ACCEPTED VERSION

Nicolas P. Rebuli, N.G. Bean, J.V. Ross
Estimating the basic reproductive number during the early stages of an emerging epidemic
Theoretical Population Biology, 2018; 119:26-36

© 2017 Elsevier Inc. All rights reserved.

This manuscript version is made available under the CC-BY-NC-ND 4.0 license
http://creativecommons.org/licenses/by-nc-nd/4.0/

Final publication at http://dx.doi.org/10.1016/j.tpb.2017.10.004

PERMISSIONS

https://www.elsevier.com/about/our-business/policies/sharing

Accepted Manuscript

Authors can share their accepted manuscript:

[12 months embargo]

After the embargo period

- via non-commercial hosting platforms such as their institutional repository
- via commercial sites with which Elsevier has an agreement

In all cases accepted manuscripts should:

- link to the formal publication via its DOI
- bear a CC-BY-NC-ND license – this is easy to do
- if aggregated with other manuscripts, for example in a repository or other site, be shared in alignment with our hosting policy
- not be added to or enhanced in any way to appear more like, or to substitute for, the published journal article

24 June 2020

http://hdl.handle.net/2440/116603

Estimating the basic reproductive number during the early stages of an emerging epidemic

Nicolas P. Rebuli[a,b,*], N. G. Bean[a,b], J. V. Ross[a,b]

[a] *Stochastic Modelling and Operations Research Group, School of Mathematical Sciences, University of Adelaide*
[b] *Australia Research Council Centre of Excellence for Mathematical and Statistical Frontiers*

Abstract

A novel outbreak will generally not be detected until such a time that it has become established. When such an outbreak is detected, public health officials must determine the potential of the outbreak, for which the *basic reproductive number* R_0 is an important factor. However, it is often the case that the resulting estimate of R_0 is positively-biased for a number of reasons. One commonly overlooked reason is that the outbreak was not detected until such a time that it had become established, and therefore did not experience initial fade out. We propose a method which accounts for this bias by conditioning the underlying epidemic model on becoming established and demonstrate that this approach leads to a less-biased estimate of R_0 during the early stages of an outbreak. We also present a computationally-efficient approximation scheme which is suitable for large data sets in which the number of notified cases is large. This methodology is applied to an outbreak of pandemic influenza in Western Australia, recorded in 2009.

Keywords: Basic reproductive number, Continuous-time Markov chain, Hybrid discrete-continuous

1. Introduction

Obtaining an accurate and reliable estimate of the basic reproductive number R_0, during the early stages of an outbreak is crucial for public health officials. The basic reproductive number characterises the transmission potential of a disease which is vital for predicting the size of the outbreak and the resources required for fighting it (Simonsen et al., 1997; Meltzer et al., 1999;

*Corresponding author
Email address: nicolas.rebuli@adelaide.edu.au (Nicolas P. Rebuli)

Lemon et al., 2007). Under these circumstances, R_0 is typically estimated from data relating to the daily number of new cases of the disease over a time period of just a few weeks (White and Pagano, 2007; Bettencourt and Ribeiro, 2008) which is often heavily influenced by incomplete reporting, population heterogeneity, and imported infectious cases (Mercer et al., 2011). Furthermore, estimation methods often over-look the probability of initial fade out during the early stages of the outbreak. This is an important factor because the fact that the outbreak is identified as a threat by authorities requires that it has effectively overcome the probability of initial fade out. Each of these factors contribute to positively-biased estimates of R_0 unless they are appropriately accommodated (Roberts and Nishiura, 2011; Nishiura et al., 2010; Pedroni et al., 2010; Mercer et al., 2011).

The impact of the probability of initial fade out on the estimate of R_0 is relatively unexplored. In the context of the Susceptible–Exposed–Infectious–Removed (SEIR) epidemic model, Mercer et al. (2011) demonstrated that estimates of R_0 from the initial stages of an outbreak are positively-biased and that this bias decreases as the outbreak progresses. This observation supports the hypothesis that the bias is at least partially influenced by the probability of initial fade out of the outbreak, which is considerable during the initial stages of an outbreak. Given that the outbreak is known to have eventually become established, this bias may be counteracted by *conditioning* the model on the event that an "established outbreak" occurs (Mercer et al., 2011; Rida, 1991).

In this paper, we present a conditioned Susceptible–Infectious–Removed (SIR) *continuous-time Markov chain* (CTMC) model which accounts for the probability of initial fade out (Bartlett, 1949). This is achieved by conditioning the usual SIR CTMC on the event that the outbreak eventually becomes established by modifying its transition rates according to Waugh (1958). We argue that it is reasonable to consider an established outbreak to be one where the cumulative number of cases eventually exceeds a predetermined *threshold*. Under this construction, we demonstrate that conditioning the SIR CTMC on the event that the outbreak eventually exceeds 50 cases reduces the resulting estimate of R_0 by around 0.2 on average.

Fundamental to estimating the basic reproductive number is the *likelihood function* of the CTMC model (Sprott, 2000). Computationally-exact methods for evaluating the likelihood function of a CTMC are typically computationally infeasible, even for moderate population sizes, hence it is common to consider the various ways in which the likelihood may be approximated (Cooper and Lipsitch, 2004). One important approximation for inference in large populations utilises the *diffusion approximation* of the CTMC (Ross et al., 2006, 2009; Ross, 2012). The diffusion approximation

provides a highly efficient and accurate approximation of the CTMC once the outbreak is established, but provides a poor approximation during the initial stages of the outbreak (Kurtz, 1970, 1971; Barbour, 1980), which Viboud et al. (2016) demonstrated is crucial for faithfully representing the early growth dynamics of emerging outbreaks. An alternative to the diffusion approximation which accounts for this is a *hybrid approximation* similar to Rebuli et al. (2016); Barbour (1975); Scalia-Tomba (1985); Safta et al. (2015). The hybrid approximation used here models the initial stages of the outbreak with a CTMC, until such a time that the population of infectious individuals is large enough for the diffusion approximation to provide a reliable approximation of the CTMC. We demonstrate that this hybrid approximation is highly accurate and provides a significant computational advantage over the CTMC model for large data sets.

We demonstrate the utility of our methodology by applying it to an outbreak of pandemic influenza from 2009 (A(H1N1)pdm09) which occurred in Western Australia (WA) (Kelly et al., 2010). During this outbreak, a thorough case ascertainment and follow-up program was conducted during the first three weeks of the outbreak until such a time that the outbreak was deemed widespread, by which stage 102 cases had been confirmed. Using the simple SIR CTMC, we demonstrate that estimates of R_0 which account for this fact are more accurate during the early stages of the outbreak.

The present paper has two objectives. The first is to present an approach for reducing the systematic bias in estimates of R_0 which are based on daily incidence counts from the initial stages of an outbreak. The second is to motivate a computationally-efficient algorithm for computing these estimates using a hybrid approximation of the underlying CTMC model. These concepts are straightforward to implement and can be generalised to more complex epidemiological models. We demonstrate the utility of our methodology by using it to estimate R_0 from an outbreak of pandemic influenza.

2. Background theory

The SIR CTMC is a population process which tracks the number of individuals in each of the susceptible (S), infectious (I) and removed (R) compartments, in a fixed population of N individuals (Keeling et al., 2000; Kermack and McKendrick, 1927). Using the relation $S + I + R = N$, the population process is completely described by the vector (S, I) where $S + I \leq N$ and $S, I \geq 0$.

For reasons which will become clear soon, it is more convenient to work with the *degree of advancement* (DA) representation of the SIR CTMC, rather than the population representation (Jenkinson and Goutsias, 2012; Rebuli et al., 2016; Black and Ross, 2015). The DA process $\{\boldsymbol{N}(t) : t \geq 0\}$ is a counting process which tracks the number of infection events (N_I) and the number of recovery events (N_R) which belong to the state space $\mathcal{S} = \{(N_I, N_R) : N_I \geq N_R, N_I, N_R = 0, 1, \ldots, N\}$. For the SIR model considered herein, we can map between representations using the relationships

$$N_I = N - S - I(0) - R(0), \qquad N_R = N - S - I - R(0).$$

The events and transition rates of the DA process are given in Table 1, where β is the effective transmission rate and γ is the recovery rate ($1/\gamma$ is the average infectious period). The basic reproductive number is $R_0 = \beta/\gamma$, which is defined as the average number of new cases of the disease, resulting from a single infectious individual in a completely susceptible population. The DA process is completely specified by its transition rates and an initial *probability mass function* (PMF) $p_{\boldsymbol{n}}(0) = \Pr(\boldsymbol{N}(0) = \boldsymbol{n})$, $\forall \boldsymbol{n} \in \mathcal{S}$, for which we assume $p_{(1,0)}(0) = 1$, herein.

Event	Transition	Rate
Infection	$\boldsymbol{n} \to \boldsymbol{n} + \boldsymbol{e}_1$	$q(\boldsymbol{n}, \boldsymbol{n} + \boldsymbol{e}_1) = \beta(S(0) - N_I)(I(0) + N_I - N_R)/(N - 1)$
Recovery	$\boldsymbol{n} \to \boldsymbol{n} + \boldsymbol{e}_2$	$q(\boldsymbol{n}, \boldsymbol{n} + \boldsymbol{e}_2) = \gamma(I(0) + N_I - N_R)$

Table 1: Events and transmission rates of the DA process where $\boldsymbol{e}_i$ is a vector of zeros with a one in the ith entry.

Define the *transition probability* $p_{\boldsymbol{m},\boldsymbol{n}}(t) = \Pr(\boldsymbol{N}(t) = \boldsymbol{n} | \boldsymbol{N}(0) = \boldsymbol{m})$, $\forall \boldsymbol{n}, \boldsymbol{m} \in \mathcal{S}$ as the probability that the DA process is in the state $\boldsymbol{n}$, given that $t > 0$ time units have elapsed since the process was in the state $\boldsymbol{m}$. Then $\forall \boldsymbol{n}, \boldsymbol{m} \in \mathcal{S}$, the transition probabilities are governed by the *Kolmogorov Forward Equations*

$$\frac{dp_{\boldsymbol{m},\boldsymbol{n}}(t)}{dt} = \sum_{i=1}^{2} p_{\boldsymbol{m},\boldsymbol{n}-\boldsymbol{e}_i}(t)\, q(\boldsymbol{n} - \boldsymbol{e}_i, \boldsymbol{n}) - p_{\boldsymbol{m},\boldsymbol{n}}(t)\, q(\boldsymbol{n}, \boldsymbol{n} + \boldsymbol{e}_i),$$

(Keeling et al., 2000; Norris, 1997; Jenkinson and Goutsias, 2012). The transition probabilities are calculated by integrating the Forward Equations numerically using the Implicit Euler scheme. Under the DA representation, Jenkinson and Goutsias (2012) showed that the implicit Euler scheme is highly computationally-efficient and achieves a global L_1-error of $\mathcal{O}(\tau)$, where τ is the length of the time step of the numerical integration.

