Simulation and Analysis Methods for Stochastic Compartmental Epidemic Models

Simulation and analysis methods for stochastic compartmental epidemic models

Tapiwa Ganyani*[1], Christel Faes[1], and Niel Hens [1,2]

[1]I-BioStat, Data Science Institute, Hasselt University, Hasselt, Belgium

[2]Centre for Health Economics Research and Modelling Infectious Diseases (CHERMID), Vaccine and Infectious Disease Institute, University of Antwerp, Antwerp, Belgium

*corresponding author: tapiwa.ganyani@uhasselt.be

Abstract

This paper considers simulation and analysis of incidence data using stochastic compartmental models in well mixed populations. Several simulation approaches are described and compared. Thereafter, we provide an overview of likelihood estimation for stochastic models. We apply one such method to a real life outbreak dataset and compare models assuming different kinds of stochasticity. We give some references of publications where detailed information can be found.

Keywords: stochastic SIR model, stochasticity, simulation, estimation, epidemic modelling

1 Introduction

Compartmental epidemic models are an important tool for understanding the dynamics of the spread of infectious diseases and hence provide quantitative support to public health decision making [1]. These models assume that a population can be divided into compartments such that at any given time an individual belongs to only one of the compartments. Ordinary differential equations are then used to describe how numbers of individuals in each compartment evolve over time. An appealing feature of these models is that they capture the epidemiological or biological mechanism of the disease such that their parameters have a natural epidemiological or biological interpretation [2]. In response to the public health threat posed by infectious diseases, these models therefore provide a useful framework with which to understand disease transmission processes as well as to interpret outbreak data and inform public health policy [1, 3]. They provide conceptual results such as, among others, reproduction numbers (average number of infections caused

by an infectious individual), herd immunity threshold (population fraction to be immunized to stop the infection from spreading in the entire population), final size (number of individuals who ultimately become infected during the entire period of the epidemic) and future course of the epidemic [2, 4, 5, 6]. These models have been used to study dynamics of many infectious disease outbreaks including the historical 1918 Spanish flu pandemic and the most recent coronavirus disease 2019 (COVID-19). The 1918 Spanish flu pandemic caused by H1N1 influenza A virus was cited as the most severe pandemic in history. It is estimated that about 500 million people or one-third of the world's population became infected with this virus and had clinical illnesses. The number of deaths was estimated to be at least 50 million worldwide. Its impact has not been limited to 1918-19, other influenza pandemics such as H1N1/09, H2N2 and H3N2 have been caused by descendants of the 1918 virus [7]. In December, 2019, a local outbreak of COVID-19 was detected in Wuhan (Hubei, China). The outbreak later spread to every province of mainland China as well as 188 other countries and regions, with more than 7 million confirmed cases and more than 400 thousand deaths as of June 9, 2020 [8].

In this paper, for illustrating purposes, we consider one example of a compartmental epidemic model, i.e., the Susceptible-Infected-Removed (SIR) model in a well mixed population. This model is important in infectious disease modelling because all compartmental epidemic models can be thought of to be extensions of the SIR model [3, 9]. For example, in the context of Ebola, greater detail has been incorporated to account for exposed cases, i.e., infected cases who are not yet infectious leading to a Susceptible-Exposed-Infected-Removed (SEIR) model. In the context of tuberculosis for which immunity is not lifelong the SEIR model has been extended to include reinfections [10]. With or without modifications, this model has been widely used to study outbreaks of human-to-human infectious diseases such as influenza, tuberculosis, measles and Ebola, see e.g. [3, 6, 10, 11, 12]. Though we focus on a basic model applied to human-to-human infections, compartmental models are more general and can be extended to incorporate, for instance, vectorborne transmission for diseases such as malaria where mosquitoes are the vectors [13].

Compartmental models can be formulated as deterministic or stochastic. Deterministic compartmental epidemic models implicitly assume that individuals in the same compartment have the same characteristics and therefore behave the same. They do not account for the fact that different individuals in the population have different nutritional, environmental or genetic status and may therefore have different epidemiological behaviour making infectious disease spread a random process [2, 4, 14]. These models are however popular in the study of infectious diseases in large populations because randomness due to individual-to-individual variability averages out making the deterministic model a reasonable approximation to describe observed data. However, for small populations randomness has a large impact on the transmission process hence

deterministic models become unsuitable [2, 4, 14]. Also, in large populations, because randomness can cause dramatic deviations from the deterministic model, average trajectories obtained from a stochastic model may not always be adequately approximated by the deterministic counterpart [4]. It is therefore well recognized that stochastic compartmental epidemic models are important when studying infectious disease dynamics [4, 15, 16]. From an estimation view point, stochastic models enable quantification of uncertainty for parameters estimated from disease outbreak data [17].

Simulation of stochastic models plays an important role in the modelling of infectious diseases using compartmental models. It allows understanding the qualitative behavior of models, e.g., sensitivity of solutions to different assumptions, see e.g. [18, 19]. It also enables quantitative study of problems that are mathematically intractable - for detailed models which are difficult or impossible to solve analytically simulations are necessary [3, 20]. Also, strategies to mitigate or delay the impact of future seasonal or pandemic outbreaks can be explored via simulation, see e.g. [21, 22, 23]. Moreover, analysis of simulated data can be used to evaluate performance of estimation algorithms as well as to investigate which model parameters or combinations of parameters are estimable from data, see e.g. [11, 12, 24]. From an estimation point of view, the ability to simulate from models comes in handy for problems where the likelihood function is not tractable [24].

Statistical analysis of infectious disease data has been geared towards estimating model parameters from observed infectious disease data [25], therefore, accurate estimation of model parameters from data is of utmost importance. Parameter estimation methods for stochastic compartmental models usually rely on calculating the likelihood. Oftentimes, due to interplay of a proposed model and observed data, the likelihood is intractable leading to estimation challenges [26]. With the advent of modern computing power, the problem of intractable likelihood has led to continual development of statistical methodologies for parameter inference such as Markov Chain Monte Carlo (MCMC) methods, Approximate Bayesian Computation methods and sequential Monte Carlo, see e.g. [24, 26, 27] and references therein.

