. Posterior simulations show in Figure 4.13.

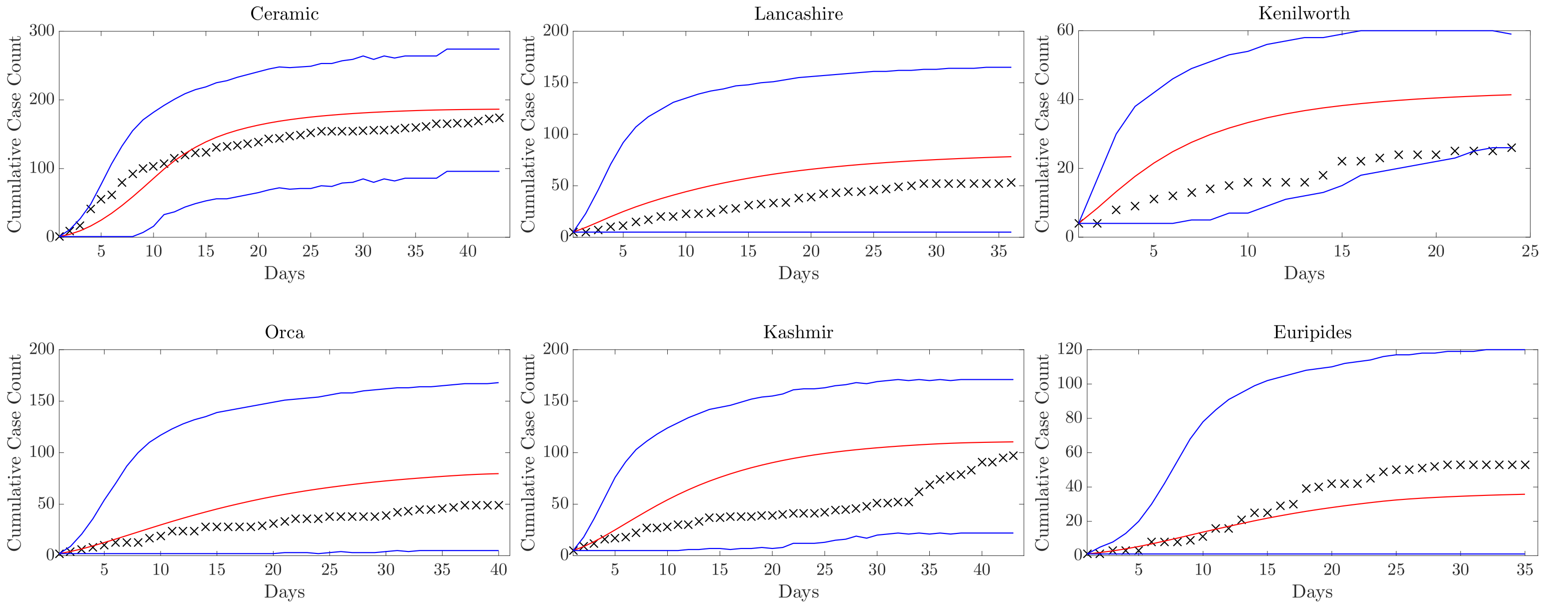


Figure 4.13: Wave 3 ship data versus posterior simulations. 10,000 ship epidemic model simulated datasets are generated using posterior sampled parameter sets. Black crosses denotes the observed cases. Red denotes the posterior simulation mean. Blue denotes the 95% HPD interval.

We wish to identify the proportion of transmission from symptomatic, pre-symptomatic, asymptomatic and non-symptomatic individuals aboard the third wave ships. The κ posterior mean and 95% HPD intervals are shown by ship in Table 4.12. We plot the κ kernel density estimates in Figure 4.14. We estimate the symptomatic, pre-symptomatic and asymptomatic mean transmission proportions as 0.23, 0.24 and 0.52 respectively. Hence, we estimate the non-symptomatic mean transmission proportion as 0.77. We see general agreement with the estimated proportion of transmission in the second wave, except for a ≈10% increase in symptomatic contribution. This appears to arise from a larger uncertainty in the symptomatic proportion of transmission than the second wave; this is not unexpected due the smaller number of datasets inferred upon. Hence, this does support our Wave 2 conclusion that non-symptomatic individuals were responsible for over half of all transmission in both waves; approximately 85% in Wave 2 and 75% in Wave 3.

Parameter	Mean (95% HPD)
κ_S	0.23 (0.01, 0.50)
κ_P	0.24 (0.04, 0.49)
κ_A	0.52 (0.12, 0.91)
κ_{NS}	0.77 (0.50, 0.99)

Table 4.12: Wave 3 ships κ estimates. Mean and 95% HPD intervals of posterior samples.

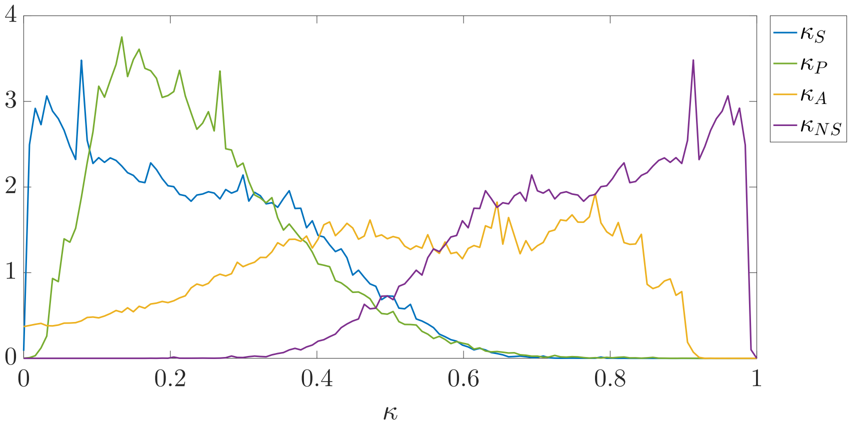


Figure 4.14: Wave 3 ship posterior R_0 contributions kernel density estimates.

Again, we must note the expected number of infections that arise from a symptomatic, pre-symptomatic or asymptomatic individual aboard the third wave ships. The E[infections] posterior mean and 95% HPD intervals are shown by ship in Table 4.13. We plot the overall E[infections] kernel density estimates in Figure 4.15. Here, we see a average expected infections from symptomatic, pre-symptomatic and asymptomatic individuals of 0.69, 0.80 and 109, respectively. Again, asymptomatic individuals have the potential to be *super-spreader* individuals that can cause a huge number of infections, but due to the small p_s values, there is on average only seven asymptomatic individuals in the duration of the epidemic.

	Ship Name	$E[I_S$ infections] Mean (95% HPD)	$E[I_P$ infections] Mean (95% HPD)	$E[I_A$ infections] Mean (95% HPD)
(10)	Ceramic	0.63 (0.02, 1.35)	0.70 (0.06, 1.32)	79.5 (0.00, 278)
(11)	Lancashire	0.57 (0.04, 1.16)	0.70 (0.02, 1.36)	112 (0.00, 444)
(12)	Kenilworth Castle	0.89 (0.04, 2.00)	1.04 (0.19, 2.29)	162 (0.00, 610)
(13)	Orca	0.60 (0.04, 1.15)	0.69 (0.20, 1.26)	109 (0.01, 391)
(14)	Euripides	0.72 (0.07, 1.31)	0.85 (0.27, 1.66)	134 (0.00, 526)
(15)	Kashmir	0.73 (0.08, 1.44)	0.85 (0.24, 1.62)	135 (0.00, 502)

Table 4.13: Wave 3 ships posterior E[infections] estimates by ship. Mean and 95% HPD intervals of posterior samples.

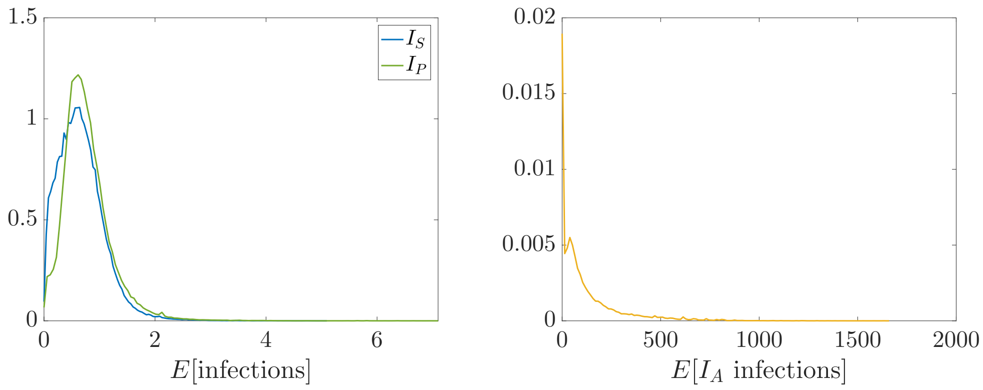


Figure 4.15: Wave 3 ships posterior E[infections] kernel density estimates.

In summary, we have characterised the epidemiological dynamics occurring aboard 6 ships during the third wave of the 1918 pandemic. We found the following statistics of note. R_0 values aboard these ships are in the range of 2.9 - 5.3, and were on average smaller than the second wave's R_0 estimates. The quarantine measures implemented upon entry to Australian quarantine stations still resulted in a 45% reduction in transmission; these estimates are consistent with Wave 2. Roughly 95% of all infections were symptomatic and these individuals were infectious on average 18 hours before development of symptoms; that is, for approximately half the incubation period. Only 10% of the population was susceptible prior to onset of outbreak in the third wave. Non-symptomatic individuals (asymptomatic or prior to onset of symptoms) were responsible for approximately 75% of transmission; where asymptomatic individuals were rare but potentially extremely transmissive.

4.4 Summary

In this chapter, we have conducted the principal investigation of this thesis in characterising the 1918 pandemic influenza strain. Firstly, we introduced the previously unstudied data source from Cumpston [20]. The data details a series of influenza outbreaks aboard ships, corresponding to two waves of the 1918 pandemic: 9 ships in the second wave and 6 ships in the third wave. The ship epidemic model was developed specifically to characterise the pandemic influenza epidemiology from the ship outbreaks. This includes mechanisms for asymptomatic infections, pre-symptomatic infectious periods, prior immunity, removal of infected and healthy individuals, and reduction in transmission caused by quarantine measures. The ship data was used to characterise the 1918 pandemic by inferring the parameters of the ship epidemic model. Parallel inference was conducted in outbreak subsets corresponding to pandemic wave; this allowed characterisation of the 1918 pandemic influenza strain and identification of differences between the second and third wave of the pandemic.