Approach to estimation. We can now specify the likelihood of observing a set of daily incidence counts $x_k\,(k = 1, 2, \ldots, n)$, given a set of parameters $\theta \in \Theta$. An additional benefit of the DA representation is that the daily incidence counts may be expressed directly in terms of N_I by considering the cumulative incidence counts $y_k = \sum_{j=1}^{k} x_j\,(k = 1, 2, \ldots, n)$. It follows that the *exact likelihood* is

$$L(y|\theta) = \prod_{k=1}^{n} L_E^k(\theta),$$

in which the transition probabilities $L_E^k(\theta)\,(k = 1, 2, \ldots, n)$ are defined as

$$\begin{aligned}
L_E^k(\theta) &= \Pr\left(N_I(t_k) = y_k \,|\, \mathcal{Y}_{k\text{-}1}\right) \\
&= \sum_{i=0}^{y_k} \Pr\left(\boldsymbol{N}(t_k) = (y_k, i) \,|\, \mathcal{Y}_{k\text{-}1}\right) \\
&= \sum_{j=0}^{y_{k\text{-}1}} \sum_{i=0}^{y_k} \Pr\left(\boldsymbol{N}(t_k) = (y_k, i) \,|\, \boldsymbol{N}(t_{k\text{-}1}) = (y_{k\text{-}1}, j)\right) \Pr\left(\boldsymbol{N}(t_{k\text{-}1}) = (y_{k\text{-}1}, j) \,|\, \mathcal{Y}_{k\text{-}1}\right) \\
&= \sum_{j=0}^{y_{k\text{-}1}} \sum_{i=0}^{y_k} \Pr\left(\boldsymbol{N}(t_k) = (y_k, i) \,|\, \boldsymbol{N}(t_{k\text{-}1}) = (y_{k\text{-}1}, j)\right) \left(\frac{\Pr\left(\boldsymbol{N}(t_{k\text{-}1}) = (y_{k\text{-}1}, j) \,|\, \mathcal{Y}_{k\text{-}2}\right)}{\Pr\left(\mathcal{Y}_{k\text{-}1} \,|\, \mathcal{Y}_{k\text{-}2}\right)}\right) \\
&= \sum_{j=0}^{y_{k\text{-}1}} \sum_{i=0}^{y_k} \Pr\left(\boldsymbol{N}(t_k) = (y_k, i) \,|\, \boldsymbol{N}(t_{k\text{-}1}) = (y_{k\text{-}1}, j)\right) \left(\frac{\Pr\left(\boldsymbol{N}(t_{k\text{-}1}) = (y_{k\text{-}1}, j) \,|\, \mathcal{Y}_{k\text{-}2}\right)}{L_E^{k\text{-}1}(\theta)}\right),
\end{aligned}$$

where $\mathcal{Y}_{k\text{-}1} = \{N_I(t_{k\text{-}1}) = y_{k\text{-}1}, N_I(t_{k\text{-}2}) = y_{k\text{-}2}, \ldots, N_I(t_0) = y_0\}\,(k = 1, 2, \ldots, n)$ is the *history* of the outbreak. Each transition probability captures the probability of observing y_k infection events by the end of the kth day, given the history of the epidemic leading up to the start of the kth day, which describes a recurrence relation that is initialised by assuming the initial state $\boldsymbol{N}(0) = (1, 0)$.

The dependence of the likelihood on the underlying parameters θ is made explicit because the likelihood is used for estimating the parameters of the underlying CTMC (Sprott, 2000). A commonly used estimate is the *maximum likelihood estimate* (MLE), defined as the set of parameters which maximise the likelihood over the parameter space Θ.

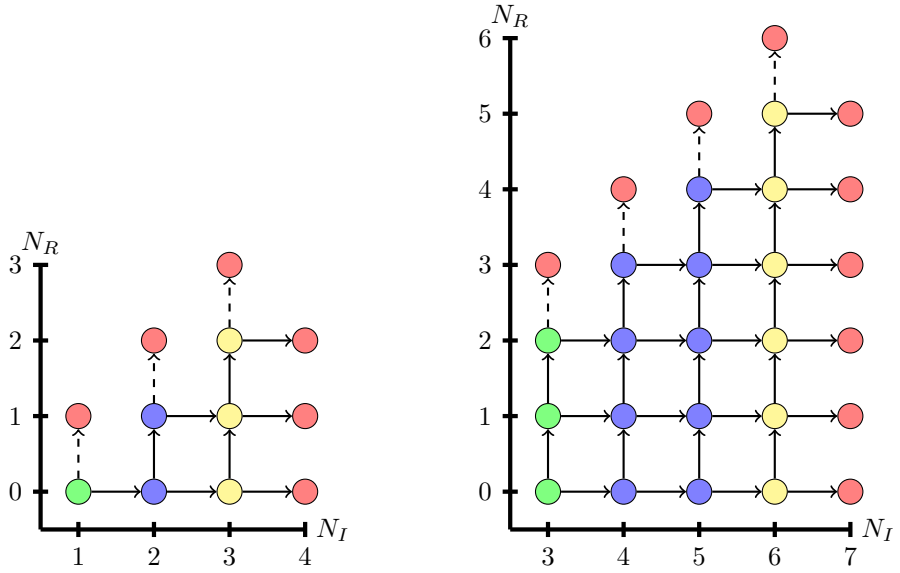
A common alternative is to adopt a Bayesian framework in which the parameters θ are treated as unknown random variables for which we seek their *posterior distribution*

$$f(\theta|y) \propto L(y|\theta) f(y),$$

where $f(\theta)$ represents our *prior* knowledge of the parameters. This is achieved herein by generating random samples from the posterior distribution using the Metropolis–Hastings Markov Chain Monte

Carlo algorithm, which are used to infer the distributional properties of the posterior (Gilks, 2005). We obtain a point estimate for θ from the Metropolis–Hastings algorithm by taking the median of the samples from the posterior distribution, commonly referred to as the *median posterior estimate* (MPE).

Illustrative example. We now provide an example to demonstrate an efficient algorithm for calculating the likelihood. Suppose an outbreak begins with two infections on the first day, and three on the second. Assuming that the outbreak started with a single infectious case, the basic reproductive number is estimated from the cumulative incidence counts $y_0 = 1$, $y_1 = 3$ and $y_2 = 6$ using the exact likelihood $L(y|\theta) = L_E^1(\theta)L_E^2(\theta)$, given a set of parameters $\theta = (\beta, \gamma)$, by sequentially calculating $L_E^1(\theta)$ and $L_E^2(\theta)$.



(a) State transition diagram for calculating the probability that $N_I(1) = 3$, assuming the initial state $N_I(0) = 1$.

(b) State transition diagram for calculating the probability that $N_I(2) = 6$, given $N_I(1) = 3$ and $N_I(0) = 1$.

Figure 1: Example of how the exact likelihood is calculated, using the data set $x_1 = 2$ and $x_2 = 3$.

The transition probability $L_E^1(\theta)$ is defined as the probability of observing three infection events in the CTMC model by time $t = 1$, assuming $\boldsymbol{N}(0) = (1, 0)$. Since N_I is monotonically increasing, the computational effort of this calculation can be reduced by truncating the state space to contain only states with $1 \leq N_I \leq 4$. The resulting state space is shown in Figure 1a, in which the green

state is the initial state, the yellow states are states with $N_I = 3$, the blue states are ordinary transient states, and the red states are absorbing states. The absorbing states with $N_I = N_R$ are extinction states which we will later condition $\boldsymbol{N}(t)$ on never reaching, hence transition into these states is denoted by a dashed arrow. It follows that the transition probability $L_E^1(\theta)$ is obtained by evolving the distribution of $\boldsymbol{N}(t)$ from time $t = 0$ to time $t = 1$ using the Kolmogorov forward equations, and then adding up the probability that $\boldsymbol{N}(1)$ is in any of the yellow states.

We now seek the transition probability $L_E^2(\theta)$, which is defined as the probability that $N_I(2) = 6$, given the history $\mathcal{Y}_1 = \{N_I(0) = 1, N_I(1) = 3\}$. In order to consider $\boldsymbol{N}(t)$ conditioned on the event $\mathcal{Y}_1$ for $t \geq 1$, the distribution of $\boldsymbol{N}(1)$ must be conditioned on being in the set of the yellow states in Figure 1a, which is given by

$$\Pr\left(\boldsymbol{N}(1) = (3, i) \,|\, \mathcal{Y}_1\right) = \frac{p_{(1,0),(3,i)}(1)}{L_E^1(\theta)}, \qquad \text{for } i = 0, 1, 2.$$

To calculate $L_E^2(\theta)$ we consider the state space truncation containing all states in $\mathcal{S}$ with $3 \leq N_I \leq 7$. This is shown by Figure 1b, in which the initial distribution across the green states is provided by $\boldsymbol{N}(1)|\mathcal{Y}_1$, and the yellow states denote states with $N_I = 6$. It follows that the transition probability $L_E^2(\theta)$ is obtained by evolving the distribution of $\boldsymbol{N}(t)|\mathcal{Y}_1$ from time $t = 1$ to time $t = 2$ using the forward equations, and then adding up the probability assigned to each of the yellow states.

The exact likelihood may now be computed from the product of the transition probabilities $L_E^1(\theta)$ and $L_E^2(\theta)$. This algorithm may be extended to include more days of observations by generalising the procedure for calculating $L_E^2(\theta)$.

A computationally-efficient approach. It is often computationally infeasible to evaluate the exact likelihood for a large population of individuals. However, under these circumstances the diffusion approximation of the underlying CTMC provides a computationally-efficient approach for approximating the exact likelihood (Ross et al., 2006, 2009; Ross, 2012). In the following discussion we provide a brief outline of how this methodology is applied to the class of *density dependent* CTMCs. For more detail, see Kurtz (1970, 1971); Ethier and Kurtz (2008).