This review provides an overview of basic ideas on data simulation as well as statistical analysis for stochastic compartmental models. We will omit technical details whenever they are not essential for the discussion. The goal of the statistical analysis is not to compare different estimation methods but to compare models with different levels of stochasticity. Using real outbreak data, we compare stochastic and deterministic compartmental models with respect to parameter uncertainty; also, we demonstrate the impact of accounting for overdispersion on model fit as well as on parameter uncertainty.

2 Deterministic SIR model

In a closed population with no births or deaths, a simplified version of the SIR model [28] assumes that a population is divided into three compartments (Figure 1), $S(t)$ - number of susceptible individuals at time t, $I(t)$ - number of infectious individuals at time t that is capable of infecting susceptible individuals, and $R(t)$ - number of recovered/removed (immune or dead) individuals at time t. The population size M is given by $S(t) + I(t) + R(t) = M$. The model postulates that initially there is a single or few infective(s), they transmit infection to susceptibles during their infectious period; thereafter an infected person is infectious for a certain amount of time before recovering or dying. The disease continues to spread in this manner until there are no more infectives.

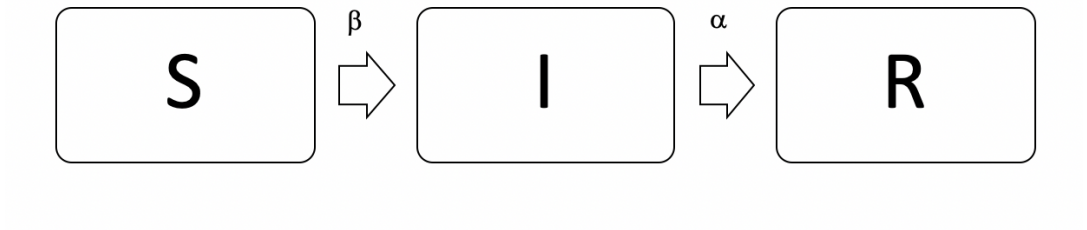

Figure 1: Flow chart for the SIR model; β represents the rate at which individuals come into effective contact per unit time, α represents the rate at which infected individual recover.

Assuming that individuals within a population mix completely and move randomly (homogeneous mixing), the total number of contacts between susceptibles and infectives is given by $S(t)I(t)$ (a principle called mass action) [29]. In reality not all contacts will lead to infection, assuming that the rate at which individuals come into effective contact per unit time is β (transmission rate), new infections (incidence) will occur at a rate $\beta S(t)I(t)$. It is common to specify the incidence rate as $\beta S(t)I(t)/M$ to reflect that the probability that a susceptible individual encounters an infected individual is independent of population density [30]. In continuous time, the SIR model is described by the following set of ordinary differential equations:

$$\begin{aligned} \frac{dS(t)}{dt} &= -\frac{\beta S(t)I(t)}{M} \\ \frac{dI(t)}{dt} &= \frac{\beta S(t)I(t)}{M} - \alpha I(t) \\ \frac{dR(t)}{dt} &= \alpha I(t), \end{aligned} \tag{1}$$

where $1/\alpha$ is the average infectious period and t represents calendar time. A common initial condition of (1)

is $S(0) = M - 1$, $I(0) = 1$, and $R(0) = 0$. The basic reproduction number R_0 is given by β/α - if $R_0 < 1$, then the epidemic will die out, while if $R_0 > 1$ the epidemic can grow [31].

In real life epidemic data are always observed in discrete rather than continuous time (e.g. daily, weekly, monthly) and they reflect aggregated information between consecutive reporting periods. A discrete approximation of (1) can be formulated as:

$$\begin{aligned} S(t+h) &= S(t) - \frac{\beta S(t) I(t)}{M} h \\ I(t+h) &= I(t) + \frac{\beta S(t) I(t)}{M} h - \alpha I(t) h \\ R(t+h) &= R(t) + \alpha I(t) h, \end{aligned} \tag{2}$$

where $h > 0$ represents the discrete time step. As $h \to 0$ (2) approaches (1), therefore, for sufficiently small h, main features of (1) also hold for (2).

3 The general stochastic SIR model

Most stochastic models for compartmental epidemic models are Markov processes [32]. Markov processes are stochastic processes whereby the future state of the population at time, $t + 1$, depends only on the current state at time t, see e.g. [33]. Let $M_I(t)$ and $M_R(t)$ be Poisson processes which denote, at time t, the number of individuals who have been infected and the number of individuals who have recovered, respectively. Following the work of [34], the standard stochastic version of (1), also known as the general stochastic model, is defined as a bivariate continuous-time Markov process (Markov jump process) $\{(S(t), I(t)) : t \geq 0\}$. For $h > 0$, it is specified by the following infinitesimal increment probabilities,

$$\begin{aligned} P\Big(\Delta M_I(t) = 1 | M_I(t)\Big) &= \frac{\beta S(t) I(t)}{M} h + o(h) \\ P\Big(\Delta M_R(t) = 1 | M_R(t)\Big) &= \alpha I(t) h + o(h), \end{aligned} \tag{3}$$

where $\Delta M_i(t) = M_i(t + h) - M_i(t)$, $i \in (I, R)$ denotes increments of $M_i(t)$ and $o(h)$ tends to zero in limit as h approaches zero. Thus, new infections and new recoveries occur at the points of two non-homogeneous Poisson processes with rates $\beta S(t)I(t)/M$ and $\alpha I(t)$, respectively. A direct consequence of assuming $M_I(t)$ and $M_R(t)$ are Poisson processes is that the amount of time until the next individual gets infected as well as the amount of time until an infected person recovers are exponentially distributed.