We estimate R_0 values in the range of 1.3 - 9.5 for the second wave and 2.9 - 5.3 for the third wave. For context, previous studies of the 1918 pandemic have estimated R_0 in the range of 1.4 - 2.8 [23, 54], and estimates from contained environments such as schools and military bases in the range of 2.8 - 5.7 [48, 49]. The exact comparison to past studies is unwise as the R_0 value would be expected to be higher aboard ships than in open populations. Although, these values are consistent with previous studies and support the conclusion that one of the contributing factors to the severity of the 1918 pandemic was a highly transmissible virus; far more infectious than reported estimates of endemic influenza and other modern pandemics [18, 23, 79]. We also note the mean Wave 2 R_0 estimate is 0.75 larger than the Wave 3 estimate. Partial responsibility for the decline in R_0 between waves falls to the significant increase in the immune population prior to the third wave. We found approximately 45% of the population aboard the Wave 2 ships were immune to infection prior to onset of outbreak and 90% in the Wave 3 ships. From R_0, the generation time was estimated as ≈ 2 days between a primary and secondary infection. The incubation period of symptomatic infections was on average 42 hours, with individuals were infectious to others for approximately half that time. Across both waves, approximately 95% of all infections developed observable symptoms. Approximately 80% of transmission occurred asymptomatically or prior to the onset of symptoms; where asymptomatic infections were rare but accounted for a large number of transmissions. We also found that the Australian quarantine measures had an approximately 35 - 50% reduction in transmission. These results are discussed further with context in Chapter 5.

Chapter 5

Discussion

The objective of the work within this thesis is to epidemiologically characterise the 1918 pandemic influenza strain. This study was not only important to the understanding of the worst influenza pandemic in history, but is informative of present day influenza transmission. Previous studies, while plentiful, were often restricted to data and epidemic models, or their deterministic approximations, that failed to account for important dynamics deserving characterisation or ignored external factors which could potentially bias results. Here, we investigated a number of ship-bound influenza outbreaks from the 1918 pandemic, as recorded within Cumpston [20]. These outbreaks presented a naturally occurring, detailed picture of influenza transmission within a contained environment. The data is unlike that typically seen in previous mathematical studies, especially for the 1918 pandemic where reliably recorded data is sparse, populations are often subject to external factors and mixing dynamics are difficult to identify and replicate. Hence, the data provides a rare chance to best capture the epidemiological characteristics of the 1918 pandemic.

Chapter 2 introduced the PMMH algorithm, which was required to conduct inference using stochastic epidemic models of the desired complexity. The algorithm uses an SMC estimate of the likelihood which had potential to cause issues with the mixing of the MH chain. SMC was shown to have difficulty estimating rare event probabilities, and typically methods to alleviate these issues were found to be not practical. We attempted to devise our own alternative method in the Hybrid algorithm, that had its own set of complications (see below). Other recent inference methods for stochastic epidemic models such as data augmentation deserve consideration and may have been better suited to the task for computational reasons [10, 51]. Chapter 3 details the investigation into the rigorous testing of the PMMH algorithm in comparison to the Metropolis-Hastings algorithm. The PMMH algorithm was found preferential such that it replicated the posterior results of the Metropolis-Hastings algorithm in a significantly shorter runtime; this validated the used of the algorithm. In this study we compared algorithms using a standard SMC number of particles, N_p, and implicit Euler precision, τ. These values were chosen by studying the variance of the likelihood estimate calculated for a single parameter set. Ideally, we would compare the PMMH and Metropolis-Hastings algorithms across a variety of standard N_p and τ values to better identify the effects of the number of particles required.

In Chapter 3, the PMMH algorithm was used to assess the parameter identifiability of a series of increasingly complex epidemic models. These models were built upon to develop the SEIpIsIaR model, which formed the foundation of the ship epidemic model. The inference tests on the SEIpIsIaR model found definitive biases and parameter identifiability concerns when conducting inference from a single dataset of symptomatic case observation. Biases such as overestimating R_0/underestimating p_s and inability to identify asymptomatic infection dynamics were noted. Parallel inference highlighted the identifiability benefits from conducting inference on multiple datasets, which reduced these biases. Although, some identifiability concerns remained; especially when allowing unique R_0 values between outbreaks. Primarily, a tendency to slightly overestimate R_0 and p_s, and an inability to adequately identify κ_P, κ_A and $E[I_A$ infections]. The conclusions of the identifiability study are limited by the fixed parameter choices and number of datasets used. Here, simulated datasets were generated using population sizes and parameter values informed by past studies to replicate the ship data. Ideally, a more comprehensive study of a series of parameters sets would be undertaken to better identify possible biases. Included in these considerations is a quantitative study of the number of datasets inferred upon. That is, how many datasets must be used within parallel inference to achieve a desired result. Here, we only used 5 datasets as similar to Wave 3 of the ship data. There are indications of possible reduction of biases/identifiability concerns given the inclusion of more datasets, but the number required deserves study. Identifiability studies of this detail are rarely published and there is little in the literature for comparison. Hence, a study of this detail could be useful to validate the conclusions of other studies.

Chapter 4 presents the principal study of parameter inference conducted in an effort to characterise the 1918 pandemic influenza strain. In short, we found evidence that supported the following epidemic narrative. The ships were under attack from a highly transmissible strain of influenza, consistent with previous studies of the 1918 pandemic. The reduced clinical attack rate between Wave 2 and Wave 3 was primarily a result of large increase in the immune population, from approximately 45% to 90% immunity. The epidemic spread quickly due to the high infectiousness and roughly one day latent period; a primary infection would transmit the disease after 2 days on average. Approximately 95% of all infections developed symptoms and roughly 80% of transmission arose from non-symptomatic individuals, that is, asymptomatically or in the 20 hours prior to development of symptoms. These results highlight key factors about the 1918 pandemic, and potentially modern influenza. Here, we provided further support to the hypothesis that the 1918 pandemic was caused by a highly transmissible virus, and the reduced impact between second and third waves was a result of significantly increased population immunity [49, 74]. We suggest that the pre-symptomatic infectious period is responsible for approximately half the incubation period; indicating individuals are infectious for a significant period of time before the development of symptoms. We provide evidence that non-symptomatic transmission was the dominant method of transmission aboard these ships. The proportion of transmission from symptomatic individuals could suggest far reduced levels of mixing/contacts from symptomatic cases, possibly due to the debilitating symptoms as described in Cumpston [20]. This result is important for two reasons: the severe lack of evidence attributed to asymptomatic and presymptomatic transmission, and its significance to the informing the choice of epidemic control measures [25, 60]. This result has the potential make influenza control measures that targeted only symptomatic individuals, such as case isolation, obsolete. Hence, we have outlined a plausible epidemiological characterisation of 1918 pandemic influenza aboard ships and drawn conclusions with consequences to the field.

Although, we note and acknowledge the well documented limitations of stochastic compartmental models in reflecting true transmission dynamics [3, 4]. There are also some concerns within this analysis relating to aforementioned identifiability/bias. The 1918 pandemic inference estimated the proportion of symptomatic cases as approximately 95 - 97%; an especially high estimate when compared to previous studies of the 1918 and other pandemics. While this is argued as a result from the mandatory reporting of all cases to the ship surgeon, it is possible that the high values are a result of the overestimating R_0/p_s bias. This is compounded by the inability to identify $E[I_A$ infections], which allows the possibility of asymptomatic individuals to be highly infectious. It is difficult to determine if these estimates are a bias of the model or illustrative of the true dynamics. Hence, we suggest the further model testing detailed above, and propose model adjustments for comparison. In particular, we wish to reduce the impact of transmission from asymptomatic infections as common of past studies. This include: using an Erlang distributed infectious compartment to reduce duration variance [48], using more informative/restrictive prior distributions educated from past studies, and truncating the support of the $E[I_A$ infections] prior distribution to realistic values. These changes could eliminate/reduce the bias potential and are worthy of investigation.

5.1 Further Areas of Study

In the next section, we detail further areas of study regarding the original work proposed here, but that falls outside the scope of this thesis. There are still questions about the epidemiology of the 1918 pandemic and more unanalysed data to be found within Cumpston [20]. Here, we detail other possible future investigations into the data. We also detail possible improvements for the Hybrid algorithm.

Cumpston [20] contains a register of all ships entering Australian during the maritime quarantine period of the 1918 pandemic. For our analysis, we extracted 15 of the most detailed accounts that support a set of modelling assumptions. These outbreaks were chosen as they met the minimum requirements as follows: daily recordings of influenza case counts with first and last case date (some missing data permitted), a closed population within a contained environment, introduction of an infected individual into a susceptible (or partially immune) population at known location, no effective means of isolation, all passenger landings are recorded with infection status, and all quarantine measures implemented are recorded. The data requirements meant each outbreak painted an accurate picture of a contained influenza epidemic with minimal external factors; permitting the explicit modelling and conclusions drawn in Chapter 4. Although, if these data requirements are relaxed, the number of ships outbreaks that meet the criteria, and therefore the data available for analysis, increases. However, the missing detail in the records reduces the strength of conclusions that are able to be drawn. That is, further analysis can be conducted on a larger set of ship outbreaks than the original 15, but the conclusions may be weakened due to possible influences of externals factors or unaccounted dynamics.

There are 75 registered ships that were observed to have influenza cases on board during the voyage to Australia. Omitting the 15 already included in analysis, there are 13 additional ships that meet the minimum data requirements except the daily resolution influenza case counts. The number of additional ships jumps to 21 if assumptions are made about inconsistencies in the recorded data, and 27 if the source-of-infection requirement is ignored. Hence, there are 42 ship outbreaks of reasonable fidelity in the records available for analysis. Alternatively, each of the 75 registered ships records a final size of the outbreak; a naive approach could be to analyse the final size distribution of all ships regardless of surrounding circumstance. Note, this summary doesn't include the 92 Australian vessels that developed influenza cases on interstate voyages, which may be more susceptible to external factors than the international voyages, but are still worthy of consideration. This brings the prospective total number of ships outbreaks for analysis to 167. Hence, there is potential for much further study of the 1918 pandemic from the other data available in Cumpston [20], given relaxed modelling assumptions.

In Chapter 4, we conduct two studies of the original 15 ship outbreaks, split into groups by wave of the 1918 pandemic. While this was informative of the differences in epidemiology between pandemic waves, further comparisons could also have been made. In particular, one potential difference worthy of further study is geographical-based inference. That is, a comparison of ship outbreaks between source-of-infection cities. During the 1918 pandemic, some countries/cities experienced noticeably more virulent influenza epidemics than others and demonstrated the pandemic waves at different times [74]. Hence, a study of the possible differences by geographical-based inference is recommended. This could be done a number of ways, such as by hemisphere, country or city comparisons. Given many of the ship outbreaks originated from England, Egypt, South Africa and New Zealand, there is ample data for this comparison. The study would not be difficult to implement as it only requires different groupings of ships to be used in parallel inference. A study conducted in this manner has the potential to be informative about dynamics that could vary significantly based on origin of population, such as prior immunity [49]. Hence, a parallel inference study of the ship data, broken into groups by location and/or local pandemic wave could be enlightening about the differences in epidemiology of the 1918 pandemic by country.