The main technical requirement of the asymptotic approximations of Kurtz (1970, 1971) is that the CTMC is *density dependent*. A CTMC is density dependent if its transition rates can be written in the form $q(\boldsymbol{n}, \boldsymbol{n} + \ell) = \nu f(\boldsymbol{n}/\nu, \ell)$, $\forall \boldsymbol{n}, \boldsymbol{n} + \ell \in \mathcal{S}$ for a suitable function f and a *scaling constant* $\nu \in \mathbb{R}$, taken to be the population ceiling N, herein. The asymptotic limits then refer to the *density process* $\boldsymbol{N}_N(t) = \boldsymbol{N}(t)/N$ which depends on the current state $\boldsymbol{n}$ only through the density $\tilde{\boldsymbol{n}} = \boldsymbol{n}/N$,

for which $\tilde{\boldsymbol{n}} \in E$ and $E = [0,1]^2$. In a slightly more general definition of density dependence, the function f is realised asymptotically (for large N), see Pollett (1990).

The first of these asymptotic approximations is the *deterministic approximation* (Kurtz, 1970) which describes the mean trajectory of the scaled process $\boldsymbol{N}_N(s)$ over a finite time interval. The deterministic approximation $\boldsymbol{n}(t) \in E$ is the unique solution to the system of *ordinary differential equations* (ODEs) $d\boldsymbol{n}(t)/dt = F(\boldsymbol{n}(t))$ where $F(\tilde{\boldsymbol{n}}) = \sum_\ell f(\tilde{\boldsymbol{n}}, \ell)$, provided $\boldsymbol{n}(0) = \boldsymbol{N}(0)/N$. The second approximation is the diffusion approximation (Kurtz, 1971) which describes the fluctuations of the density process about its deterministic approximation. The centred diffusion approximation $\boldsymbol{Z}(t) = \sqrt{N}\,(\boldsymbol{N}_N(t) - \boldsymbol{n}(t))$ is a Gaussian diffusion process with expected value $\mathbf{0}$ and covariance matrix $\Sigma(t)$, given by the unique solution to the system of ODEs

$$\frac{d\Sigma(t)}{dt} = B(t)\Sigma(t) + \Sigma(t)B^T(t) + G(t), \qquad \Sigma(0) = \mathbf{0},$$

where $B(t) = \nabla F(\tilde{\boldsymbol{n}}(t))$ and $[G(t)]_{i,j} = \sum_\ell \ell_i \ell_j f(\tilde{\boldsymbol{n}}(t), \ell)$. It follows that the diffusion approximation of the population process $\boldsymbol{N}(t)$ is a Gaussian diffusion process with mean value $N\boldsymbol{n}(t)$ and
130 covariance matrix $N\Sigma(t)$. We now ground these ideas by applying them to the DA process.

The DA process is density dependent with the relevant functions

$$\begin{aligned} f(\tilde{\boldsymbol{n}}, \boldsymbol{e}_1) &= \beta(s_0 - n_I)(i_0 + n_I - n_R), \\ f(\tilde{\boldsymbol{n}}, \boldsymbol{e}_2) &= \gamma(i_0 + n_I - n_R), \end{aligned}$$

where $s_0 = S(0)/N$ and $i_0 = I(0)/N$. As such, the density process $\boldsymbol{N}_N(t)$ is a CTMC taking values $\tilde{\boldsymbol{n}} = \boldsymbol{n}/N$ for all $\boldsymbol{n} \in \mathcal{S}$. The density n_I denotes the proportion of individuals who have been infected and the density n_R denotes the proportion of individuals who have recovered.

The deterministic approximation of the density DA process, $\tilde{\boldsymbol{n}}(t) \in E$, is the unique solution to the system of ODEs

$$\begin{aligned} \frac{dn_I}{dt} &= \beta(s_0 - n_I)(i_0 + n_I - n_R), \\ \frac{dn_R}{dt} &= \gamma(i_0 + n_I - n_R), \end{aligned}$$

provided $\tilde{\boldsymbol{n}}(0) = (N_I(0)/N, N_R(0)/N)$. The fluctuations of the density process about the deterministic trajectory $\tilde{\boldsymbol{n}}(t)$ are captured by the centred diffusion approximation $\boldsymbol{Z}(t)$ which is a Gaussian diffusion process with mean $\mathbf{0}$ and covariance matrix $\Sigma(t) = (\sigma_{i,j}(t) : i, j = 1, 2)$ whose elements

are the unique solutions to the system of ODEs

$$\begin{aligned}
\frac{d\sigma_1}{dt} &= 2\beta\sigma_1\left(s_0 - i_0 + n_R - 2n_I\right) - 2\beta\sigma_{1,2}\left(s_0 - n_I\right) + \beta\left(s_0 - n_I\right)\left(i_0 + n_I - n_R\right), \\
\frac{d\sigma_{1,2}}{dt} &= \gamma\left(\sigma_1 - \sigma_{1,2}\right) + \beta\sigma_{1,2}\left(s_0 - i_0 + n_R - 2n_I\right) - \beta\sigma_2\left(s_0 - n_I\right), \\
\frac{d\sigma_2}{dt} &= \gamma\left(i_0 + n_I - n_R + 2\sigma_{1,2} - 2\sigma_2\right).
\end{aligned}$$

Therefore, a working approximation of the DA process $\boldsymbol{N}(t)$ is a Gaussian diffusion process with mean $N\tilde{\boldsymbol{n}}(t)$ and covariance matrix $N\Sigma(t)$. Hence, the diffusion approximation of the transition probability $p_{\boldsymbol{m},\boldsymbol{n}}(\theta\,;t)$ is the *transition density* $f_N(\theta\,;\boldsymbol{n},\boldsymbol{m},t)$, $\forall \boldsymbol{n},\boldsymbol{m} \in \mathcal{S}$, which is given by

$$f_N(\theta;\boldsymbol{n},\boldsymbol{m},t) = \frac{1}{2\pi N\sqrt{|\Sigma(t)|}} \exp\left(-\frac{1}{2}\left(\boldsymbol{m}/N - \tilde{\boldsymbol{n}}(t)\right)^T \Sigma^{-1}(t)\left(\boldsymbol{m}/N - \tilde{\boldsymbol{n}}(t)\right)\right).$$

We are now able to write down the *diffusion likelihood* as

$$L(y|\theta) = \prod_{k=1}^{n} L_D^k(\theta)$$

in which the transition probabilities are replaced by the transition densities $L_D^k(\theta)\,(k = 1, 2, \ldots, n)$ defined by

$$L_D^k(\theta) = \sum_{j=0}^{y_{k\text{-}1}} \sum_{i=0}^{y_k} f_N\left(\theta\,;(y_{k\text{-}1}, j),(y_k, i), 1\right) \Pr\left(\boldsymbol{N}(t_{k\text{-}1}) = (y_{k\text{-}1}, j)\,|\,\mathcal{Y}_{k\text{-}1}\right).$$

The diffusion likelihood is calculated by using the transition densities as a crude midpoint approximation to the transition probabilities. In the context of Figure 1a, the transition probability $L_D^1(\theta)$ is an approximation of $L_E^1(\theta)$ in which the probability of each of the yellow states is approximated by the transition densities $f_N(\theta;(1,0),(3,i),1)$, for $i = 0, 1, 2$. It follows that the initial distribution over the green states in Figure 1b can be approximated by normalising the density of each state as follows,

$$\Pr\left(\boldsymbol{N}(1) = (3,i)\,|\,\mathcal{Y}_1\right) = \frac{f_N(\theta;(1,0),(3,i),1)}{L_D^1(\theta)}, \qquad \text{for } i = 0, 1, 2.$$

Provided the population N is large, the diffusion approximation is highly accurate and more computationally-efficient than the exact likelihood. However, the diffusion approximation breaks down if the population of at least one compartment of the population process (S, I) is close to zero, as is the case during the early stages of an emerging epidemic (Kurtz, 1971; Barbour, 1980).

3. Accounting for bias during the early stages of an emerging epidemic

The DA process discussed up to this point imposes no restrictions on the trajectory of the incidence count N_I. However, if the probability of initial fade out is not considered appropriately, then the resulting estimates of R_0 are positively-biased. In this section we condition the DA process on eventually reaching a particular set of states $\mathcal{T} \subset \mathcal{S}$, such that once the process hits a state in $\mathcal{T}$ it may be considered an established outbreak. Note that we now refer to the original DA process as the *unconditioned* DA process.

The conditioned DA process is a CTMC taking values (N_I, N_R) in the state space $\mathcal{S}$. Define $u_{\boldsymbol{n}}$, $\forall \boldsymbol{n} \in \mathcal{S}$ as the probability that the unconditioned DA process ever hits a state in $\mathcal{T}$, starting from the state $\boldsymbol{n}$ (Norris, 1997). Then $\forall \boldsymbol{n}, \boldsymbol{m} \in \mathcal{S}$, with $\boldsymbol{m} \neq \boldsymbol{n}$, the conditioned DA process has the transition rates

$$\tilde{q}(\boldsymbol{n}, \boldsymbol{m}) = \begin{cases} (u_{\boldsymbol{m}}/u_{\boldsymbol{n}})\, q(\boldsymbol{n}, \boldsymbol{m}) & \text{if } \boldsymbol{n} \notin \mathcal{T}, \\ q(\boldsymbol{n}, \boldsymbol{m}) & \text{otherwise,} \end{cases}$$

with the condition that $\tilde{q}(\boldsymbol{n}, \boldsymbol{n}) = -\sum_{\boldsymbol{m} \neq \boldsymbol{n}} \tilde{q}(\boldsymbol{n}, \boldsymbol{m})$ (Waugh, 1958).

The set $\mathcal{T}$ may be defined as any subset of $\mathcal{S}$ provided there is a non-zero probability of reaching $\mathcal{T}$ from at least one state in $\mathcal{S} \setminus \mathcal{T}$. We define $\mathcal{T}$ as the set of all states with $N_I > n_T$, where n_T is defined as the threshold number of infection events. The threshold may be determined a-priori with the understanding that once N_I exceeds n_T, there must be a high probability that the outbreak is established. Hence, the *conditioned likelihood* is

$$L(y|\theta) = \prod_{k=1}^{k_T \wedge n} L_C^k(\theta) \prod_{k=k_T+1}^{n} L_E^k(\theta),$$

where $k_T = \min\{k | y_k > n_T\}$, $k_T \wedge n = \min\{k_T, n\}$, and $L_C^k(\theta)\,(k = 1, 2, \ldots, n)$ are the transition probabilities of the conditioned DA process. Clearly, a more rigorous choice for $\mathcal{T}$ would reflect the number of infectious individuals, rather than the number of infection events. However, this choice is less convenient to implement because it depends on the difference $N_I - N_R$, where N_I is observed but N_R is not. Furthermore, it is generally safe to assume that an outbreak is established if a large number of individuals have become infected, unless $R_0 < 1$.