In ecological and epidemiological modelling, the nature of stochasticity represented by model (3) is usually referred to as demographic stochasticity - it represents unpredictable event times due to, e.g., individual-to-individual differences in nutrition, environment or genetic status. Such effects average out in large populations which essentially means that the role of this kind of stochasticity diminishes with increasing population size [27].

4 Simulation

4.1 Kolmogorov Forward Equations

From the stochastic model (3) differential equations for the infinitesimal increment probabilities can be derived. The equations often known as 'forward Kolmogorov equations' or 'master equations' can be used for predicting future dynamics of the SIR model. Using a large set of deterministic differential equations for the probability of being in each possible state (S, I), they provide a complete description of behaviour of the stochastic system, see e.g. [32]. For an SIR model the master equations are given by,

$$\frac{dp_{S,I}(t)}{dt} = p_{S+1,I-1}(t)\Big(\frac{\beta}{M}(S+1)(I-1)\Big) + p_{S,I+1}(t)\Big(\alpha(I+1)\Big) - p_{S,I}(t)\Big(\frac{\beta}{M}SI + \alpha I\Big), \tag{4}$$

where $p_{S,I}(t)$ is the probability of having, at time t, S susceptibles and I infectives. The master equations are computationally feasible for sufficiently small populations because computational time increases proportionally to the number of states or even faster (there are as many $p_{S,I}$'s as there are states). For a SIR model, the number of possible states that the process can be is $(1/2)(M+1)(M+2)$; the number of states grows like $1/k\,!M^k$ where k is the number of possible compartments that an individual can be in, see [32, 35] and references therein. Moreover, for models with multiple compartments (see Section 1), it is often difficult to find closed form solutions for the infinitesimal increment probabilities (4) [35]. For large populations stochastic simulation is handy. A commonly used stochastic simulation algorithm (SSA) for Markovian models is the Gillespie algorithm [36]. The algorithm simulates an exact stochastic version of the trajectory that would be obtained by solving the corresponding master equations.

4.2 The Gillespie SSA

The algorithm assumes that in a well mixed population of a fixed size, at a given time, an individual belongs to one of k compartments; changes in numbers of individuals in each compartment are a result of 'reactions'

between interacting compartments. In the context of the 3-compartment SIR model, the algorithm proceeds in two steps:

1. Simulate time at which the next event will occur; in this case an event refers to the movement of one individual either from S to I or from I to R. The time Z until the next event occurs follows an exponential distribution with rate equal to the sum of the rates over all possible events (in this case two events are possible, i.e., movement from S-to-I and I-to-R). Denoting $c_1 = \beta S(t)I(t)/N$ and $c_2 = \alpha I(t)$ (note there are as many c's as there are reactions), Z is distributed as:

$$\mathrm{g}_Z(z) = \left(\sum_{i=1}^{2} c_i\right) \exp\left(-z \sum_{i=1}^{2} c_i\right). \tag{5}$$

2. When the event time has been simulated, next simulate which event occurs. Event rates c_i are converted into probabilities, one of the events is then selected at random according to:

$$P(\text{Event} = v) = \frac{c_v}{\sum_{i=1}^{2} c_i} \tag{6}$$

   Using these distributions, the algorithm proceeds as follows:

   i. Set initial population numbers;

   ii. Calculate c_1 and c_2;

   iii. Simulate time to next event from (5);

   vi. Simulate which event occurs from (6);

   v. Update the population sizes in line with the event that occurred;

   vi. Update the time and return to the second step.

Since a single simulation is insufficient to represent the average behaviour of the process, many replicates are required to obtain a representative picture.

The Poisson τ - leap method

The drawback of the Gillespie SSA is that it requires a great amount of computation time if the number of individuals in at least one compartment is large - transition rates change with each jump of the process; which happens very often. Several approximate procedures known as 'τ-leap' methods have been proposed to improve computational speed with acceptable losses of accuracy. The principle behind these methods is that if the time axis can be divided into contiguous small subintervals such that we can determine, in each subinterval, the

number of movements of a given type, then we can do without the exact time at which the movements occurred. Substantial computational speed can be gained by leaping along the time axis from one subinterval to the next instead of moving from one event to the next [37]. Accuracy is achieved for small τ (leap condition) [37, 38, 39].

In the so-called Poisson τ - leap method, it is assumed that for each given type of event, the number of events occurring in a small subinterval independently follow a Poisson distribution. The number of events of a given type is simulated independently and then numbers of individuals in each compartment are updated accordingly. The algorithm proceeds as follows, for a fixed time step τ:

i. Set initial population numbers;

ii. Calculate c_1 and c_2;

iii. Simulate the number of events of each type from $\text{Po}(c_1\tau)$ and $\text{Po}(c_2\tau)$;

iv. Update the population sizes in line with the events that occurred;

v. Update the time to $t + \tau$ and return to the second step.

A disadvantage of this method is that numbers of individuals in the compartments may become negative due to unboundedness of Poisson random variables; the binomial distribution can be used to remedy this problem [39]. Another disadvantage is that a particular τ may not yield similar accuracy along the entire time axis. A remedy to this problem is to allow τ to vary along the time axis in such a way that at each time step a trade-off is made between accuracy and speed [37, 38].