Another study worth consideration is adjustments to the ship epidemic model to account for a heterogeneous population. In Chapter 4, we argue the assumption of a homogeneously-mixing population for mathematical simplicity, as common of mathematical epidemic studies [4, 41]. We note that this assumption is an approximation to the realistic heterogeneity in the population, as evident by the differing attack rates of passengers/troops and crew aboard some ships (see Appendix). Hence, a consideration of *levels of mixing* is warranted, similar to epidemic *patch* or *household* models [3, 36, 47]. These models account for levels of mixing within a population, allowing contact between individuals to occur within and between subgroups of the population at different rates. This concept could be extended to the ship epidemic model to account for the different levels of mixing between personnel, such as crew and passengers. The subgroups could even have unique observations processes as records show the distribution of cases between crew and passengers for some ships. The modelling of multiple levels of mixing could be expanded beyond just crew and passenger subgroups. Some ships describe the progression of disease transmission along the length of the ship, spreading between the sleeping quarters. Unfortunately the exact structure of the ship is difficult to determine, and so a number of assumptions would be required about the proposed subgroups. Modelling multiple levels of mixing aboard the ship could greatly alter the characterisation of ship outbreaks. With this approach, we may be able to better understand the reason for unusual epidemic characteristics in the ship data, such as the observed cases slow initial development, sharp spike, followed by a long tail of cases that could last weeks. The inclusion of a heterogeneously-mixing population could greatly improve the fit of the model. Although, this study would require a significantly more complex model which may hinder inference.

Here, we have discussed multiple areas of further study into the 1918 pandemic from data collated within Cumpston [20]. We detail the vast amount of ship outbreaks that have not been included within our analysis, due to lacking the minimum data requirements such as daily observation of influenza case counts and agreement with modelling assumptions. These datasets could be used for analysis by relaxing the required data assumptions. The inclusion of the remaining data may complicate epidemic modelling, but the sheer number of outbreaks is unlike past studies of 1918 pandemic influenza and worthy of further investigation. We also detail two possible studies into the ship data worthy of undertaking. The first is an alternate inference scheme that compares the 1918 pandemic influenza epidemiology on a source-of-infection location basis. The study could identify possible key geographical differences that explain why the progression of the 1918 pandemic was unique city-to-city and country-to-country. The second is an alternate model that allows for multiple levels of mixing within the on-board population. This inclusion within the model could explain the characteristics of the epidemic progression on some vessels, and potentially illustrate different epidemiological dynamics than in Chapter 4. There are many possible avenues of future work that can be investigated from the ship data. Admittedly, the future areas of study documented here is not an exhaustive list of possible ways the data can be exploited. As such, Cumpston [20] is a rich resource for mathematical modelling and should be utilized to the fullest extent in characterising the 1918 pandemic.

Hybrid Algorithm

In Chapter 3, we present a Hybrid algorithm that combines SMC and "exact" (using the implicit Euler method to solve the master equation) methods in likelihood calculation. We present a formulation of the algorithm which uses a *probability threshold*, δ, which determines the time steps where exact methods are used. We highlight the variance reduction benefits of the Hybrid algorithm, and noted the computational cost of the exact methods made it an infeasible choice for our study. Although, the Hybrid algorithm can be effective for simpler models and there exist other implementations/improvements to the method that may prove beneficial in the field.

There are multiple ways the Hybrid algorithm can be implemented by using different conditions to determine the time steps where the exact methods are used. In Chapter 3, we use the exact method for time steps where the estimated SMC likelihood is below a threshold δ. The choice of δ can be chosen to elicit certain behaviours of the algorithm. That is, δ can be chosen so exact methods are only used for time points suffering likelihood failure, minimising the number of observations where exact methods are used. A larger δ could mean exact methods are used when the estimated likelihood is too small to be reliable and an increased precision is required. This implementation is specifically designed to remove the possibility of likelihood failure/particle degeneracy and reduce the estimator variance. This comes at the cost of using the exact methods for an unknown number of time steps, which can be computationally costly if δ is poorly chosen. Alternatively, the algorithm can be implemented where exact methods are used for predetermined time steps or time steps chosen by any arbitrary condition. For an epidemiology example, the exact methods could be used for the first few days of the outbreak. This period is often highly variable and has the highest probability of epidemic die-out; hence the Hybrid algorithm can increase the precision of the estimator for the most variable time steps. Implementing the Hybrid algorithm in a problem-specific manner is favourable if it can reduce the number of time steps where exact methods are required and while maintaining the benefits over SMC.

As discussed in Chapter 3, the efficiency of the Hybrid algorithm is dependent on the efficiency of the exact methods, hence the implicit Euler method, even if only used for a small number of time points. If the implicit Euler method is infeasible or an order of magnitude larger in runtime than SMC, the benefits of the Hybrid algorithm are mostly lost by the huge increase in runtime. This will limit the use of the algorithm to simpler models, but there are techniques to speed up the implicit Euler method. The most common is truncation of the CTMC's Q matrix to only consider states of interest; a method known as finite state projection. We neglect the documented mathematical background here for brevity but it can be found in Munsky and Khammash [57] and Jahnke and Sunkara [38]. Truncation can be quite effective in reducing the runtime of the implicit Euler method as it can greatly reduces the size of the state space considered. The gains in efficiency are directly proportional to the decrease in considered state space. Hence, for the SIR and SEIR models truncation is recommend, as it is easy to greatly reduce the state space considered. For some models such as the SEIpIsIaR, the number of states able to be removed from consideration is limited and so truncation doesn't have as large an effect. Although, the combination of the SMC and exact methodologies with Q matrix truncation could result in a Hybrid algorithm that is feasible efficiency for analysis and greatly reduces the estimator variance compared to SMC.

Appendix A

Appendix

We provide detailed summaries of the ships, passengers and outbreaks that constitute the 1918 pandemic ship data, collated from Cumpston [20]. There are 15 ships in total:

(1) Niagara
(2) Mataram
(3) Devon
(4) Marathon
(5) Atua
(6) Manuka
(7) Medic
(8) Boonah
(9) Nestor
(10) Ceramic
(11) Lancashire
(12) Kenilworth Castle
(13) Orca
(14) Kashmir
(15) Euripides

A summary table of the ships is given in Chapter 4. Here, within each summary is a brief description of (where available): ship dimensions, travel route and reason for travel, population breakdown, case breakdown, summary of quarantine measures implemented, full timeline of epidemic including arrival/departure/quarantine dates, and any other relevant miscellaneous information. Note, all blue writing denotes modelling assumptions made or adjudicated inconsistencies in the records. All dates refer to September - December 1918, and January - April 1919.

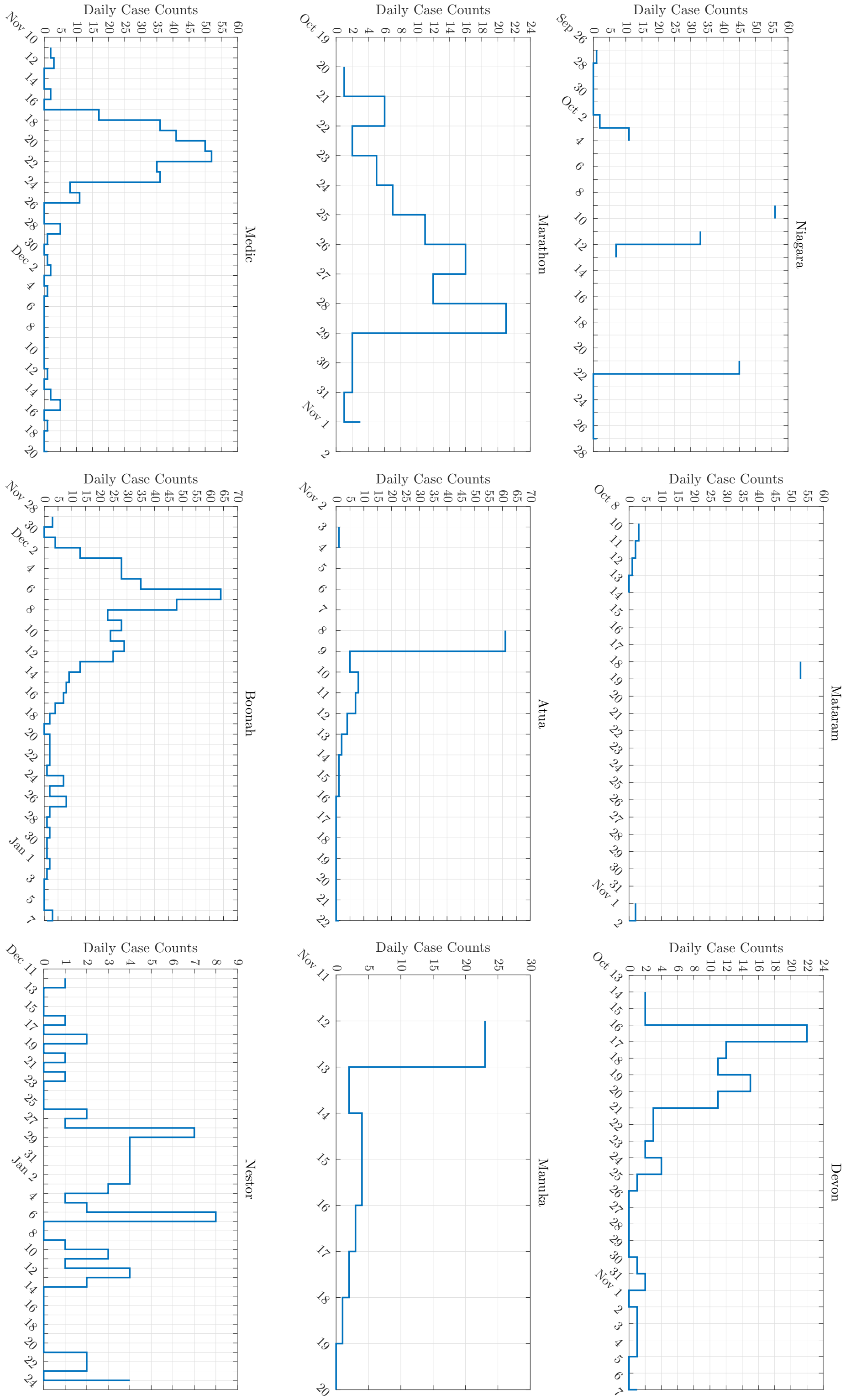


Figure A.1: Daily case observations. No line represents missing data.