The conditioned likelihood is calculated via the same algorithm as the exact likelihood, with the natural generalisation that the conditioned transition rates are used in place of the unconditioned transition rates for $k = 1, 2, \ldots, k_T$. In particular, the dashed transitions in Figures 1a and 1b are

removed from the model and the remaining transition rates are adjusted so that the CTMC will eventually reach the set $\mathcal{T}$ with probability one.

4. A computationally-efficient approach

The conditioned likelihood is computed via the forward equations which are computationally prohibitive for large N. Under the assumption that the outbreak is established by the time the number of infection events reaches n_T, it is reasonable to assume that the diffusion approximation will provide an accurate approximation of the process thereafter. Hence, we define the conditioned hybrid approximation as the hybrid discrete–continuous process which has the dynamics of the conditioned DA process while $N_I \leq n_T$, and the dynamics of the diffusion approximation otherwise. It follows that the *conditioned hybrid likelihood* is

$$L(y|\theta) = \prod_{k=1}^{k_T \wedge n} L_C^k(\theta) \prod_{k=k_T+1}^{n} L_D^k(\theta).$$

Computing the conditioned hybrid likelihood is achieved in the same way as the conditioned likelihood and the diffusion likelihood, with the exception that the initial distribution on the (k_T+1)th day is computed from the final distribution of the conditioned DA process at the end of the k_Tth day.

5. Results

In this section we demonstrate the accuracy and utility of our methodology by using it to estimate R_0 from daily incidence data from the first two weeks of an outbreak. Our analysis is comprised of two parts. First we demonstrate that conditioning reduces bias in estimates of R_0. Second, we demonstrate that the hybrid approximation provides an accurate and computationally-efficient means for estimating R_0 during the initial stages of an outbreak. To achieve this, we consider the four different parameter regimes displayed in Table 2. The values of R_0, γ and N have been selected to be representative of an influenza-like outbreak in a realistic population. The value of N also guarantees that the susceptible pool will not be depleted during the first two weeks of the epidemic, so we will be estimating R_0 from data during the early stages of the outbreak. We vary R_0 between Regimes 1 and 2 to investigate the effect of the underlying value of R_0 on the estimated R_0. We vary the threshold between Regimes 1 and 3, and Regimes 2 and 4 to investigate

the sensitivity of the conditioned likelihood to the threshold. In each regime, we consider $1,000$ independent simulated realisations of the SIR CTMC, each of which starts with a single infectious case, has a total duration of two weeks, and exceeds 50 infection events by the final day of the outbreak. We then illustrate the utility of our methodology by applying our conditioned hybrid process to an outbreak of pandemic influenza.

Parameter	Regime 1	Regime 2	Regime 3	Regime 4
R_0	1.2	1.4	1.2	1.4
n_T	50	50	20	20
γ	1/3	1/3	1/3	1/3
N	10^7	10^7	10^7	10^7
$I(0)$	1	1	1	1

Table 2: Parameters used for investigating our methodology. γ and R_0 are representative of influenza and N ensures that the susceptible pool is not depleted during the first two weeks of the epidemic.

In each regime we obtain the MLE and MPE of R_0 under the parameterisation $\theta = (1/\gamma, R_0)$, for $\theta \in \Theta$, where Θ contains all $1/\gamma, R_0 \geq 1/10$. To calculate the MPEs we use the exponential prior

$$f(1/\gamma, R_0) = \frac{1}{c_1 c_2} e^{-(1/c_1\gamma) - R_0/c_2},$$

which favours small values of $1/\gamma$ and R_0, but provides support to all $1/\gamma, R_0 > 0$. We selected $c_1 = 5$ and $c_2 = 1.3$ to provide a reasonable amount of weight to values of $1/\gamma$ and R_0 which are realistic for an influenza-like outbreak, see Figure 2. Our proposal density is a truncated bivariate Gaussian with support Θ and fixed covariance structure $\text{var}(1/\gamma) = 1$, $\text{var}(R_0) = 1/2$ and $\text{cov}(1/\gamma, R_0) = 0$. For each simulated data set, we generate four independent Markov chain Monte Carlo realisations on Θ consisting of $200,000$ iterations, and discard the initial $20,000$ iterations as burn-in.

To calculate the MLEs we maximise the log likelihood function $\ell(y|\theta) = \log(L(y|\theta))$ on Θ using MATLAB's `fmincon` function. We found that in some cases a MLE did not exist because the optimisation routine failed to converge. These cases were characterised by realisations where the number of infection events remained low for the first week before growing rapidly in the second week. These realisations have were dropped from the following analysis on the basis that they do

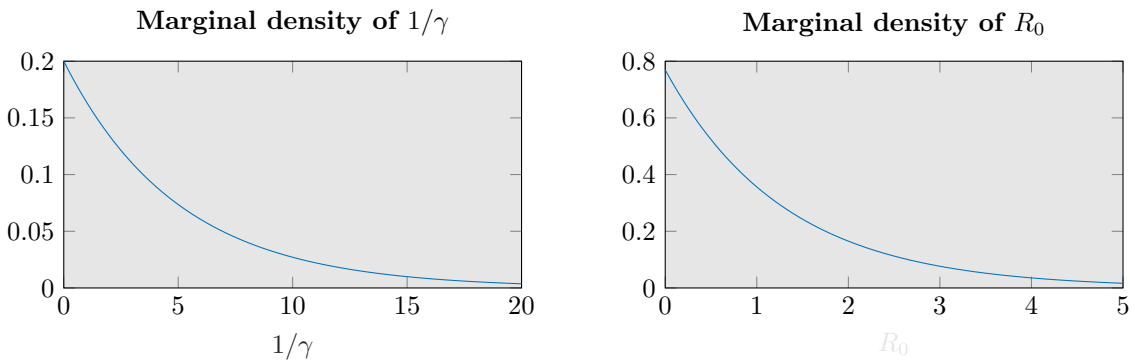


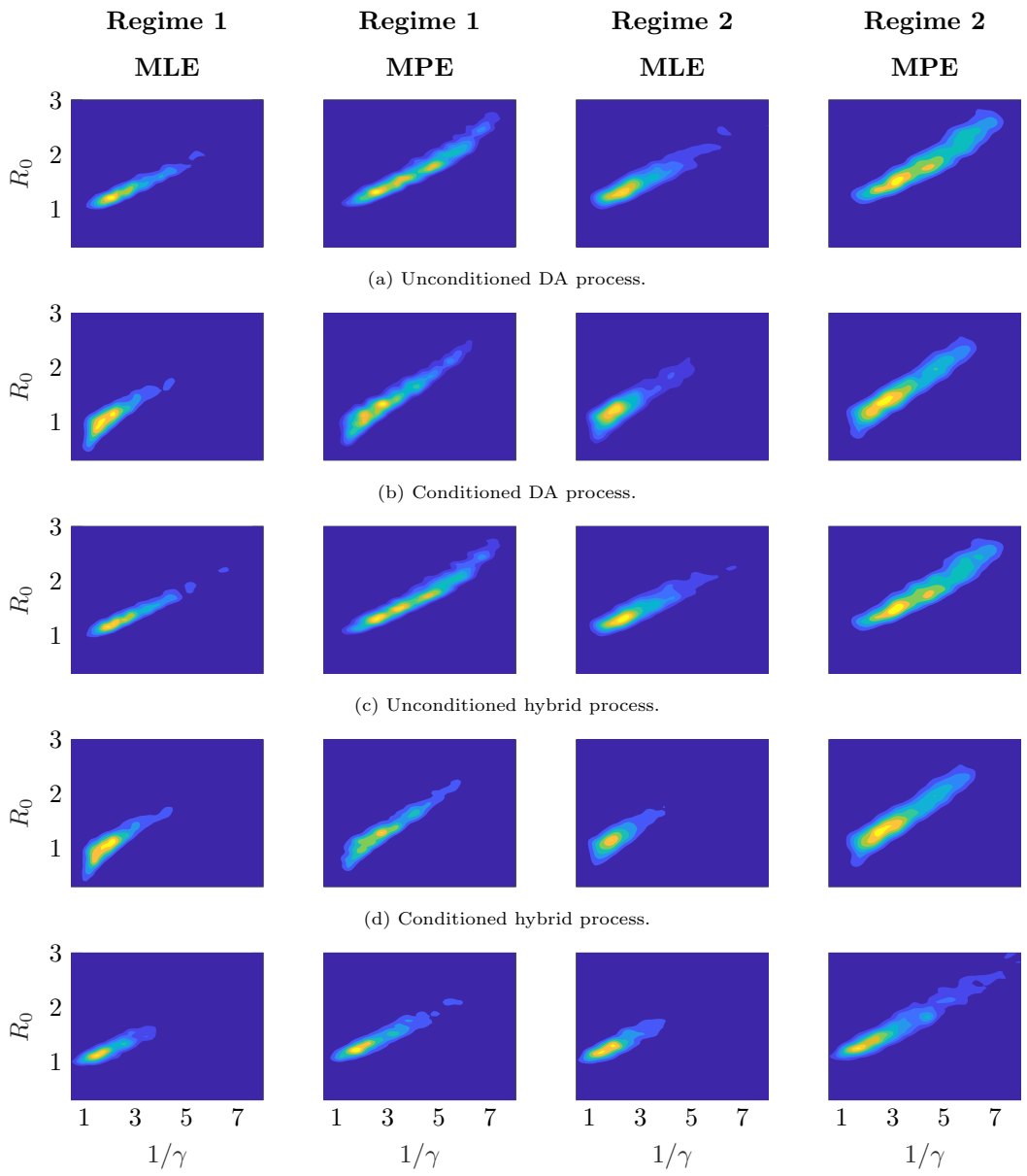
Figure 2: Marginal densities of the prior distribution of $1/\gamma$ and R_0.

not contain enough information to provide a reliable MLE.