4.3 Stochastic differential equations

Another way to incorporate stochasticity in compartmental models is via stochastic differential equations SDEs. SDEs are used to model systems of continuous variables (states) that fluctuate in continuous time due to randomness, where randomness might be due to random coefficients or dependence on a stochastic force [40]. An SDE of a continuous process Y is usually of the form:

$$dY(t) = \mu\big(Y(t)\big)dt + \Psi\big(Y(t)\big)dW(t), \tag{7}$$

where μ is the deterministic component and W is the random component (diffusion process) intensified by Ψ. In practice it is common to approximate Markov jump processes (3) by Markov processes such as diffusion processes which have continuous state space and almost surely continuous sample paths. The justification is

that jump sizes in the Markov jump processes are infinitesimally small such that the discontinuities in the trajectory of the process can be fairly approximated by continuous curves [38, 40]. The SDE for the SIR are given by,

$$\begin{aligned} dS(t) &= -\big(\beta S(t)I(t)/M\big)dt - \sqrt{\beta S(t)I(t)/M}dW_1(t) \\ dI(t) &= \big(\beta S(t)I(t)/M - \alpha I(t)\big)dt + \sqrt{\beta S(t)I(t)/M}dW_1(t) - \sqrt{\alpha I(t)}dW_2(t) \\ dR(t) &= \alpha I(t)dt + \sqrt{\alpha I(t)}dW_2(t), \end{aligned} \tag{8}$$

where $W(t) = (W_1(t), W_2(t))^T$ is a vector of two independent standard Wiener processes, i.e., $W_i \sim N(0,t)$ or $dW_i \sim N(0,dt)$, see derivations in [35]. The choice of the normal distribution in this case preserves the Markov property.

However, analytical solutions are often difficult to obtain. When solutions are not obtainable, simulating the SDE at discrete time points is straightforward since the random components are simply normally distributed. Numerical procedures are typically used to approximate the solution on a regular grid of points; a popular simple procedure is the Euler-Maruyama approximation [40]. The approximation proceeds on a regular grid $t_0, t_0 + \Delta, t_0 + 2\Delta, \ldots, t_n$ as follows,

$$\begin{aligned} \Delta S(t) &= -\big(\beta S(t)I(t)/M\big)\Delta t - \sqrt{\beta S(t)I(t)/M}\Delta W_1(t) \\ \Delta I(t) &= \big(\beta S(t)I(t)/M - \alpha I(t)\big)\Delta t + \sqrt{\beta S(t)I(t)/M}\Delta W_1(t) - \sqrt{\alpha I(t)}\Delta W_2(t) \\ \Delta R(t) &= \alpha I(t)\Delta t + \sqrt{\alpha I(t)}\Delta W_2(t), \end{aligned} \tag{9}$$

where ΔW_i ($\sim N(0, \Delta t)$) is finite and sufficiently small.

4.4 Comparison of simulation procedures

Table 1 shows computational time required to simulate one realisation of the stochastic SIR model for different population sizes using the Gillespie SSA and the Poisson τ-leap method ($\tau = 1/1000$). Simulations are performed using the `R` package `Gillespie SSA`. Computational time for SDE is not included for two reasons. First, the Euler-Maruyama method used to simulate SDE is closely related to τ-leaping approaches, hence, it is expected that computational time for SDEs would be similar to that of τ-leaping approaches. Secondly, simulation methods for SDEs are not implemented in the `R` package `GillespieSSA`; though it is straightforward to program SDEs in `R`, a neat comparison can be made when both τ-leaping and Euler-Maruyama methods

are coded in a similar format.

Table 1: Time required to simulate the stochastic SIR model (3) using: i) Gillespie SSA; ii) Poisson τ-leap ($\tau = 1/1000$). Simulation are perfomed using the `R` package `GillespieSSA` on a 3.1 GHz PC.

M	time to perform simulation (seconds)	
	Gillespie SSA	Gillespie τ-leap
10,000	33.17	32.60
100,000	66.39	67.14
500,000	1405.01	65.99
750,000	3218.06	77.70
1,000,000	5456.66	64.13

For the Gillespie SSA, computational time is high and increases rapidly with increasing population size. On the other other hand, for τ-leaping there is huge gain in computation speed, moreover, computational time does not increase for larger population sizes [14].

Table 2 shows, for epidemics simulated using the Gillespie SSA, Poisson τ-leap ($\tau = 1/1000$) and SDE ($\Delta = 1/1000$), a descriptive summary of peak time (amount of time from epidemic onset to peak), peak incidence (number of cases at the peak of the epidemic), epidemic duration (amount of time from epidemic onset until every infective is recovered) and final size. Also shown are the true values obtained from the deterministic model. Initial conditions are specified as: $S(0) = M - 1$, $I(0) = 1$, and $R(0) = 0$ with $M = 100,000$; $R_0 = 1.8$ is chosen similar to R_0 for influenza [41] with a recovery rate $\alpha = 1/4.1$ days [42] corresponding to a transmission rate $\beta = \alpha R_0$. For each method 650 epidemics were simulated.

Table 2: Descriptive summary of peak time (days), peak incidence, epidemic duration (duration) and final size for epidemics simulated using the Gillespie SSA, Poisson τ-leap ($\tau = 1/1000$) and SDE ($\Delta = 1/1000$).

Simulation procedure	2.5%, 50% and 97.5% percentiles			
	peak time (days)	peak incidence	epidemic duration (days)	final size
Gillespie SSA	(43, 50, 64)	(3054, 3197, 3358)	(114, 129, 150)	(72500, 73246, 73875)
Gillespie τ-leap	(43, 51, 67)	(3055, 3193, 3350)	(119, 137, 165)	(72537, 73224, 73993)
SDE	(53, 54, 54)	(3153, 3159, 3165)	(144, 145, 147)	(73223, 73244, 73268)
True values	54	3159	145	73244

As expected, summary statistics for the Gillespie SSA and the τ-leap approach are in the same ranges since the τ-leap algorithm approximates the Gillespie SSA well for small τ [37, 39]. Summaries for these two are also fairly close to the true values, this is also expected since demographic stochasticity averages out in large populations. On the other hand, summaries for SDE are very close to the true values. This is because the noise term is very small i.e., $\sim N(0, 1/1000)$, moreover, the noise is smoothed out in large populations. If the noise term is very large, it can obscure the deterministic component resulting in a pure random walk (Section 5.2). For small population sizes stochasticity can have a greater effect. Notably, for the Gillespie methods, peak time and epidemic duration, and hence peak incidence and final size, are more variable compared to

SDE. This is due to the manner in which stochasticity is incorporated, it plays a role at epidemic take off, i.e., some simulated epidemics will take off slower or faster than the deterministic model; moreover, the Gillespie methods aim at simulating all possible behaviour of the stochastic model (Section 4.1). In SDE, stochasticity causes variations about the deterministic component.