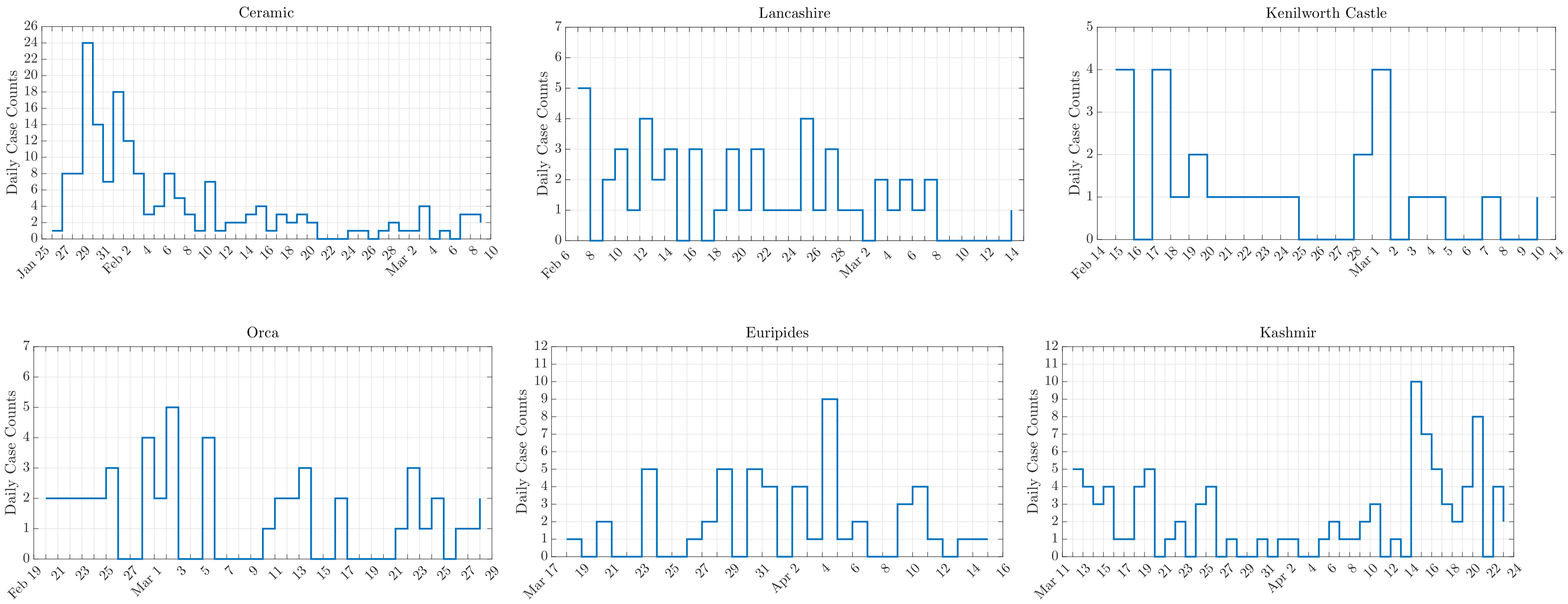


Figure A.2: Daily case observations. No line represents missing data

Niagara

Ship Information

The *Niagara* was an Australian trading vessel measuring 160m by 20m [76]. It travelled from Vancouver (Canada) to Honolulu (USA), Suva (Fiji), Auckland (New Zealand) before arriving into quarantine in Sydney (Australia). It brought on the influenza virus in Vancouver.

Passenger Breakdown

The passengers aboard the vessel totalled 566, broken down in Table A.1.

Crew	Number
Officers	9
Engineers	15
Deck Workers	46
Engine Workers	26
Providers	158
Total	254

Passengers	Number
First Class	153
Troops	106
Total	312

Table A.1: Niagara passenger record.

Case Breakdown

There was a total of 156 cases out of the 566 personal, broken down by group in Table A.2.

Crew	Number
Stewards	94
Deck Workers	26
Engine Workers	25
Total	145

Passengers	Number
Saloon	7
Second class	4
Third class	0
Total	11

Table A.2: Niagara case record.

There was also a large number of crew with mild cases who continued to work but with no record of infection. Assume these are asymptomatic infections. The infection began in the “after peak” (back end) section of the ship with 14 out of 16 becoming infected within 7 days. Afterwards the infection spread up through the ship, infecting crew/passengers in different sections. During the stay in Auckland, the ship captain brought on board a doctor, 4 nurses and a dispenser and attempted to isolate all cases to section of ship. Assume standard transmission rate. After implementing these measures the epidemic “practically ceased” after 9 days (when the ship departed Auckland). No record of epidemic decline or last case. Discussion states 45 cases took place in Auckland. The Niagara entered quarantine in Sydney on November 13. Upon entry to Sydney, there were 4 ongoing cases on board that had occurred since leaving Auckland. There were 5 deaths as a result of influenza. Three crew members died on Oct 13, Oct 21 and Nov 11. One nurse and one passenger died in Auckland.

Ship	Source	Arrival Date	Days at sea	Cases before arrival	Cases after arrival	Deaths	Crew	Troop	Total
Niagara	Canada	Oct 25	30	155	1	5	254	312	566

Table A.3: Niagara summary table.

Timeline Information

Sep 24 Niagara departs Vancouver. Between Sep 19 - 24, an individual brings infection on board as *Exposed* class.

Sep 27 Crew member taken ill with “severe cold” and goes off duty. Later confirmed to be influenza. First observed case.

Sep 30 Arrives in Honolulu. Assume no infection brought on board.

Oct 1 Departs Honolulu.

Oct 2 Two cases observed. Ship surgeon diagnosis “Feverish cold”. Note, these men shared same quarters as first case.

Oct 3 Men from the same ship section as previous cases became infected. 7 men came down with disease in morning and by afternoon, 14 out of 16 men from section came down with disease. As 3 cases prior, 11 cases observed on this day.

Oct 4 - 8 Infection spreads to midship peak and forward to stewards and sailors. From there to the firemen further forward on the lower deck. No information on timing of these cases. At some point Captain agreed with surgeon and doctors to isolate worst cases to hospital and in quarters on bridge deck, but was quickly made redundant by large number of cases. Assume standard rate of infection.

Oct 9 Arrives and Departs Suva. 70 crew members were down with disease. Assumes 70 cases observed so far. Too many cases on ship so isolation was infeasible.

Oct 11 103 cases so far. Day of first passenger cases as disease moves up ship. No numbers or details.

Oct 12 Arrives in Auckland. 110 cases so far. Discussion states 110 cases occurred between Oct 2 - 12 This ignores the first case. Assume 110 includes first case. 29 cases landed into Auckland hospital. Land 29 infected individuals. Captain attempted isolation of patients not sent to hospital by boarding off section of ship. Also brought on board a doctor and 5 nurses. A nurse became sick after a few days and was replaced. Assume standard rate of infection during this time. Discussion says isolation was not complete and termination of epidemic wasn't due to an external force.

Oct 21 After 9 days epidemic has "practically ceased". No information if last case occurred in Auckland or on route to Sydney in appendix. Discussion states 45 cases occurred between Oct 12 - 21. Assume 155 cases so far. Departs Auckland. On board: 280 crew and 312 passengers (592 total) after taking on board 103 passengers in Auckland. Note, 10 individuals were too ill to rejoin the ship after hospital and one died. Hence, permanently landed 29 - 11 = 18 at Auckland. Ignore population size comment and rejoined convalescent as numbers are not consistent. E.g. Prior to arrival 254+312=566, landed 18 = 548, take 103 on board = 651.

Oct 25 Arrives in Sydney. States 155 cases in total. One other case recorded after quarantine. Assume the 45 cases difference take place in Auckland consistent with Chapter 2 discussion. Arrives in quarantine with 4 cases on board. Begin quarantine transmission rate and remove cases at onset.

Oct 27 Last case observed. End of cases.

Oct 28 Healthy individuals released from quarantine station. No new observed cases on the 28th. End of data.

Mataram

Ship Information

The *Mataram* was an Australian trading vessel measuring 102m by 13m [35]. The ship travelled from Singapore (Singapore), where it brought on board the the influenza virus, to Samarang (Indonesia) and then to Darwin (Australia).

Passengers Information

The passengers aboard the vessel totalled 189, broken down in Table A.4.

Crew	Number
Officers	15
Engine Workers	29
Deck Workers	21
Providers	23
Total	88

Passengers	Number
First Class	42
Second Class	43
Third class	16
Total	101

Table A.4: Mataram passenger record.

Case Information

There was a total of 61 cases out of the 189 personal, broken down by group in Table A.5.

Crew	Number	Cases	Percentage
Malay sailors	21	11	52%
Indian firemen	29	20	69%
Chinese stewards	22	10	45%
European stewardess	1	1	100%
European officers	15	9	60%
Total	88	51	58%
Passengers	Number	Cases	Percentage
Saloon	42	7	17%
Second class	43	0	0%
Deck	16	1	6%
Total	101	8	8%

Table A.5: Mataram case record.

The Mataram entered quarantine in Darwin on Oct 18. Some individuals suspected to have had suffered influenza prior to embarking, due to heavy prevalence in Singapore. The prior immunity is thought to have contributed to the small number of cases, but this is not fully supported by the distribution of cases by crew origins. Although, no newly-shipped Singapore men or Greek steerage passengers were infected; they had all had the disease previously. After leaving Darwin, daily inspection and zinc-sulphate inhalation was undertaken and only 2 further cases appeared. Discussion says 6 cases on arrival to Sydney. Assume includes 4 previously observed cases from prior to Darwin. The quarantine measures were thought to have been just a contributing factor as most cases were already convalescent upon arrival to Darwin. Due to wide spread of disease, any susceptible was more than likely already exposed.

Ship	Source	Arrival Date	Days at sea	Cases before arrival	Cases after arrival	Deaths	Crew	Troop	Total
Mataram	Singapore	Oct 18	9	59	2	0	88	101	189

Table A.6: Mataram summary table.

Timeline Information

Oct 9 Mataram departs Singapore. Between Oct 4 - 9, an individual brings infection on board as *Exposed* class.

Oct 10 First 3 cases observed amongst Malay crew (living in front section of ship).

Oct 11 2 cases observed.

Oct 12 One more case before arrival to Samarang. Assume occurs on 12th.

Oct 13 Arrives in Samarang. Unknown departure time. Assume no infection brought on board.