Validation of the conditioned approach. We begin by presenting the MLEs and MPEs of R_0, across all Regimes. Figure 3 contains density estimates of the MLEs and MPEs under Regimes 1 and 2, plotted on the $(1/\gamma, R_0)$ axes. Each row contains parameter estimates according to a different model: unconditioned/conditioned DA process, unconditioned/conditioned hybrid process, and diffusion process. Figure 4 contains density estimates of the MLEs and MPEs under Regimes 3 and 4 for the conditioned DA process and conditioned hybrid process. Note that the density estimates of the MPEs are clearly different to the prior distribution, suggesting that our MPEs are not sensitive to the choice of prior distribution.

The density estimates of $1/\gamma$ and R_0 appear unimodal with a strong correlation between $1/\gamma$ and $R_0(= \beta/\gamma)$. The distributions appear non-symmetric, with a higher density associated with estimates which have smaller values of $1/\gamma$ and R_0. Under all regimes, the distributions obtained via maximum likelihood and Bayesian inference appear similar. The unconditioned estimates appear to favour higher values of R_0 and $1/\gamma$ than their conditioned counterparts, which we now investigate in more detail.

In the following analysis we use bean plots to compare independent data sets. The bean plot is comprised of horizontal side-by-side box plots for which the whiskers represent the 2nd and 98th percentiles. The outliers are shaded according to their distance away from the median. The box plots are accompanied by the corresponding density estimates which provide a more informative view of the distribution of the data.



(a) Unconditioned DA process.

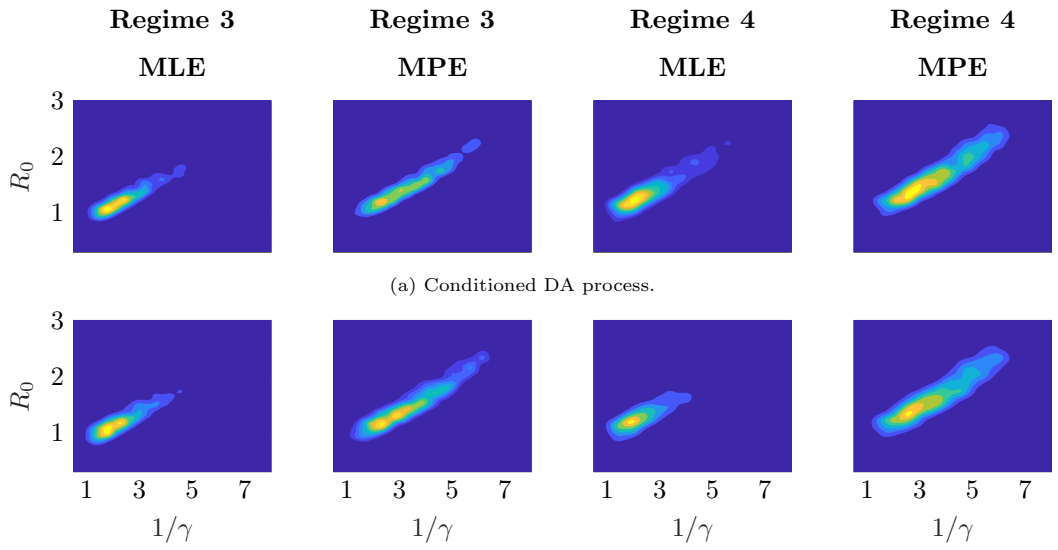
(b) Conditioned DA process.

(c) Unconditioned hybrid process.

(d) Conditioned hybrid process.

(e) Diffusion process.

Figure 3: Density estimates of the MLEs and MPEs of $(1/\gamma, R_0)$ obtained under Regimes 1 and 2. The rows contain estimates from the: unconditioned/conditioned DA process, unconditioned/conditioned hybrid process, and diffusion process. The density estimates demonstrate broad agreement between estimates of R_0



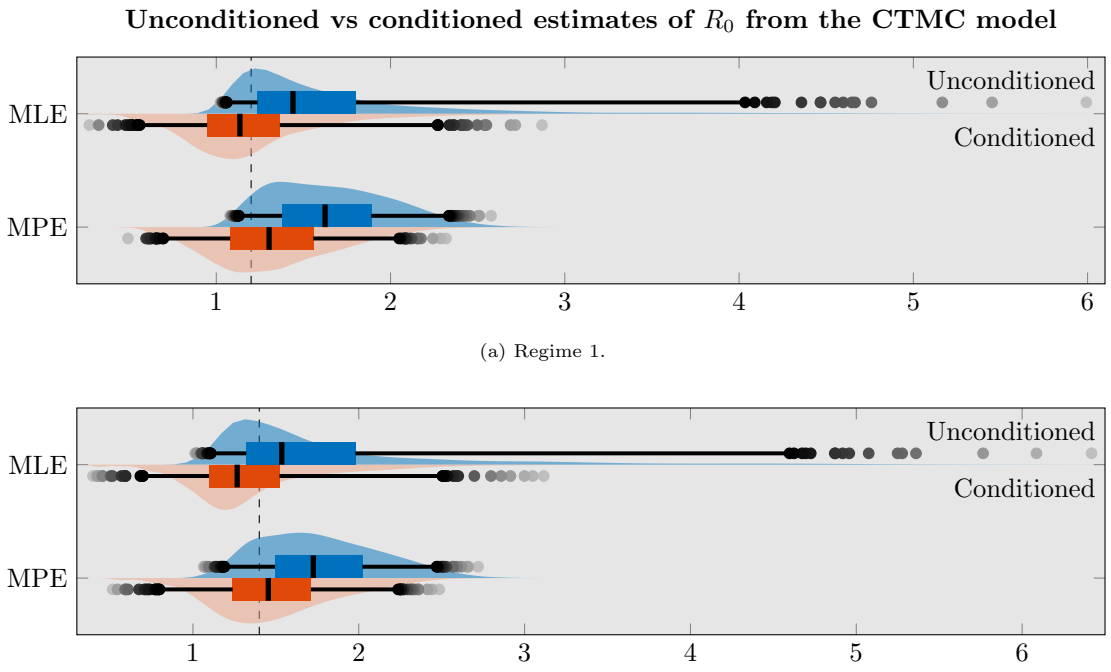
(a) Conditioned DA process.

(b) Conditioned hybrid process.

Figure 4: Density estimates of the MLEs and MPEs of $(1/\gamma, R_0)$ obtained under Regimes 3 and 4 from the conditioned DA process and conditioned hybrid process.

Figure 5 contains bean plots of the MLEs and MPEs of R_0 from the unconditioned DA process against the conditioned DA process, with the vertical dashed black line representing its true value. The unconditioned estimates are biased towards higher values of R_0 than the conditioned estimates and have a larger IQR. The unconditioned estimates show more bias in Regime 1 than Regime 2, presumably because the lower value of R_0 leads to a higher chance of extinctions and hence conditioning has a more significant impact on the transition rates. The conditioned MPEs show less bias than the MLEs though both MLEs and MPEs have a similar IQR in each regime. The MLEs appear more susceptible to outliers. We determined the cause of these outliers to be relatively uninformative realisations which do not provide enough information to obtain a reliable estimate of the underlying values of $1/\gamma$ and R_0.

Figure 6 contains bean plots of the paired difference between estimates from the unconditioned DA process and the conditioned DA process from Regime 1, plotted against Regime 2, where Figure 6a shows the difference in estimates of R_0, and Figure 6b shows the difference between estimates of the expected proportion of individuals who experience infection. Here, we have defined the difference to be the value of the unconditioned estimate minus the conditioned estimate. Figure 6a



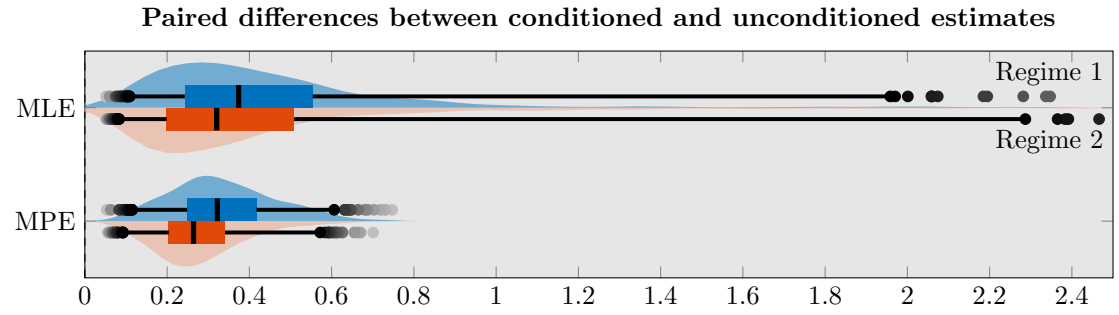
(a) Regime 1.

(b) Regime 2.

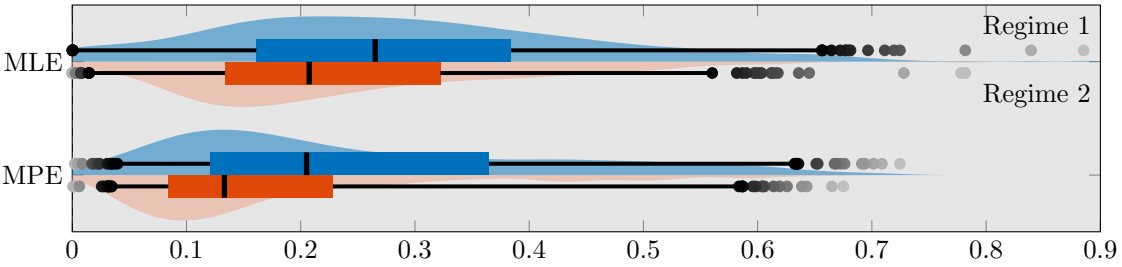
Figure 5: Bean plots of the estimated R_0 under Regimes 1 and 2. Bean plots are comprised of side-by-side box plots (where the whiskers represent the 2nd and 98th percentiles) plotted on top of a kernel density estimate. The conditioned estimate is smaller than the unconditioned estimate in every case. The unconditioned estimates in Regime 1 appear more biased than the unconditioned estimates in Regime 2.

shows that the unconditioned estimates of R_0 are always larger than the unconditioned estimates. On average, the unconditioned estimates are approximately 0.3 higher than the corresponding conditioned estimates. In addition, the MLEs appear more variable than the MPEs, although both distributions have a similar median.