5 Estimation

Infectious disease outbreak data are typically available as time series of newly reported cases aggregated over some time period, usually daily or weekly. Statistical analysis typically focuses on inferring parameters from these data via likelihood-based inference. The recovery rate as well as other epidemiological parameters are typically measured directly from appropriate studies such as laboratory experiments or surveys of infected individuals in a population. On the other hand, the transmission rate which combines many biological, social and environmental factors is often inferred from time series of case counts [2].

Parameter inference in stochastic compartmental models is often complicated by data incompleteness. Only partial information about the disease dynamics is observed. The underlying disease transmission process is continuous in time but the exact moment at which events occur are never observed - recorded data are an aggregation of events that occurred between consecutive reporting periods with exact time of occurrence for each event being unknown. Also, it is often the case that data on at least one compartment is unobserved. Data incompleteness presents an estimation problem when performing likelihood-based inference as it leads to an intractable likelihood [12, 26]. In this section we first introduce, following [43], the partially observed Markov process (POMP) modelling framework which is a practical way to connect a postulated dynamic epidemic model with observed data. Next we discuss overdispersion whose role is important in the statistical analysis of infectious disease data [44, 45]. Thereafter we provide an overview of the key aspects regarding likelihood-based inference.

5.1 Relating models to data: the POMP model

Let Y_{t_n} denote observed incidence counts at discrete times $t_1 < t_2 < \cdots < t_N$. The observed counts are modeled as noisy and incomplete observations of a Markov state process $\{X_t,\ t \geq t_0\}$ which can either be discrete or continuous (Section 2). In this case $\{X_t,\ t \geq t_0\}$ represents the incidence trajectory obtained from the SIR model (process model) which can either be deterministic (Section 2) or stochastic (Section 3). A state process $\{X_t\}$ is Markovian if, given the current value of the process, the history of the process is uninformative about the future of the process, i.e., $P\big(X_n|X_{n-1}, \cdots, X_0\big) = P\big(X_n|X_{n-1}\big)$. Thus, a POMP model consists

of two components: an unobserved continuous time process model which describes the dynamics of disease spread at the population level, and a measurement model, which describes how data Y_{t_n} collected at discrete time points $\{t_1, \cdots, t_N\}$ are connected to the process model via the state process $\{X_t,\ t \geq t_0\}$ (Figure 2). The measurement model is specified by a distributional assumption; for count data, a natural choice is the Poisson model. Stochasticity implied by a distributional assumption is often referred to as observational or measurement noise. It represents data recording uncertainties, e.g. incomplete reporting of cases or mis-diagnosis [14, 46].

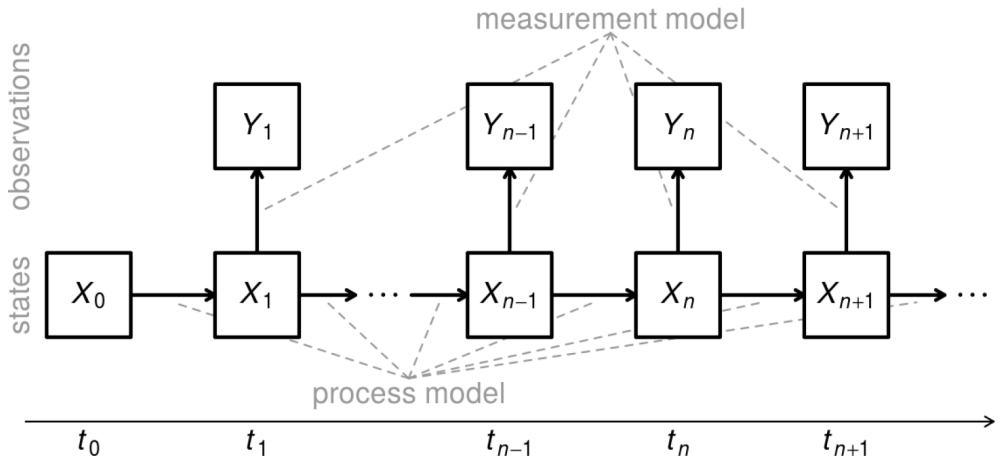

Figure 2: POMP model: Y_{t_n} denotes observed incidence data at discrete time points which connected to the continuous state process X_t at that time point. Source of diagram: [43]

At any given time t_n, the measurements Y_{t_n} depend on X_n. Also, conditional on X_n, the distribution of Y_{t_n} is independent of all other variables. Figure 3 shows, on the left panel, $\{X_t,\ t \geq t_0\}$ and, on the corresponding right panel, observations Y_{t_n} simulated via a Poisson measurement model with conditional mean equal to X_n. A population of size $N = 5,000$ is assumed; initial conditions and parameters are specified as in Section 4.4. We use a Gillespie τ leap algorithm ($\tau = 1/100$).