Oct 14 Infection breaks out amongst engine-room staff (living in back section of ship). No information detailing outbreak.

Oct 18 Arrives in Darwin. 59 cases have occurred prior to arrival. Enters quarantine. Begin quarantine transmission rate. End of Data. (Truncated due to no landing information and inconsistent cases.)

Nov 1 Arrives in Sydney with 56 crew and 55 passengers. No details of personal landing to account for change in passenger numbers. Arrives with 6 cases. Discussion states only 2 more cases after reaching Darwin.

Devon

Ship Information

The *Devon* was an New Zealand transport vessel measuring 144m by 18m [35]. On voyage back from London (England), the Devon stopped at Port Said (Egypt), Suez (Egypt) and Colombo (Sri Lanka) before arriving in Fremantle (Australia). It brought on board the influenza virus in Suez.

Passengers Information

The passengers aboard the vessel totalled 1,096, broken down in Table A.7.

Crew	Number
Total	110

Passengers	Number
First Class	62
Second Class (Troops)	829
Third Class (Troops)	829
Total	986

Table A.7: Devon passenger record.

Case Information

There was a total of 95 cases out of the 1,096 personal, broken down in Table A.8.

Personnel	Number	Cases	Percentage
Crew	110	14	12%
First Class Passengers	0	0	0%
Troops	920	81	8.8%
Total	1096	95	8.7%

Table A.8: Devon case record.

Ship	Source	Arrival Date	Days at sea	Cases before arrival	Cases after arrival	Deaths	Crew	Troop	Total
Devon	Egypt	Nov 15	36	95	0	0	986	110	1,096

Table A.9: Devon summary table.

Timeline Information

Oct 10 Devon departs Port Said after embarking troops.

Oct 10 Arrives in Suez.

Oct 13 Departs Suez. Between Oct 10 - 13, an individual brings infection on board as *Exposed* class.

Oct 14 First 2 (troop) cases observed.

Oct 16 Large spike as 22 observed. Unrealistic spike in cases likely caused by observation error. Smooth 22 cases over day prior. Condition on 24 cases in 2 days. No information about data collection. Assume first day of close monitoring caused previous cases to be recorded.

Oct 19 First (crew) case observed.

Oct 29 Arrives in Colombo. Assume no introduced infection.

Nov 1 Departs Colombo.

Nov 2 Last (crew) case.

Nov 7 Last (troop) case. End of cases.

Nov 15 Arrives in Fremantle. No new observed cases up to this date. End of data.

Marathon

Ship Information

The *Marathon* was an Australian transport vessel measuring 154m by 17m [76]. The Marathon landed at Devonport (England), where the influenza virus was bought on board, before travelling to Cape Town (South Africa) and onto Albany (Australia) and Melbourne (Australia).

Passengers Information

The passengers aboard the vessel totalled 1,041, with 920 troops and 121 crew.

Case Information

There was a total of 89 cases out of the 1,041 personal. The Marathon arrived at Devonport and embarked troops on Oct 20. Prior to arrival, there had been an influenza outbreak amongst the crew. These cases are not recorded and outbreak ceased prior to arrival. Previous cases ignored. This caused caused the departure date to be pushed to Nov 7. During this time, the infection broke out amongst the military personal. As the virus was already on board, infection was likely introduced to the embarking troops. The cases were observed from Oct 21 to Nov 1, so the last case occurred before leaving Devonport. The captain had attempted isolation of cases, but due to inadequate facilities, the isolation area was still commonly used by healthy individuals. Assume standard transmission rate. On Oct 28, twice daily spraying of formaldehyde was conducted amongst troops until arrival to Capetown. Assume quarantine transmission rate. This is thought to have stopped the spread of infection. The Marathon entered quarantine in Albany on November 13.

Ship	Source	Arrival Date	Days at sea	Cases before arrival	Cases after arrival	Deaths	Crew	Troop	Total
Marathon	England	Dec 23	65	89	0	4	920	121	1,041

Table A.10: Marathon summary table.

Timeline Information

Oct 20 Marathon enters Devonport. Troops are embarked. As only troop cases reported, use troop population. One case on Oct 20 to support total case number.

Oct 21 First recorded case. Between the Oct 15 - 20, an individual brings infection on board as *Exposed* class. Extend due to possible influenza virus on board prior to embarking troops.

Oct 28 Begin formaldehyde spray. Begin quarantine transmission rate (unique to other ships due to formaldehyde treatment).

Nov 1 Last case. End of cases.

Nov 17 Departs Devonport. No new observed cases up to this date. End of data.

Atua

Ship Information

The *Atua* was an New Zealand transport vessel 97m by 13m [76]. The Atua left Suva (Fiji) before calling into Auckland (New Zealand) on way to Sydney (Australia). It brought on the influenza virus in Auckland.

Passengers Information

The passengers aboard the vessel totalled 163, broken down in Table A.11.

Crew	Number
Officers	7
Engineers	6
Deck Workers	13
Engine Workers	20
Providers	31
Total	77

Passengers	Number
First Class	67
Troops	19
Total	86

Table A.11: Atua passenger record.

Case Information

There was a total of 88 cases out of the 163 personal, broken down in Table A.12. Data shows 91 cases.

Personnel	Number	Cases	Percentage
Crew	77	56	72%
Passengers	86	32	37%
Total	163	88	54%

Table A.12: Atua case record.

The first case occurred the day after leaving Auckland. Upon arrival arrives to Sydney with 54/77 crew and 8/86 passengers “suffering from most virulent type of influenza”. Assume 54+8 = 62 cases so far. The Atua entered quarantine in Sydney on Nov 8. The cases could be broken down by severity upon entering quarantine in Table A.13.

Crew	Number
Dangerously ill	10
Seriously ill	13
Ill (but trying to work)	13
Convalescent	18
Well	23
Total	77

Passengers	Number
Dangerously ill	4
Seriously ill	4
Convalescent	15
Doubtful	1 (later confirmed.)
Well	55
Total	79 (86.)

Table A.13: Atua case severity breakdown.

After entering Sydney, all sick were removed to hospital and everyone else removed to isolation on the shore. Land all 62 cases that have taken place so far. Further cases continued to occur in isolation. The last case on Nov 22 can be ignored due to extended period since last infection and likely transmission from quarantine station staff. 29 cases take place after quarantine, 91 cases in all. This was the first cases where the Commonwealth coryza vaccine was first tried with successful results.

				Cases	Cases				
		Arrival	Days	before	after				
Ship	Source	Date	at sea	arrival	arrival	Deaths	Crew	Troop	Total
Atua	New Zealand	Nov 8	6	62	26	16	77	86	163

Table A.14: Atua summary table.

Timeline Information

Oct 26 Atua departs Suez.

Oct 31 Arrives in Auckland.

Nov 2 Departs Auckland. Between the Oct 31 - Nov 2, an individual brings infection on board as *Exposed* class.

Nov 3 First case observed.

Nov 8 Arrives in Sydney. 62 total cases have occurred by this time. Sick landed to hospital. Healthy landed to isolation on shore. Land 62 infected individuals. Begin quarantine transmission rate. Remove cases at onset.

Nov 22 1 cases observed. End of Cases. End of data.

Manuka

Ship Information

The *Manuka* was an New Zealand transport vessel measuring 112m by 14m [35]. It travelled from Wellington (New Zealand), where it brought the influenza virus on board, to Sydney (Australia).

Passengers Information

The passengers aboard the vessel totalled 203, broken down in Table A.15.

Crew	Number
Officers	7
Engineers	7
Deck Workers	13
Engine Workers	27
Providers	41
Total	95

Passengers	Number
First Class	63
Second Class	45
Total	108

Table A.15: Manuka passenger record.

Case Information

There was a total of 42 cases out of the 203 personal, with 32 crew and 9 passengers cases. The date of the first case is unknown but occurred between the Nov 7 - 10. The Manuka entered quarantine in Sydney on November 13. Upon entry into quarantine all sick were isolated to cabins with stewards due to the quarantine station being occupied. Assume quarantine transmission rate. All patients were landed by Nov 15. Land all cases to date on this day. These patients and 10 healthy passengers were inoculated with the Commonwealth coryza vaccine before supply ran out.

Ship	Source	Arrival Date	Days at sea	Cases before arrival	Cases after arrival	Deaths	Crew	Troop	Total
Manuka	New Zealand	Nov 13	6	23	19	1	95	108	203

Table A.16: Manuka summary table.

Timeline Information

Nov 7 Medic departs Wellington. Between the Nov 2 - 7, an individual brings infection on board as *Exposed* class.

Nov 13 Arrives in Sydney. 23 cases on board. No date of first infection. First observation.

Nov 15 4 observed cases. All cases landed. Land all infected individuals. Remove cases at onset.

Nov 21 3 new cases amongst contacts on quarantine station. Ignored. 0 observed cases. End of cases.

Nov 26 Ship released with healthy contacts. No new observed cases up to this date. End of data.

Medic

Ship Information

The *Medic* was an Australian transport vessel measuring 174m by 19m [35]. It travelled from Sydney (Australia) to Wellington (New Zealand), where it brought on board the influenza virus, before returning to Sydney.

Passengers Information

The passengers aboard the vessel totalled 989, broken down in Table A.17.

Crew	Number
Officers	10
Engineers	8
Deck Workers	36
Engine Workers	41
Providers	61
Total	156

Passengers	Number
First Class	4
Troops	829
Total	833

Table A.17: Medic passenger record.

Case Information

There was a total of 313 cases out of the 989 personal, broken down in Table A.18.

Personnel	Number	Cases	Percentage
Crew	156	52	33%
Passengers	4	4	100%
Australian Troops	670	166	24%
Italian Troops	159	91	57%
Total	989	313	32%

Table A.18: Medic case record.

The Medic entered quarantine in Sydney on Nov 13. The Medic began using inhalation of zinc-sulphate on Dec 3, after the epidemic was in decline and until extinction. Inoculation of troops was carried out on Nov 21. Isolation of cases to shore began on Nov 21 and was completed on Nov 27. After this date all new cases are removed at onset.

Ship	Source	Arrival Date	Days at sea	Cases before arrival	Cases after arrival	Deaths	Crew	Troop	Total
Medic	New Zealand	Nov 21	10	203	103	22	833	156	989

Table A.19: Medic summary table.

Timeline Information

Nov 2 Medic departs Sydney.

Nov 7 Arrives in Wellington. Between the Nov 7 - 11, an individual brings infection on board as *Exposed* class.

Nov 11 Departs Wellington. First 2 cases observed.

Nov 15 Arrives back at Wellington. No communication between shore and ship. Assume no infection brought onboard.

Nov 16 Departs Wellington.