Figure 6b translates the differences in estimates of R_0 into differences in the expected proportion of individuals of who experience infection, which provides an indication of the extent to which the unconditioned DA process overestimates the size of the outbreak. The median difference in the MLE (MPE) of the expected final epidemic proportions are 26% (20%) and 20% (13%) in Regime 1 and Regime 2. Meaning that even the most conservative estimate (MPE in Regime 2), overestimates the size of the outbreak by 13% of the total population, in 50% of realisations. This may have a significant impact on how public heath authorities perceive an emerging epidemic.



(a) Difference in estimate of R_0.

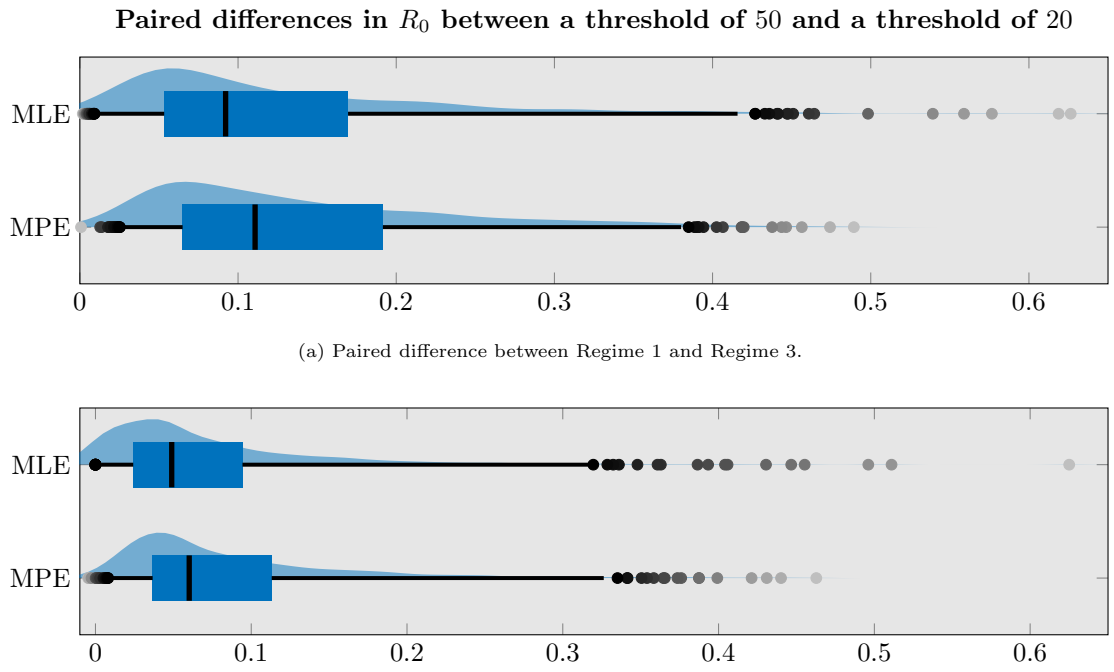


(b) Difference in estimate of the expected proportion of individuals who experience infection.

Figure 6: Bean plots of the paired difference between estimates from the unconditioned DA process and the conditioned DA process in Regime 1 plotted against Regime 2, where the difference is defined as the unconditioned estimate minus the conditioned estimate. In all cases the conditioned estimates are smaller than the unconditioned estimates.

Figure 7 contains bean plots of the paired difference between the conditioned DA process estimate of R_0 in Regimes 1 and 2 against Regimes 3 and 4. On average, the estimates in Regimes 3 and 4 are higher than those of Regimes 1 and 2, suggesting that the probability of extinction is considerable even after N_I has exceeded 20. However, the paired differences exhibited here are smaller than the paired differences exhibited in Figure 6a, suggesting that conditioning on a threshold of 20 is preferable to not conditioning at all. It is also clear that the change in the estimated R_0 is lower if the underlying value of R_0 is higher.

Validation of the hybrid approach. We now define the paired unconditioned hybrid (diffusion) difference as the estimate of R_0 from the unconditioned hybrid (diffusion) process minus the corresponding estimate from the unconditioned DA process. Figure 8 contains bean plots of the paired unconditioned hybrid differences against the paired diffusion differences, under Regimes 1 and 2.



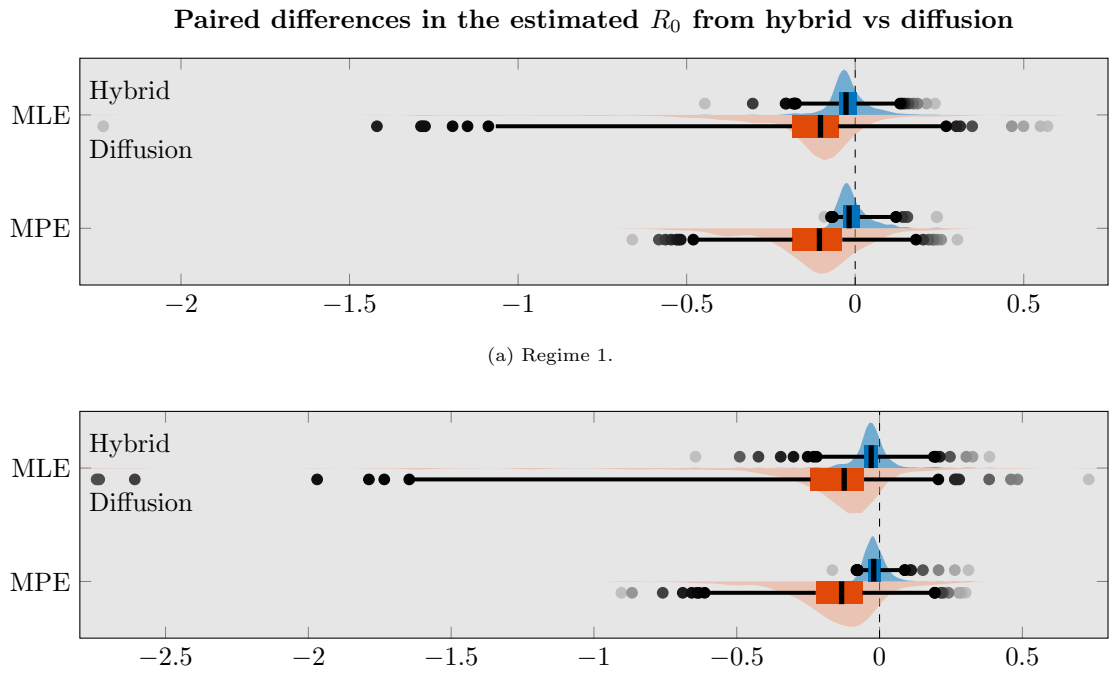
(a) Paired difference between Regime 1 and Regime 3.

(b) Paired difference between Regime 2 and Regime 4.

Figure 7: Bean plots of the paired difference in the conditioned DA process estimate of R_0 when the threshold is decreased from 50 to 20, where the difference is defined as the estimate from a threshold of 20 minus the estimate from a threshold of 50. The smaller conditioning level in Regimes 3 and 4 do less to reduce the positive-bias of the unconditioned estimate of R_0.

The paired diffusion differences demonstrate more bias and variation than the paired unconditioned hybrid differences, suggesting that the hybrid approximation is more reliable than the diffusion approximation in this context. This is unsurprising because the diffusion approximation is not suitable during the initial stages of an outbreak. However, since the hybrid approximation utilises the diffusion approximation only once the outbreak has become established, the difference exhibited here may be thought of as the amount of error accumulated by the diffusion approximation in modelling the initial stages of the outbreak.

Figure 9 shows bean plots of the paired differences between the estimate of R_0 from the conditioned DA and the conditioned hybrid, where the difference is defined as the conditioned hybrid estimate minus the conditioned DA estimate. The median bias in the MLE of R_0 is approximately -0.05, and the median bias for the MPE of R_0 is approximately -0.03. This indicates that the



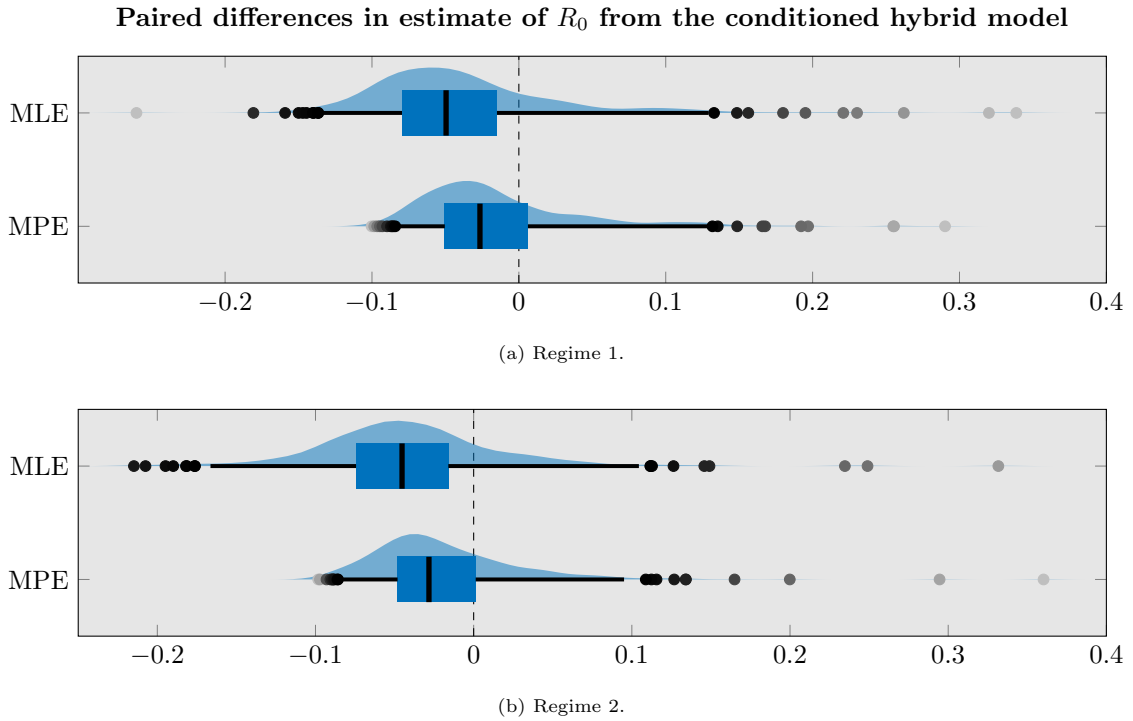
(a) Regime 1.