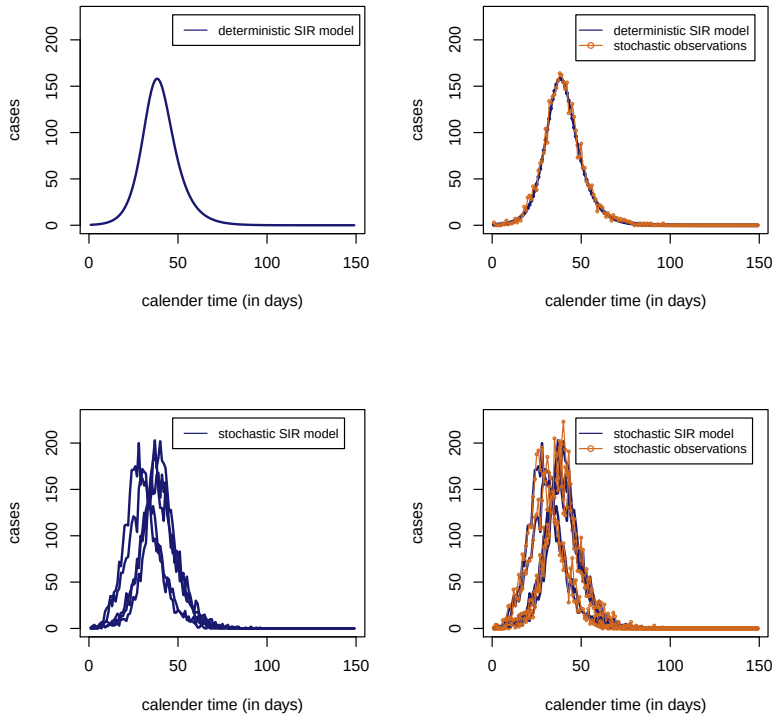

Figure 3: *Top row: left* - deterministic state process; *right* - observations simulated at discrete time points (dots) via a Poisson measurement model with mean given by values of the deterministic state process at the corresponding time points. *Bottom row: left* - 5 realisations of the stochastic state process; *right* - for each realisation of the stochastic state process, observations simulated at discrete time points (dots) via a Poisson measurement model with mean given by values of the stochastic state process at the corresponding time points.

5.2 Overdispersion

So far the nature of stochasticity considered is demographic (Section 3) and observational (Section 5.1). Another variation of stochasticity is environmental stochasticity. This type of stochasticity can affect all individuals equally or can be independent of population levels - its role overtakes that of demographic stochasticity in large populations [45]. Examples of factors that contribute to environmental stochasticity include external unpredictable events [14], e.g. temperature, rainfall or humidity, individual variation (e.g. variability in contacts between infectives and susceptibles, super-spreaders). These factors translate into fluctuations in the transmission parameter, a concept that can be related to overdispersion of the Poisson distribution [44, 45]. As in generalized linear models, failure to account for overdispersion may lead to severe underestimation of standard errors [44]. Environmental stochasticity is typically modelled by randomizing the transmission rate with white noise. One way to incorporate it in the SIR model is to multiply the transmission rate by a Lévy white noise process ξ_t which fluctuates around one. In [44], the process ξ_t is chosen to be $\frac{d\Gamma(t)}{dt}$ where marginally, $\Gamma(t+h) - \Gamma(t) \sim \text{Gamma}(h/\sigma^2, \sigma^2)$. The parameter σ^2 is known as the infinitesimal

variance parameter; it specifies the intensity of the increments of $\Gamma(t)$.

In order to separate environmental stochasticity from possibly overdispersed observational stochasticity, it is common to specify for the measurement model a distribution which accounts for overdispersion. Failure to do so or including measurement process overdispersion in the estimates of environmental stochasticity can lead to biased estimates of key model parameters and may also mask the structure of the underlying process model, see e.g. [47]. As such, accounting for overdispersion associated with the measurement process can facilitate distinguishing observational stochasticity from all other sources of variability [45, 47, 48]. To allow for extra variability in the measurement model it is common to specify a distribution which allows variability to be greater than the Poisson mean. A typical choice is the negative binomial (NB) distribution but other choices are possible, e.g., Poisson-Lognormal model, Poisson-Inverse Gaussian.

5.3 Maximum likelihood inference

Maximum likelihood inference is the standard inference approach in the statistical literature owing to its useful statistical properties, for instance, consistency and asymptotic efficiency, see for example [49]. The rationale behind this inference approach is to find values in the parameter space for which observed data are most likely under the proposed model. Following [43], we provide an overview of the likelihood methodology for estimating parameters governing the POMP model.

The joint density of the state and measurement processes is given by:

$$f\big(x_{0:N}, y_{1:N}; \Theta\big) = f\big(x_0; \Theta\big) \prod_{n=1}^{N} f\big(y_n|x_n; \Theta\big) f\big(x_n|x_{n-1}; \Theta\big), \tag{10}$$

where, $x_{0:N} = \big(x_0, x_1, \cdots, x_N\big)$, $y_{1:N} = \big(y_1, y_2, \cdots, y_N\big)$, Θ is the parameter vector, $f\big(x_n|x_{n-1}; \Theta\big)$ is the transition density, $f\big(y_n|x_n; \Theta\big)$ is the measurement density and, $f\big(x_0; \Theta\big)$ is the initial density.

5.3.1 Inference for deterministic state process

When the state process $\{X_t\}$ is deterministic, noise is confined to the observation process, as such, inference closely resembles nonlinear regression [43, 44]. Assuming that the initial values of the model are given, $\{X_t\}$ is non-random and the likelihood function (12) is given by:

$$\mathcal{L}\big(\Theta\big) = \prod_{n=1}^{N} f\big(y_n|x_n; \Theta\big). \tag{11}$$

Here, since the measurement density is known, the likelihood can be easily evaluated given the value of the state process X_t at each time point t_n. The maximum likelihood estimate $\hat{\Theta}$ can be obtained by optimizing $\mathcal{L}$ or simply $\log \mathcal{L}$ via a non-Bayesian approach using standard numerical methods or via Bayesian Markov chain Monte Carlo (MCMC) approaches.