Nov 21 Enters Sydney Head (Quarantine zone). Every person aboard ship inoculated. Begin quarantine transmission rate. 32 cot cases and 50 mild cases from troops landed into *Quarantine hospital*. Land 82 infected individuals.

Nov 22 24 cases observed + 11 (Italian) cases found later to occur on this day. Data records 35 cases. 14 cot cases and 43 mild cases landed to quarantine hospital. Land 57 infected individuals.

Nov 23 36 cases observed. 26 troops + 8 Italians = 34 landed. Land 34 healthy individuals.

Nov 24 8 observed cases. 151 Italians landed (8 remaining Italians previously landed). Land 151 healthy individuals.

Nov 25 11 observed cases. 38 cot cases landed. Land 38 infected individuals

Nov 27 108 cases landed. Land 108 infected individuals. End of landing period. 293 cases observed so far, but 285 cases landed. Some infected individuals still on board. Begin post-landing period of immediate landing of cases at onset. Remove cases at onset.

Nov 30 2nd round of inoculations. 128 troops left on ship. 128 remaining troops inconsistent with data. Assuming referring to only Australian troops then original 670 - 285 total cases landed - 185 troops landed (including Italians) - 6 cases landed at onset = 194 remaining. Ignore this comment and allow other data to determine remaining individuals. End of data. (Truncated to avoid unknown numbers of troops on board and inoculations altering the transmission rate.)

Dec 3 Begin zinc-sulphate inhalation.

Dec 4-11 Healthy crew and troops released as 7 days with no new cases.

Dec 12 - 20 A series of observed cases amongst troops onshore. Unclear if within troops at quarantine hospital or released troops developing after 7 day period.

Boonah

Ship Information

The *Boonah* was an Australian transport vessel measuring 138m by 18m [35]. Originally from Adelaide (Australia), it travelled from Fremantle (Australia) to Durban (South Africa), where it brought on board the influenza virus. It returned to Fremantle and underwent maritime quarantine measures, before journeying to Albany (Australia) and then back to Adelaide. Note, structure of the vessel did not allow for isolation between different sections or companies.

Passengers Information

The passengers aboard the vessel totalled 164 crew and 931 troops, total 1095. The state-of-origin distribution of the troops was recorded and compared to the distribution amongst cases in Table A.20. The uniform distribution of cases supports no latent immunity by region. Upon landing in Durban, no passengers was allowed on shore due to the influenza outbreak on land. 15 men escaped but were not allowed back on board. Some local naval officers were allowed on board, along with a large number of natives who restocked the coal and mixed with the troops. First case outbreak amongst troops.

	VIC	SA	NSW	WA
Onboard distribution	40%	22%	19%	19%
Cases distribution	43%	23%	18%	16%

Table A.20: Boonah passenger origin record.

Case Information

There was a total of 470 cases out of the 1095 personal, broken down by group in Table A.21. Only have data of 433 troop infections. Use troop population and troop cases. Ignore crew infections.

Personnel	Number	Cases	Percentage
Crew	164	37	22%
Troops	931	433	47%
Total	1095	470	43%

Table A.21: Boonah Case Record

Preventative measures were attempted early into outbreak, an unspecified number of days after leaving Durban. These include; use of zinc-sulphate inhalation chambers, thermometer parades individuals suffering headaches or sore throats. Also, some areas were attempted to be implemented as isolations areas, with little success in effective isolation. Assume standard transmission rate. The Boonah entered quarantine in Fremantle on Dec11. Assume quarantine transmission rate. Inoculation of troops was carried out on Dec 11, well after the peak of the outbreak. Landing of cases to shore began on Dec 11 and was completed on Dec 13. As ship later leaves for Albany then Adelaide, there are multiple landing periods. New cases are removed at onset where specified.

Ship	Source	Arrival Date	Days at sea	Cases before arrival	Cases after arrival	Deaths	Crew	Troop	Total
Boonah	South Africa	Dec 11	17	298	172	18	931	164	1095

Table A.22: Boonah summary table.

Timeline Information

Oct 29 Departs Fremantle.

Nov 16 Arrives in Durban.

Nov 24 Departs Durban. Between Nov 16 - 24, an individual brings infection on board as *Exposed* class.

Nov 29 Three cases observed. First observed cases.

Nov 30 - Dec 10 Epidemic builds to peak.

Dec 11 Arrives in Fremantle. Enters Quarantine with 298 cases on board. Troops aboard ship inoculated. Begin quarantine transmission rate. 150 patients landed to Fremantle quarantine station. Land 150 infected individuals. 29 new cases.

Dec 12 86 patients landed to Fremantle quarantine station. Land 86 infected individuals. 25 new cases.

Dec 13 Total cases at Fremantle 303. Removal of cases completed. Land remaining infected individuals. 150 + 86 = 236 cases landed previously, so 303 - 236 = 67 infected individuals landed on this day. Remove cases at onset until departure Fremantle. 13 new cases.

Dec 14 9 new cases.

Dec 15 8 new cases on ship.

Dec 16 7 new cases on ship and ashore. Ship and shore location of troops treated identically. Record in data as 9, 8 and 7 troop cases respectively.

Dec 17 4 new cases.

Dec 18 2 new cases.

Dec 19 No new cases. End of data. (Truncated to avoid ship leaving in quarantine with unknown population and future inconsistencies.)

Dec 20 No new cases. Departed for Albany. Convalescent healthy crew returned to ship. Contradiction with data; 2 new cases recorded and inconsistent population size after landing. See Dec 22 *"Arrives in Albany with 434 troops and 86 crew"*. Troop population - total cases up to leaving Fremantle = 931 - 397 = 534.

Dec 22 Arrives in Albany with 434 troops and 86 crew. 4 cases landed.

Dec 24 2 new cases in Fremantle. Ignore Fremantle cases. 7 cases are recorded in data.

Dec 25 1 new case. Not recorded in data. Departs Albany for Adelaide with 427 troops and 86 crew. Difference between arrival and departure populations shows 7 troops have been landed and left in Albany. Two new cases on Boonah.

Dec 28 1 new case on Boonah. Arrives in Adelaide with 14 new cases since Albany (13 troops, 1 crew). These cases landed upon arrival. 13 cases recorded on journey from Albany to Adelaide..

Dec 29 2 new cases discovered and removed at onset. All remaining troops landed into camp on Torrens Island, and new cases sent to hospital.

Dec 30 - Jan 7 Some new cases amongst troops at camp. Not recorded in data. End of Outbreak.

Nestor

Ship Information

The *Nestor* was an Australian vessel, approximately 175m by 20m, trading between the England and Australia [35]. The ship left London (England) before stopping at Post Said (Egypt), Suez (Egypt), Colombo (Sri Lanka) before arriving in Albany (Australia). The ship was quarantined in Albany for 3 days before travelling to Melbourne (Australia). It brought on board the influenza virus in London.

Passengers Information

The *Nestor* left London with 176 crew and 1,727 passengers for a total of 1,903 personal. The personal aboard were broken down in Table A.23.

Crew	Number
Officers	8
Engineers	10
Deck Workers	36
Engine Workers	64
Providers	58
Total	176

Passengers	Number
First Class	103
Second Class	78
Third Class	1,546
Total	1,727

Table A.23: Nestor passenger record.

All troops had received one dose of influenza vaccine prior to leaving England. Some received two and those who hadn't were administered a second one on commencement of the voyage.

Case Information

There was a total of 69 cases out of the 1,903 personal. No information regarding if crew or troop infections. The Nestor entered quarantine on Jan 18. Assume quarantine transmission rate. Use unique λ to account for prior inoculation. For a period of 3 days, zinc-sulphate inhalation and thermometer parades were conducted on the passengers. During this time 192 troops were landed. The vessel was granted leave after these 3 days on Jan 20, when it left for Melbourne. Upon landing in Melbourne, the vessel was quarantined for a further 7 days with zinc-sulphate inhalation and thermometer parades.

Ship	Source	Arrival Date	Days at sea	Cases before arrival	Cases after arrival	Deaths	Troop	Crew	Total
Nestor	England	Jan 18	37	69	0	0	833	1,727	1,903

Table A.24: Nestor summary table.

Timeline Information

Dec 12 Medic departs London. First case observed. Between Dec 7 - 12, an individual brings infection on board as *Exposed* class.

Dec 23 Arrives and departs from Port Said. Assume no infection introduced.

Dec 24 Arrives in Suez. Assume no infection introduced.

Dec 25 Departs Suez.

Jan 5 Arrives in Colombo. Assume no infection introduced. Two passengers taken on board. Assume to be susceptible.

Jan 7 Departs Colombo.

Jan 18 - 20 Arrives in Albany. Enters Quarantine. Begin quarantine transmission rate. Begin landing of 192 troops. No information of healthy or infected, assume healthy. Unspecified landing dates. Landed 192 healthy individuals on Jan 20.

Jan 20 - 24 Journeys to Melbourne. Continue quarantine transmission rate.

Jan 24 Arrives in Melbourne. Enters quarantines. Last Cases. Continue quarantine transmission rate. Remove cases at onset.

Jan 30 Detained in quarantine for 7 days. No new cases. No new observed cases over next 7 days. End of data.

Ceramic

Ship Information

The *Ceramic* was an Australian transport vessel measuring 200m by 21m [73]. It left London (England), stopped at Devonport (England), Port Said (Egypt), Suez (Egypt), Colombo (Sri Lanka), then arrived into quarantine in Albany (Australia). After quarantine, the ship sailed to Melbourne via Adelaide and Hobart. It brought on board the influenza virus in London or Devonport.

Passengers Information

The Ceramic left London with 246 crew and 2,115 troops, for total of 2361 passengers.

Case Information

There was a total of 194 cases out of the 2361 personal. No information regarding if crew or troop infections. The first case was recorded the day leaving Devonport and by arrival to Port Said (10 days later) there was 99 cases on board. 33 cases were disembarked at Port Said and 20 at Suez. Of these cases, 41 occurred post-entry to quarantine. Preventative measures were attempted early into outbreak, an unspecified number of days after leaving Durban. These include; use of zinc-sulphate inhalation chambers, thermometer parades of headache or sore throat symptomatic individuals.

Also, some areas were attempted to be implemented as isolations areas to little success. Assume standard transmission rate. Two secondary cases developed at Albany (one nurse and one soldier).

Ship	Source	Arrival Date	Days at sea	Cases before arrival	Cases after arrival	Deaths	Troop	Crew	Total
Ceramic	England	Mar 3	39	90	41	2	2,115	246	2,361

Table A.25: Ceramic summary table.

Timeline Information

Jan 22 Departs London.

Jan 23 Arrives in Devonport.

Jan 26 Departs Devonport. First observed case. Between Jan 21 - 26, an individual brings infection on board as *Exposed* class.