(b) Regime 2.

Figure 8: Bean plots of the paired differences in the estimated R_0 from the unconditioned hybrid against the diffusion. The difference is defined as the estimate from the approximation minus the estimate from the unconditioned DA process. The hybrid approximation is more accurate than the diffusion approximation.

conditioned hybrid approximation adds a slight (0.03 to 0.05) downwards bias on top of the 0.3 downwards correction of the conditioned DA process, when compared to the unconditioned DA process.

All computations have been carried out with the supercomputing resources provided by the Phoenix HPC service at the University of Adelaide, which is comprised of a Lenovo NeXtScale system consisting of 120 nodes, comprised of 2.3 GHz Intel Xeon E5-2698 v3 CPUs. The Bayesian analysis utilised 3GB of memory and was parallelised over 4 cores. To assess the computational-efficiency of the hybrid approximation we calculated the median runtime (in hours) to compute the MPE, averaged over all $1,000$ realisations. In Regimes 1 and 2 the computational runtime of the conditioned DA process was 1.27h and 1.55h, compared to 1.17h and 1.17h from the conditioned hybrid likelihood, indicating that the hybrid model did not have the opportunity to take full advantage of the computational-efficiency of its diffusion dynamics. In Regime 3 the median



(a) Regime 1.

(b) Regime 2.

Figure 9: Bean plots of the paired differences between the conditioned DA estimate of R_0 and the conditioned hybrid estimate of R_0, where the difference is defined as the conditioned hybrid estimate minus the conditioned DA estimate. The hybrid approximation exhibits a small amount of bias.

computational runtime of the conditioned DA process was 0.72h compared to 0.5h from the conditioned hybrid likelihood. In this case the threshold is lower so the hybrid approximation utilised its diffusion dynamics more than in Regimes 1 and 2, hence the hybrid approximation was noticeably faster than the DA process, on average. It's worth noting that the hybrid approximation scales better than the DA process with respect to the total number of observed infection events because its diffusion dynamics are relatively inexpensive, compared to CTMC dynamics.

Application to pandemic influenza. The first human infected with A(H1N1)pdm09 was recorded in the United States on the 15th of April 2009 (Gibbs et al., 2009; Team, 2009). Australia's initial response was to delay the entry and spread of the disease by enhanced case-finding, isolation, testing and treatment of incoming travellers with influenza-like illnesses; and prophylactic treatment and home quarantine of the close contacts of suspected/confirmed cases. The first confirmed

case in Australia was detected in a traveller returning home from the United States on the 9th of May. Subsequently, the first confirmed case in WA was detected in a traveller returning home from Canada via the United States on the 24th of May. On the 13th of June the WA government deemed the outbreak to be widespread and asked doctors to cease active case-finding, and prioritise influenza testing only to persons with severe influenza-like illness or established medical risk conditions (Weeramanthri et al., 2010). Prior to the 13th of June, all suspected or confirmed cases were actively followed-up and travel histories were recorded. This resulted in 102 confirmed cases and follow-up of 232 household contacts, plus a large number of aeroplane and school contacts. Of these 102 cases, 53% either originated in Victoria or were directly related to cases originating in Victoria. By the 30th of June, a total of 247 cases had been reported.

We are now considering a single outbreak so instead of reporting the distribution of the MLEs and MPEs, we now report the marginal distribution of R_0. We do so by sampling from the posterior distribution of R_0, as before, except this time we report the (2,25,50,75,98) percentiles of the samples from this distribution, rather than just its median. To achieve this, we use the same parameters as the previous analysis (4 chains of $200,000$ iterations with $20,000$ iterations as burn-in) with the exception that the population size is now assumed to be $2,040,000$, the population of Perth, and the mean of the marginal prior distribution of $1/\gamma$ is set to 3. We changed the mean of $1/\gamma$ to be consistent with other estimates of the mean serial interval of A(H1N1)pdm09 of 2.8 days (Nishiura et al., 2009a,b; Munayco et al., 2009). To assess the consistency of our methodology, we estimate the distribution of R_0 at a weekly resolution from the 24th of May to the 1st of August. Since the total number of cases by the 1st of August is prohibitively large for the DA process, we use the hybrid process instead. To demonstrate the impact of conditioning, we estimate the distribution of R_0 with and without conditioning, at the weekly intervals.

Figure 10 shows the number of notified cases of A(H1N1)pdm09, and box plots of the estimated distribution of R_0 from the conditioned hybrid in yellow and the unconditioned hybrid in ochre. The metrics of the conditioned distribution are always lower than the corresponding metrics of the unconditioned distribution. This difference is most prominent during the first few weeks of the outbreak and gradually subsides as the outbreak progresses because the impact of initial fade out decreases. The variability in the estimated distribution of R_0 can also be observed to decrease as the outbreak progresses. The MPE of R_0 from the conditioned model appears more stable than the MPE of the unconditioned model, which is influenced more heavily by a spike in cases

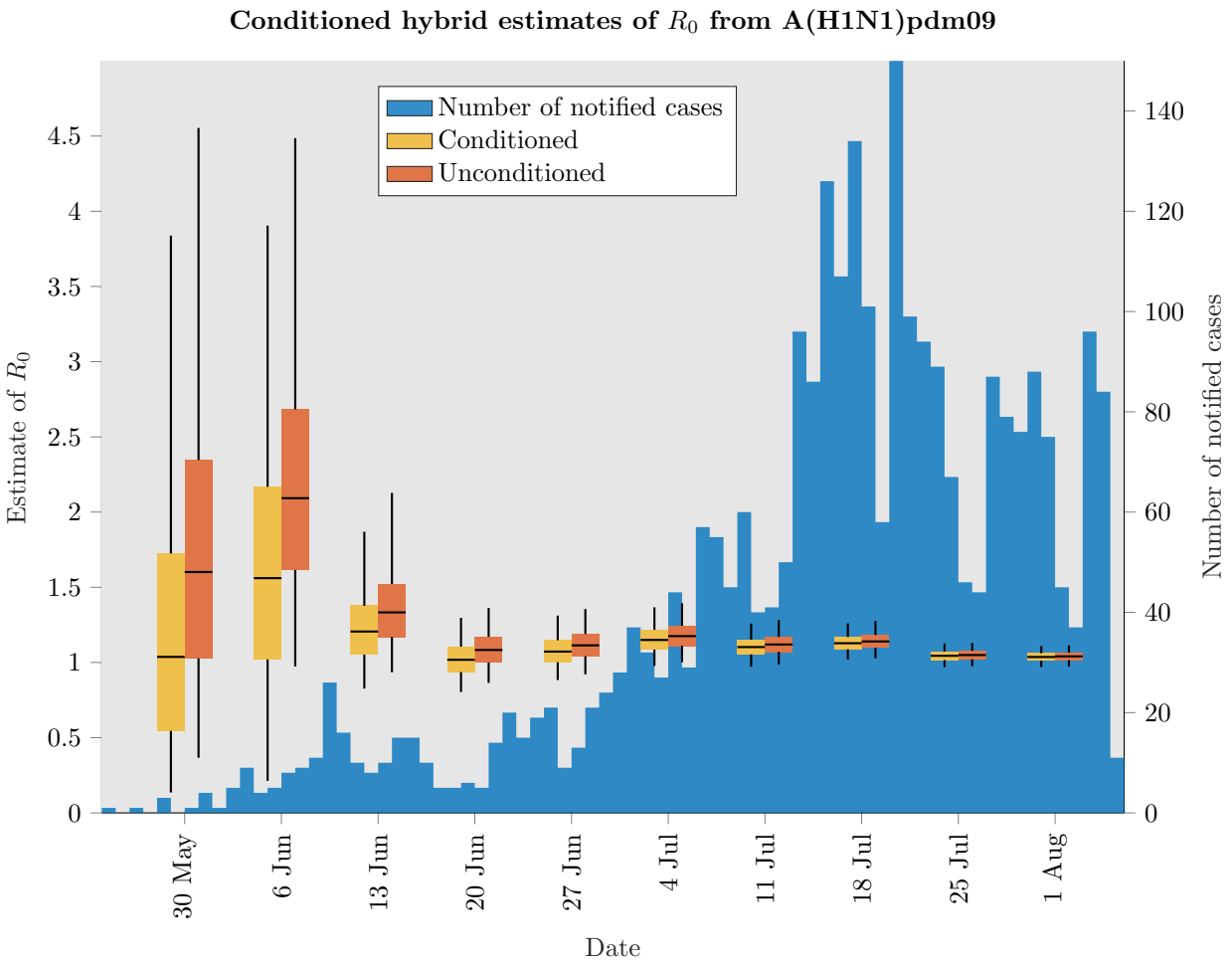


Figure 10: Number of notified cases of A(H1N1)pdm09 from WA with box plots of the estimated distribution of R_0 from the conditioned and unconditioned hybrid process. The conditioned hybrid process estimates a lower R_0 than the unconditioned.

which occurred during the third week of the outbreak. Our MPEs of R_0 from the conditioned hybrid process vary between 1 and 1.1, which are consistent with those in the literature for this outbreak (Kelly et al., 2010). The computational runtime of this analysis was under 1.5h for the first three weeks of the outbreak, and over a day for week 8 onwards.

6. Discussion

We have presented an approach to estimating R_0 from the SIR CTMC using daily incidence count data from the early stages of an emerging outbreak. This approach is conditioned on the observed number of infection events exceeding a predetermined threshold, at which stage the outbreak is regarded as established by public health officials. We also presented a highly accurate and computationally-efficient approximation applicable when the population size under consideration is computationally forbidding. We illustrated the utility of these approaches by estimating R_0 from multiple simulated outbreaks with influenza-like parameters and found our conditioned estimates of R_0 to be 0.3 smaller than the unconditioned estimate, on average. In addition, we demonstrated that the hybrid approach is more computationally-efficient than the standard CTMC approach and more accurate than the diffusion approximation.