5.3.2 Inference for stochastic state process

When the state process is stochastic the likelihood function is given by the following high dimensional integral:

$$\begin{aligned} \mathcal{L}(\Theta) &= f(y_{1:N};\Theta) \\ &= \int_{x_{0:N}} f(x_0;\Theta) \prod_{n=1}^{N} f(y_n|x_n;\Theta) f(x_n|x_{n-1};\Theta) dx_{0:N}. \end{aligned} \tag{12}$$

The likelihood function (12) is a high dimensional integral which cannot be solved analytically except in simple cases - its dimensionality depends on numbers of compartments and observations. In the statistical literature data augmentation via Bayesian Markov chain Monte Carlo (MCMC) is popular for parameter estimation in situations of an intractable likelihood [50]. The rationale behind data augmentation is to make the likelihood tractable through introducing parameters which represent the unobserved data; the joint posterior distribution of parameters and augmented data is then explored via MCMC sampling, see e.g. [12, 51, 52, 53]. In principle, data augmentation via Bayesian MCMC is applicable to a model assuming stochastic state processes (Section 3). All unobserved state variables $\{X_t,\ t \geq t_0\}$ can be treated as augmented data, however, doing so results in a large state space which can lead to slow convergence when using standard MCMC algorithms. Moreover, MCMC algorithms can quickly become computationally infeasible because designing and implementing efficient algorithms for high dimensional problems characterized by strong dependencies between states and parameters is methodologically as well as computationally challenging [27, 54]. Considering a discrete state process whose time step is equal to the reporting interval simplifies the problem at the expense of accuracy [12, 55].

Statistical methods for POMP models fall under two main classes namely, state space and information reduction. State space methods work on the unobserved state process $\{X_t,\ t \geq t_0\}$ to estimate both the model parameters and the state process itself - one example is sequential Monte Carlo. On the other other hand, information reduction methods perform inference without having to calculate the likelihood of the observed data - one example is Approximate Bayesian Computation (ABC).

Sequential Monte Carlo

Monte Carlo methods simulate the unobserved model state process and therefore make likelihood evaluation possible. The likelihood function (12) can be rewritten as:

$$\begin{aligned}\mathcal{L}(\Theta) &= \int_{x_{0:N}} \prod_{n=1}^{N} f\big(y_n|x_n;\Theta\big) f\big(x_0;\Theta\big) f\big(x_n|x_{n-1};\Theta\big) dx_{0:N} \\ &= \int_{x_{0:N}} \prod_{n=1}^{N} f\big(y_n|x_n;\Theta\big) f\big(x_{0:N};\theta\big) dx_{0:N} \\ &= \mathbb{E}\bigg[\prod_{n=1}^{N} f\big(y_n|X_n;\Theta\big)\bigg], \end{aligned} \tag{13}$$

where the expectation is taken with respect to $X_{0:N} \sim f\big(x_{0:N};\theta\big)$. By the law of large numbers, the expectation can be approximated by the average:

$$\mathcal{L}(\Theta) \approx \frac{1}{J}\sum_{j=1}^{J}\prod_{n=1}^{N} f\big(y_n|X_n^j;\Theta\big), \tag{14}$$

where $\{X_{0:N}^j, j = 1, 2, \ldots, J\}$ is a Monte Carlo sample of size J drawn from $f\big(x_{0:N};\theta\big)$. Thus, given simulated trajectories $\{X_{0:N}^j, j = 1, 2, \ldots, J\}$, a Monte Carlo estimate of the likelihood can be obtained by evaluating the measurement density of the data at each trajectory and then take the average [43]. However, this approach is inefficient as it is unconditional on the data $y_{1:N}$ - simulated trajectories that diverge from the data will make a negligible contribution to the likelihood estimate and a large number of trajectories will be needed to obtain a precise likelihood estimate useful for estimation.

As an alternative, sequential Monte Carlo directs simulated trajectories of the state process towards values which are consistent with observed measurements. The likelihood function (12) can be rewritten as:

$$\begin{aligned}\mathcal{L}(\Theta) &= f\big(y_{1:N},\Theta\big) \\ &= \prod_{n=1}^{N} f\big(y_n|y_{1:n-1}\big) \\ &= \prod_{n=1}^{N} \int f\big(y_n|x_n;\Theta\big) f\big(x_n|y_{1:n-1};\Theta\big) dx_n \qquad (15) \\ &= \prod_{n=1}^{N} \mathbb{E}\Big[f\big(y_n|x_n;\Theta\big)\Big], \qquad (16) \end{aligned}$$

where the expectation is now taken with respect to the conditional distribution $X_n|Y_{1:n-1} \sim f\big(x_n|y_{1:n-1};\Theta\big)$;

with the understanding that $f(x_1|y_{1:0}) = f(x_1)$. As before, by the law of large numbers, the likelihood can be approximated by the average:

$$\mathcal{L}(\Theta) \approx \prod_{n=1}^{N} \frac{1}{J} \sum_{j=1}^{J} f\big(y_n|X_n^j;\Theta\big), \tag{17}$$

where X_n^j is drawn from $f\big(x_n|y_{1:n-1};\Theta\big)$. On the basis of the Monte Carlo likelihood estimate, standard optimization methods (e.g. Nelder-Mead algorithm) can be applied to obtain parameter estimates. However, as the likelihood estimate is variable, standard optimizers may be susceptible to convergence problems since they converge to local maxima [56]. Stochastic optimization methods, e.g. iterated filtering [57], offer a way to perform a global optimization. In iterated filtering, unknown parameters are included in the state process as time varying and are treated as if they follow a random walk with: $\mathbb{E}\left(\Theta_t|\Theta_{t-1}\right) = \Theta_{t-1}$ and $\mathrm{Var}\big(\Theta_t|\Theta_{t-1}\big) = \phi^2$. As the sequential iterations progress the intensity of the random walk is successively reduced (limit $\phi \to 0$) and the algorithm converges towards the maximum likelihood estimate, see more details in [57]. Iterated filtering has been successfully tested on a variety of complex epidemiological models some of which were computationally intractable for available Bayesian methods [54]. This method is less computationally intensive than its comparable alternatives and has potential to yield more precise results [45, 54]. The method can be easily implemented in the `R pomp` package [24].