Feb 3 Arrives in Port Said. 33 cases landed during stay. Land 33 infected individuals. Assume landed on last day (5th). Assume no infection introduced.

Feb 5 Departs Port Said. Arrives in Suez. 20 cases landed during stay. Land 20 infected individuals. Assume landed on last day (9th). Assume no infection introduced.

Feb 9 Departs Suez.

Feb 18 Arrives in Colombo. Assume no infection introduced.

Feb 20 Departs Colombo.

Mar 3 Arrives in Albany. All Western Australians troops and 13 cases landed. Land 96 healthy and 13 infected individuals.

Mar 6 Arrives in Adelaide. 6 cases landed. 183 healthy south Australian troops landed. Land 183 healthy and 6 infected individuals.

Mar 8 Departs Adelaide.

Mar 9 Last case. End of cases..

Mar 11 Arrives in and Departs Hobart. Arrives with 246 crew and 1763 troops. Use to match number of troops landed; → 183 Western Australian troops landed. 9 cases landed at Hobart with onset dates. Land 9 infected individuals. No new observed cases between Mar 9 - 11. End of data. (Truncation due to unclear population aboard ship).

Mar 12 Arrives in Melbourne. 4 cases, 13 convalescents and 719 troops landed. First cases occur on Mar 16 (1) and Mar 17 (3), Not recorded. Population distribution and origin of cases unclear.

Lancashire

Ship Information

The *Lancashire* was an Australian transport vessel, approximately the same size as the *Marathon* (154m by 17m) [20, 76]. It travelled from Devonport (England) to Post Said (Egypt), Suez (Egypt), Colombo (Sri Lanka) before arriving in Fremantle (Australia). After Fremantle, it travelled to Adelaide, Hobart, Melbourne, Sydney and then Brisbane. It is thought to have brought on board the influenza virus in Devonport.

Passengers Information

The passengers aboard the vessel totalled 1,643, with 1,465 troops and 178 crew.

Case Information

There was a total of 53 cases out of the 1,643 personal. The Lancashire entered quarantine in Fremantle on Mar 14.

Ship	Source	Arrival Date	Days at sea	Cases before arrival	Cases after arrival	Deaths	Crew	Troop	Total
Lancashire	England	Mar 14	34	53	0	0	1465	178	1643

Table A.26: Lancashire summary table.

Timeline Information

Feb 7 First 5 cases. Sent ashore before sailing from Devonport. Between Feb 2 - 7, an individual brings infection on board as *Exposed* class. Observe 5 cases on this day. Land 5 infected individuals at end of day.

Feb 8 Lancashire departs Devonport.

Feb 17 Arrives in Port Said. Assume no infection brought on board.

Feb 18 Departs Port Said.

Feb 19 Arrives and Departs from Suez. Assume no infection brought on board.

Mar 1 Arrives in Colombo. Assume no infection brought on board.

Mar 3 Depart Colombo.

Mar 14 Arrives in Fremantle. Landed 132 healthy troops. Last observed case. Left same day. Begin quarantine transmission rate. Land 132 healthy individuals. End of cases.

Mar 19 Arrives in Adelaide. Land one case. Land one infected individual. No new observed cases between Mar 14 - 19. End of data.

Mar 22 Arrives in Hobart. Landed 46 unknown. Also landed one case.

Mar 24 Arrives in Melbourne with 176 crew, 1159 passengers.

Mar 15-29 Further travel around Australia with no new cases (2 suspected found to not be influenza).

Kenilworth Castle

Ship Information

The *Kenilworth Castle* was an Australian transport vessel measuring 174m by 19m [65]. It travelled from Liverpool (England), where it brought on board the influenza virus, to Albany passing through Madeira (Portugal), Cape Town (South Africa) and Durban (South Africa).

Passengers Information

The passengers aboard the vessel totalled 831 with 326 crew and 505 passengers.

Case Information

There was a total of 29 cases out of the 831 personal. Cases were amongst the passengers and the crew was not affected. Use passenger population size. The Kenilworth Castle entered quarantine in Sydney on April 1.

Ship	Source	Arrival Date	Days at sea	Cases before arrival	Cases after arrival	Deaths	Crew	Total	
Kenilworth Castle	England	Apr 1	46	29	0	0	326	505	831

Table A.27: Kenilworth Castle summary table.

Timeline Information

Feb 14 Kenilworth Castle departs Liverpool. Between Feb 14 - 9, an individual brings infection on board as *Exposed* class.

Feb 15 First 4 cases observed.

Feb 19 Arrives and departs Madeira. 2 cases observed. Assume no infection brought on board.

Mar 5 Arrives at Cape Town with 29 cases having occurred. 27 cases landed. Data shows 24 cases taken place so far. The 26th and final cases is observed on Mar 10. Assume landed 26 cases on Mar 10. The ship enters quarantine for 5 days. Assume no infection brought on board. Unknown quarantine measures in Cape Town. Assume standard transmission rate.

Mar 10 Last case observed. Land 26 infected individuals. End of cases.

Mar 12 Departs Cape Town.

Mar 15 - 18 Arrives in Durban.

Apr 1 Arrives in Albany. No new observed cases up to this date. End of data.

Orca

Ship Information

The *Orca* was an Australian transport vessel measuring 174m by 20m [73]. It brought on board the influenza virus in Liverpool (England) before travelling to Cape Town (South Africa), then Adelaide (Australia) and Melbourne (Australia).

Passengers Information

The passengers aboard the vessel totalled 1698, with 209 crew and 1489 troops.

Case Information

There was a total of 48 cases out of the 1698 personal. The Orca entered quarantine in Adelaide on Mar 29. The last case was observed the day before arrival into Adelaide.

Ship	Source	Arrival Date	Days at sea	Cases before arrival	Cases after arrival	Deaths	Troop	Crew	Total
Orca	England	Mar 29	19	48	0	0	1489	209	1698

Table A.28: Orca summary table.

Timeline Information

Feb 19 Orca departs Liverpool.

Feb 20 First cases. Between the Feb 15 - 19, an individual brings infection on board as *Exposed* class.

Mar 10 Arrives in Cape Town. Assume no infection brought on board.

Mar 12 Departs Cape Town.

Mar 28 Last case. End of cases.

Mar 29 Arrives in Adelaide. Land 155 troops, include 4 cases and one crew case. Land 151 healthy and 4 infected individuals.

Mar 30 Departs Adelaide.

Mar 31 Arrives in Melbourne. Land 484 non-specified troops at quarantine station. Assume healthy troops. Land 484 healthy individuals. No new observed cases up to arrival in Melbourne. End of Data. Some time later ship arrives into Sydney with remaining passengers. No cases upon arrival to Sydney and no further cases.

Kashmir

Ship Information

The *Kashmir* was an Australian transport vessel measuring 146m by 18m [76]. After bringing on board the influenza virus in Southhampton (England), it travelled via Cape Town (South Africa) to Adelaide (Australia) before reaching Hobart (Australia).

Passengers Information

The passengers aboard the vessel totalled 1500, 1281 troops and 219 crew.

Case Information

There was a total of 97 cases out of the 1500 personal. The Kashmir entered quarantine in Adelaide on April 17.

Ship	Source	Arrival Date	Days at sea	Cases before arrival	Cases after arrival	Deaths	Troop	Crew	Total
Kashmir	England	April 17	39	97	23	0	1281	219	1500

Table A.29: Kashmir summary table.

Timeline Information

Mar 9 Kashmir departs Southhampton.

Mar 12 First 5 cases. During Mar 7-9, an individual brings infection on board as *Exposed* class.

Mar 28 Arrives in Cape Town. Assume no infection brought on board.

Mar 31 Departs Cape Town.

Apr 17 Arrives in Adelaide. Begin quarantine transmission rate. 22 cases landed. Land 22 infected individuals. Assume landed on last day (19th).

Apr 19 Departs Adelaide.

Apr 21 Arrives and Departs Hobart.

Apr 23 Last case. End of cases. End of data. Unknown time later arrives in Melbourne.

Euripides

Ship Information

The *Euripides* was an Australian transport vessel measuring 167m by 20m [35]. After leaving London (the reported source-of-infection city) it visited Portland (Amsterdam), Port Said (Egypt), El Kantara (known as El-Qantara, Egypt), Suez (Egypt), Colombo (Sri Lanka) before arriving in Fremantle (Australia). After Fremantle it travelled to Adelaide, Hobart and then Melbourne.

Passengers Information

The passengers aboard the vessel totalled 1323, 1132 troop and 191 crew.

Case Information

There was a total of 53 cases out of the 1323 personal. There was an unrecorded first case the day after leaving London (Feb 28), which is why it is the recorded source of infection, but no further cases appeared until after leaving Egypt (left Suez on Mar 16) with next recorded case on Mar 17. Hence, source of infection far more likely to be Egypt, due to to the 3 weeks between observed cases. Assume Egypt is source of infection. The Euripides entered quarantine in Fremantle on Apr 10.

Ship	Source	Arrival Date	Days at sea	Cases before arrival	Cases after arrival	Deaths	Troop	Crew	Total
Euripides	England	Apr 10	42	53	0	0	1132	191	1323

Table A.30: Euripides summary table.

Timeline Information

Feb 27 Euripides departs London.

Feb 28 Arrives in Portland. Assume no infection brought on board. First unrecorded case. Assume single case was not cause of outbreak. Case ignored.

Mar 3 Departs Portland.

Mar 14 Arrives and departs Port Said. Possible infection brought on board.

Mar 15 Arrives El Kantara. Possible infection brought on board.

Mar 16 Departs El Kantara. Arrives and departs Suez. Possible infection brought on board.

Mar 17 First case recorded. Unknown if infection brought on board in Port Said, El Kantara or Suez. During Mar 14 - 16, an individual brings infection on board as *Exposed* class.

Mar 28 Arrives in Colombo. Assume no infection brought on board.

Mar 30 Departs Colombo. Assume no infection brought on board.

Apr 10 Arrives in Fremantle. Begin quarantine transmission rate.

Apr 14 Last case. End of cases.

Apr 15 Arrives in Adelaide.

Apr 18 Arrives in Hobart.

Apr 20 Arrives in Melbourne. Lands 357 into quarantine station. Land 357 healthy individuals. In quarantine until the Apr 25 with no new cases. No new observed cases up to this date. End of data.

Bibliography

[1] R. Albert, K. Ostheimer, and J. Breman. The last smallpox epidemic in boston and the vaccination controversy, 1901-1903, 2001.

[2] L. Allen. An introduction to stochastic epidemic models. In *Mathematical epidemiology*. Springer, 2008.