We applied our methodology to an outbreak of A(H1N1)pdm09 in WA. We found that the conditioned hybrid process provides a more consistent estimate of R_0 during the initial stages of the outbreak, compared to the unconditioned hybrid, and that our estimates agree with those in the literature. However, our assumption that the outbreak is established by the time that the number of infectious individuals exceeds 50 may potentially be inaccurate in this case, considering that the number of notified cases is low for the first five weeks of the outbreak. Furthermore, a significant proportion of the notified cases during the initial stages of the outbreak are originated outside of WA, thereby positively biasing our estimates of R_0. Hence, it would be more appropriate to model this outbreak as one in which the number of infectious individuals eventually exceeds 102, considering that this is the number of notified cases at the time that the relevant authorities deemed the outbreak to be established (Kelly et al., 2010). In addition, the model should also allow infectious individuals to enter the population, rather than modelling the population as a closed system.

In general terms, the simple SIR CTMC used here is not a biologically plausible model. It makes unrealistic assumptions about the dynamics of the disease, such as the assumption that it has no incubation period, and the assumption that each individual's infectious period is exponentially distributed. Furthermore, it does not account for other sources of bias such as incomplete reporting, reporting rates which change over time, population heterogeneity (such as spatial variation, age-specific or household clustering of contacts), imported infectious cases, and pre-existing immunity. However, the salient point of the methodology presented here is that conditioning is a simple

mathematical tool which may be applied to a wide range of CTMC models as a means of obtaining less-biased estimates of R_0 using case incidence data from the early stages of an outbreak.

We are currently generalising the methodology presented here to a partially observed SEIR model. The inclusion of an incubation period and an unobserved infectious class should make this model more suitable for estimating the parameters of real outbreaks.

References

References

Barbour, A., 1975. The duration of the closed stochastic epidemic. Biometrika doi:10.1093/biomet/62.2.477.

Barbour, A.D., 1980. Biological Growth and Spread. Springer, Berlin, Heidelberg. volume 38 of *Lecture Notes in Biomathematics*. chapter Density dependent Markov population processes. pp. 36–49. doi:10.1007/978-3-642-61850-5_4.

Bartlett, M.S., 1949. Some evolutionary stochastic processes. J. R. Statist. Soc. .

Bettencourt, L.M.A., Ribeiro, R.M., 2008. Real time Bayesian estimation of the epidemic potential of emerging infectious diseases. PLOS ONE 3, 1–9. doi:10.1371/journal.pone.0002185.

Black, A.J., Ross, J.V., 2015. Computation of epidemic final size distributions. J. Theoret. Biol. 367, 159–165. doi:10.1016/j.jtbi.2014.11.029.

Cooper, B., Lipsitch, M., 2004. The analysis of hospital infection data using hidden Markov models. Biostatistics 5, 223–238. doi:10.1093/biostatistics/5.2.223.

Ethier, S.N., Kurtz, T.G., 2008. Markov Processes: Characterisation and Convergence. John Wiley and Sons, Inc., New Jersey. doi:10.1002/9780470316658.

Gibbs, A.J., Armstrong, J.S., Downie, J.C., 2009. From where did the 2009 'swine-origin' influenza A virus (H1N1) emerge? Virol. J. 6, 207–218. doi:10.1186/1743-422X-6-207.

Gilks, W.R., 2005. Markov Chain Monte Carlo. John Wiley and Sons, Ltd. doi:10.1002/0470011815.b2a14021.

Jenkinson, G., Goutsias, J., 2012. Numerical integration of the master equation in some models of stochastic epidemiology. PLoS ONE 7, 1–9. doi:10.1371/journal.pone.0036160.

Keeling, M.J., Wilson, H.B., Pacala, S.W., 2000. Reinterpreting space, time lags, and functional responses in ecological models. Science 290, 1758–1761. doi:10.1126/science.290.5497.1758.

Kelly, H.A., Mercer, G.N., Fielding, J.E., Dowse, G.K., Glass, K., Carcione, D., Grant, K.A., Effler, P.V., Lester, R.A., 2010. Pandemic (H1N1) 2009 influenza community transmission was established in one Australian state when the virus was first identified in North America. PLOS ONE 5, 1–9. doi:10.1371/journal.pone.0011341.

Kermack, W.O., McKendrick, A.G., 1927. A contribution to the mathematical theory of epidemics. Proc. R. Soc. Lond. A 138, 55–83. doi:10.1098/rspa.1927.0118.

Kurtz, T.G., 1970. Solutions of ordinary differential equations as limits of pure jump Markov processes. J. Appl. Prob. 7, 49–58. doi:10.2307/3212147.

Kurtz, T.G., 1971. Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. J. Appl. Prob. 8, 344–356. doi:10.2307/3211904.

Lemon, S.M., Hamburg, M.A., Sparling, P.F., Choffnes, E.R., Mack, A., 2007. Ethical and Legal Considerations in Mitigating Pandemic Disease: Workshop Summary. The National Academic Press, Washington, D.C.. chapter Strategies for Disease Containment. pp. 76–153. doi:10.17226/11917.

Meltzer, M.I., Cox, N.J., Fukuda, K., 1999. The economic impact of pandemic influenza in the United States: priorities for intervention. Emerg. Infect. Dis. 5, 659–671. doi:10.3201/eid0505.990507.

Mercer, G.N., Glass, K., Becker, N.G., 2011. Effective reproduction numbers are commonly overestimated early in a disease outbreak. Stat. Med. 30, 984–994. doi:10.1002/sim.4174.

Munayco, C.V., Gomez, J., Laguna-Torres, V.A., Arrasco, J., Kochel, T.J., Fiestas, V., Garcia, J., Perez, J., Torres, I., Condori, F., Nishiura, H., Chowell, G., 2009. Epidemiological and transmissibility analysis of influenza A(H1N1)v in a southern hemisphere setting: Peru. Euro. Surveill. 14.

Nishiura, H., Castillo-Chavez, C., Safan, M., Chowell, G., 2009a. Transmission potential of the new influenza A(H1N1) virus and its age-specificity in Japan. Euro. Surveill. 14.

Nishiura, H., Chowell, G., Safan, M., Castillo-Chavez, C., 2010. Pros and cons of estimating the reproduction number from early epidemic growth rate of influenza A (H1N1) 2009. Theor. Biol. Med. Model. 7, 1–13. doi:10.1186/1742-4682-7-1.

Nishiura, H., Wilson, N., Baker, M.G., 2009b. Estimating the reproduction number of the novel influenza A virus (H1N1) in a southern hemisphere setting: preliminary estimate in New Zealand. N. Z. Med. J. 122.

Norris, J.R., 1997. Markov Chains. Cambridge University Press. doi:10.1017/CBO9780511810633.

Pedroni, E., García, M., Espínola, V., Guerrero, A., González, C., Olea, A., Calvo, M., Martorell, B., Winkler, M., Carrasco, M.V., Vergara, J.A., Ulloa, J., Carrazana, A.M., Mujica, O., Villarroel, J.E., Labraña, M., Vargas, M., González, P., Cáceres, L., Zamorano, C.G., Momberg, R., Muñoz, G., Rocco, J., Bosque, V., Gallardo, A., Elgueta, J., Vega, J., 2010. Outbreak of 2009 pandemic influenza A(H1N1), Los Lagos, Chile, April-June 2009. Euro. Surveill. 15.

Pollett, P.K., 1990. On a model for interference between searching insect parasites. J. Austral. Math. Soc. Ser. B 32, 133–150. doi:10.1017/S0334270000008390.

Rebuli, N.P., Bean, N.G., Ross, J.V., 2016. Hybrid Markov chain models of SIR disease dynamics. J. Math. Biol. 74. doi:10.1007/s00285-016-1085-2.

Rida, W.N., 1991. Asymptotic properties of some estimators for the infection rate in the general stochastic epidemic model. J. R. Stat. Soc. B 53, 269–283. URL: http://www.jstor.org/stable/2345741.

Roberts, M.G., Nishiura, H., 2011. Early estimation of the reproduction number in the presence of imported cases: Pandemic influenza H1N1-2009 in New Zealand. PLOS ONE 6, 1–9. doi:10.1371/journal.pone.0017835.

Ross, J.V., 2012. On parameter estimation in population models III: Time-inhomogeneous processes and observation error. Theor. Popul. Biol. 82, 1–17. doi:10.1016/j.tpb.2012.03.001.

Ross, J.V., Pagendam, D.E., Pollett, P.K., 2009. On parameter estimation in population models II: Multi-dimensional processes and transient dynamics. Theor. Popul. Biol. 75, 123–132. doi:10.1016/j.tpb.2008.12.002.

Ross, J.V., Taimre, T., Pollett, P.K., 2006. On parameter estimation in population models. Theor. Popul. Biol. 70, 498–510. doi:10.1016/j.tpb.2006.08.001.

Safta, C., Sargsyan, K., Debusschere, B., Najm, H.N., 2015. Hybrid discrete/continuum algorithms for stochastic reaction networks. J. Comput. Phys. 281, 177–198. doi:10.1016/j.jcp.2014.10.026.

Scalia-Tomba, G., 1985. Asymptotic final-size distribution for some chain-binomial processes. Adv. Appl. Prob. doi:10.2307/1427116.

Simonsen, L., Clarke, M.J., Williamson, G.D., Stroup, D.F., Arden, N.H., Schonberger, L.B., 1997. The impact of influenza epidemics on mortality: introducing a severity index. Am. J. Public Health 87, 1944–1950. doi:10.2105/AJPH.87.12.1944.

Sprott, D.A., 2000. Statistical Inference in Science. volume 1 of *Springer Series in Statistics*. Springer-Verlag New York. doi:10.1007/b98955.

Team, N.S.O.I.A.H.V.I., 2009. Emergence of a novel swine-origin influenza A (H1N1) virus in humans. N. Engl. J. Med. 25, 2605–2615. doi:10.1056/NEJMoa0903810.

Viboud, C., Simonsen, L., Chowell, G., 2016. A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks. Epidemics 15, 27–37. doi:10.1016/j.epidem.2016.01.002.

Waugh, W.A.O., 1958. Conditioned Markov processes. Biometrika 45, 241–250. doi:10.1093/biomet/45.1-2.241.

Weeramanthri, T.S., Robertson, A.G., Dowse, G.K., Effler, P.V., Leclercq, M.G., Burtenshaw, J.D., Oldham, S.J., Smith, D.W., Gatti, K.J., M., G.H., 2010. Response to pandemic (H1N1) 2009 influenza in australia – lessons from a state health department perspective. Aust. Health Rev. 4, 477–486. doi:10.1071/AH10901.

White, L.F., Pagano, M., 2007. A likelihood-based method for real-time estimation of the serial interval and reproductive number of an epidemic. Stat. Med. 27, 2999–3016. doi:10.1002/sim.3136.