[3] L. Allen, F. Brauer, P. Van den Driessche, and J. Wu. *Mathematical epidemiology*. Springer, 2008.

[4] H. Andersson and T. Britton. *Stochastic Epidemic Models and Their Statistical Analysis*. Lecture Notes in Statistics. Springer New York, 2012.

[5] C. Andrieu and J. Thoms. A tutorial on adaptive mcmc. *Statistics and Computing*, 2008.

[6] C. Andrieu, A. Doucet, and R. Holenstein. Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2010.

[7] M. Bedard. Optimal acceptance rates for metropolis algorithms: Moving beyond 0.234. *Stochastic Processes and Their Applications*, 2008.

[8] M. Begon, M. Bennett, R. Bowers, N. French, S. Hazel, and J. Turner. A clarification of transmission terms in host-microparasite models: numbers, densities and areas. *Epidemiology and infection*, 2002.

[9] A. Black and J. Ross. Computation of epidemic final size distributions. *Journal of Theoretical Biology*, 2015.

[10] A. Black, N. Geard, J. McCaw, J. McVernon, and J. Ross. Characterising pandemic severity and transmissibility from data collected during first few hundred studies. *Epidemics*.

[11] M. Bootsma and N. Ferguson. The effect of public health measures on the 1918 influenza pandemic in u.s. cities. *Proceedings of the National Academy of Sciences*, 2007.

[12] Z. Botev, J. Grotowski, D. Kroese, *et al.* Kernel density estimation via diffusion. *The Annals of Statistics*, 2010.

[13] T. Britton. Stochastic epidemic models: A survey. *Mathematical Biosciences*, 2010.

[14] B. Carlin and T. Louis. *Bayesian methods for data analysis.* CRC Press, 2008.

[15] F. Carrat, E. Vergu, N. Ferguson, M. Lemaitre, S. Cauchemez, S. Leach, and A. Valleron. Time lines of infection and disease in human influenza: A review of volunteer challenge studies. *American Journal of Epidemiology*, 2008.

[16] G. Chowell and H. Nishiura. Transmission dynamics and control of ebola virus disease (evd): a review. *BMC medicine*, 2014.

[17] G. Chowell, C. Ammon, N. Hengartner, and J. Hyman. Estimation of the reproductive number of the spanish flu epidemic in geneva, switzerland. *Vaccine*, 2006.

[18] G. Chowell, M. Miller, and C. Viboud. Seasonal influenza in the united states, france, and australia: transmission and prospects for control. *Epidemiology and infection*, 2008.

[19] B. Coburn, B. Wagner, and S. Blower. Modeling influenza epidemics and pandemics: insights into the future of swine flu (h1n1). *BMC medicine*, 2009.

[20] J. Cumpston. *Influenza and maritime quarantine in Australia.* Melbourne : Albert J. Mullett, Government Printer, 1919.

[21] P. Del Moral. Feynman-kac formulae. In *Feynman-Kac Formulae.* Springer, 2004.

[22] A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice.* Information Science and Statistics. Springer, 2001.

[23] N. Ferguson, D. Cummings, C. Fraser, J. Cajka, P. Cooley, and D. Burke. Strategies for mitigating an influenza pandemic. *Nature*, 2006.

[24] M. Forrester. *Epidemic models and inference for the transmission of hospital pathogens.* PhD thesis, Queensland University of Technology, 2006.

[25] C. Fraser, S. Riley, R. Anderson, and N. Ferguson. Factors that make an infectious disease outbreak controllable. *Proceedings of the National Academy of Sciences of the United States of America*, 2004.

[26] C. Fraser, D. Cummings, D. Klinkenberg, D. Burke, and N. Ferguson. Influenza transmission in households during the 1918 pandemic. *American journal of epidemiology*, 2011.

[27] J. Geweke. Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities. In *Computing science and statistics: Proceedings of the 23rd symposium on the interface.* Citeseer, 1991.

[28] W. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in Practice.* Taylor & Francis, 1995.

[29] D. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 1977.

[30] A. Golightly and D. Wilkinson. Bayesian parameter inference for stochastic biochemical network models using particle markov chain monte carlo. *Interface focus*, 2011.

[31] N. Gordon, D. Salmond, and A. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. In *IEE Proceedings F (Radar and Signal Processing).* IET, 1993.

[32] H. Haario, E. Saksman, and J. Tamminen. An adaptive metropolis algorithm. *Bernoulli*, 2001.

[33] K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 1970.

[34] D. Henderson, F. Dunston, D. Fedson, V. Fulginiti, R. Gerety, F. Guerra, K. Johnson, E. Marcuse, D. Metzgar, and R. Saldarini. The measles epidemic: the problems, barriers, and recommendations. *Jama*, 1991.

[35] J. Hoskin. Flotilla australia. URL `http://www.flotilla-australia.com`.

[36] T. House and M. Keeling. Deterministic epidemic models with explicit household structure. *Mathematical biosciences*, 2008.

[37] Y. Hsieh, C. Tsai, C. Lin, J. Chen, C. King, D. Chao, and K. Cheng. Asymptomatic ratio for seasonal h1n1 influenza infection among schoolchildren in taiwan. *BMC Infectious Diseases*, 2014.

[38] T. Jahnke and V. Sunkara. *Error Bound for Hybrid Models of Two-Scaled Stochastic Reaction Systems*. Springer International Publishing, 2014.

[39] G.t Jenkinson and J. Goutsias. Numerical integration of the master equation in some models of stochastic epidemiology. *PloS One*, 2012.

[40] N. Johnson. *Britain and the 1918-19 influenza pandemic: a dark epilogue*. Routledge, 2006.

[41] M. Keeling and P. Rohani. *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, 2008.

[42] M. Keeling and J. Ross. On methods for studying stochastic disease dynamics. *Journal of The Royal Society Interface*, 2008.

[43] W. Kermack and A. McKendrick. A contribution to the mathematical theory of epidemics. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 1927.

[44] D. Kroese, T. Taimre, and Z. Botev. *Handbook of Monte Carlo Methods*. Wiley Series in Probability and Statistics. Wiley, 2013.

[45] T. Kurtz. Solutions of ordinary differential equations as limits of pure jump markov processes. *Journal of applied Probability*, 1970.

[46] T. Kurtz. Limit theorems for sequences of jump markov processes approximating ordinary differential processes. *Journal of Applied Probability*, 1971.

[47] A. Lloyd and R. May. Spatial heterogeneity in epidemic models. *Journal of theoretical biology*, 1996.

[48] J. Mathews, C. McCaw, J. McVernon, E. McBryde, and J. McCaw. A biological model for influenza transmission: Pandemic planning implications of asymptomatic infection and immunity. *PLoS ONE*, 2007.

[49] J. Mathews, E. McBryde, J. McVernon, P. Pallaghy, and J. McCaw. Prior immunity helps to explain wave-like behaviour of pandemic influenza in 1918-19. *BMC infectious diseases*, 2010.

[50] T. McKinley, A. Cook, and R. Deardon. Inference in epidemic models without likelihoods. *The International Journal of Biostatistics*, 2009.

[51] T. McKinley, J. Ross, R. Deardon, and A. Cook. Simulation-based bayesian inference for epidemic models. *Computational Statistics & Data Analysis*, 2014.

[52] G. Mercer, K. Glass, and N. Becker. Effective reproduction numbers are commonly overestimated early in a disease outbreak. *Statistics in medicine*, 2011.

[53] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 1953.

[54] C. Mills, J. Robins, and M. Lipsitch. Transmissibility of 1918 pandemic influenza. *Nature*, 2004.

[55] C. Moler and C. Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM review*, 2003.

[56] D. Morens, J. Taubenberger, and A. Fauci. The persistent legacy of the 1918 influenza virus. *New England Journal of Medicine*, 2009.

[57] B. Munsky and M. Khammash. The finite state projection algorithm for the solution of the chemical master equation. *The Journal of chemical physics*, 2006.

[58] J. Norris. *Markov chains.* Cambridge university press, 1998.

[59] Joint United Nations Programme on HIV/AIDS *et al. 2008 report on the global AIDS epidemic.* Unaids, 2008.

[60] E. Patrozou and L. Mermel. Does influenza transmission occur from asymptomatic infection or prior to symptom onset? *Public health reports*, 2009.

[61] H. Phillips and D. Killingray. *The Spanish influenza pandemic of 1918-19: new perspectives.* JSTOR, 2003.

[62] M. Pitt, R. Silva, P. Giordani, and R. Kohn. Auxiliary particle filtering within adaptive metropolis-hastings sampling. *arXiv preprint arXiv:1006.1914*, 2010.

[63] A. Plourde and E. Bloch. A literature review of zika virus. *Emerging infectious diseases*, 2016.

[64] M. Porta. *A dictionary of epidemiology.* Oxford university press, 2014.

[65] Wartime Memories Project. Wartime memories project. URL `http://www.wartimememoriesproject.com`.

[66] C. Robert. *Monte carlo methods.* Wiley Online Library, 2004.

[67] K. Roberts, H. Shelton, P. Stilwell, and W. Barclay. Transmission of a 2009 h1n1 pandemic influenza virus occurs before fever is detected, in the ferret model. *PLoS ONE*, 2012.

[68] S. Ross. *Introduction to probability models.* Academic press, 2014.

[69] P. Selby. *Influenza models: prospects for development and use.* Springer Science & Business Media, 2012.

[70] G. Sertsou, N. Wilson, M. Baker, P. Nelson, and M. Roberts. Key transmission parameters of an institutional outbreak during the 1918 influenza pandemic estimated by mathematical modelling. *Theoretical Biology and Medical Modelling*, 2006.

[71] R. Sidje. Expokit: a software package for computing matrix exponentials. *ACM Transactions on Mathematical Software (TOMS)*, 1998.

[72] A. Smith and A. Gelfand. Bayesian statistics without tears: a sampling–resampling perspective. *The American Statistician*, 1992.

[73] B. Solem. Norway heritage. URL `http://www.norwayheritage.com/`.

[74] J. Taubenberger and D. Morens. 1918 influenza: the mother of all pandemics. 2006.

[75] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. Stumpf. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 2009.

[76] Caledonian Maritime Research Trust. Clyde-built ship database. URL `http://www.clydeships.co.uk`.

[77] J. Van-Tam. *Pandemic Influenza.* CABI, 2012.

[78] R. Webster, A. Monto, T. Braciale, and R. Lamb. *Textbook of influenza.* John Wiley & Sons, 2014.

[79] Y. Yang, J. Sugimoto, E. Halloran, N. Basta, D. Chao, L. Matrajt, G. Potter, E. Kenah, and I. Longini. The transmissibility and control of pandemic influenza a (h1n1) virus. *Science*, 2009.