Approximate Bayesian Computation

In ABC the goal is to approximate the posterior density $\mathcal{P}(\Theta|y_n)$,

$$\begin{aligned}
\mathcal{P}\big(\Theta|y_n\big) &\propto \mathcal{L}\big(\Theta\big)\mathcal{P}(\Theta) \\
&\propto \prod_{n=1}^{N} f\big(y_n|y_{1:n-1}\big)\mathcal{P}(\Theta) \\
&\propto \prod_{n=1}^{N} \int f\big(y_n|x_n;\Theta\big) f\big(x_n|y_{1:n-1};\Theta\big) dx_n \; \mathcal{P}(\Theta),
\end{aligned} \tag{18}$$

where $\mathcal{P}(\Theta)$ is the prior distribution of the model parameters. The procedure proceeds as follows, see e.g. [24, 27, 58, 59, 60]. Simulate a candidate vector of parameters Θ^* from $\mathcal{P}(\Theta)$. Next, simulate $y_{1:N}^j$ from $f_{y_{1:N}}(\,.\,;\Theta)$. Transform observed data $y_{1:N}$ and simulated $y_{1:N}^j$ into summary statistics z^0 and z^j, respectively. Using a suitably chosen distance measure d, accept Θ^* when $y_{1:N}$ and simulated $y_{1:N}^j$ are sufficiently close, i.e., if $d(z^j, z^0) \leq \varepsilon$ where $\varepsilon \geq 0$, otherwise reject it. The output of the algorithm is an estimate of the posterior density $\mathcal{P}(\Theta|z^0)$; for an appropriate d and small ε, $\mathcal{P}(\Theta|z^0)$ provides a good approximation for $\mathcal{P}(\Theta|y_n)$. Choice of summary statistics and distance measures is a subject of active research, see [24, 27, 58, 59, 60] for main issues and challenges. One ABC approach is implemented in the `R pomp` package [24].

ABC and sequential Monte Carlo are examples of estimation approaches from a range of approaches which have been developed to perform parameter inference when the likelihood is intractable. Other examples include, particle MCMC, synthetic likelihood, nonlinear forecasting, trajectory matching. These methods are described and their application demonstrated in [24, 27, 58]; their weakness and strengths have been studied in [24, 54].

6 Real data example: 1918 influenza pandemic in San Francisco

The city of San Francisco (Northern California, USA) was significantly affected by the 1918 influenza pandemic. At that time the city had a population of approximately 550,000 and 28,310 infected cases were recorded over a period of 63 days between September and November [42]. We use this dataset to compare estimates obtained from different transmission models incorporating varying levels of stochasticity. Models considered are: i) deterministic SIR model, ii) stochastic SIR model and iii) stochastic SIR model incorporating environmental stochasticity; each of these models is fitted with Poisson and negative binomial measurement models. We do not assert that the SIR model is fully adequate for this dataset. Several underlying assumptions of this model could be at odds with the data, e.g., it does not account for the latent period, also, the mass action principle may be an oversimplification because in practice, within a population an individual has a finite number of contacts which are not necessarily random. Nevertheless, we use the model for illustration purposes.

We estimate the transmission rate β and use it to compute R_0; the removal rate is taken to be $1/4.1$ days [42]; where applicable we also estimate the dispersion parameter of the NB measurement model or the infinitesimal standard deviation of the transmission model (Section 5.2). We consider these models on a discrete scale (h=0.01). For the stochastic models, a τ-leap algorithm is used for simulating the underlying transmission model ($\tau = 0.01$). The models are fitted to the first 28 data points of the ascending phase using sequential Monte Carlo for evaluating the likelihood and iterated filtering for optimizing the likelihood function. We use the Akaike Information Criterion (AIC) to compare the different models; in the analysis of outbreak data, this model comparison tool has been found to perform well in detecting potential mispecification in the transmission model [61]. The AIC is calculated as,

$$\text{AIC} = -2 \log L\big(\Theta\big) + 2p, \tag{19}$$

where p is the dimension of Θ. It acts as a penalized log-likelihood criterion, providing a trade off between a

good fit (high value of log-likelihood) and complexity (models with larger p are penalized more than those with smaller p). Among a set of candidate models the 'best' model is the one with the smallest AIC [62]. A general rule of thumb is that models that differ in AIC by more than two units are generally considered to 'differ' in terms of fit.

Table 3 shows parameter estimates, 95% confidence intervals, AIC values and computation time for six different models fitted to the dataset. In terms of goodness of fit, for all three transmission models, a NB measurement model provides a better fit than a Poisson measurement model. All the three models assuming a NB measurement model yield a similar fit to these data (see also Figure 4). Estimates of β and R_0 are all within the same range for all the six models.

Table 3: Parameter estimates, 95% confidence intervals, AIC values and computation time. The models were fitted on a 3.1 GHz PC with 4 cores. θ denotes the dispersion parameter of the negative binomial (NB) measurement model ($\theta = 0 \equiv$ Poisson model). σ represents environmental stochasticity (Section 5.2). Stochastic SIR* represents the stochastic SIR model incorporating environmental stochasticity.

Transmission model	Measurement model	Estimate (95% confidence interval)				AIC	time (seconds)
		β	R_0	θ	σ		
Deterministic SIR	Poisson	0.494 (0.492, 0.496)	2.026 (2.019, 2.034)	-	-	442.380	250
	NB	0.495 (0.491, 0.500)	2.029 (2.011, 2.048)	0.034 (0.011, 0.063)	-	313.147	259
Stochastic SIR	Poisson	0.477 (0.422, 0.501)	1.955 (1.730, 2.053)	-	-	337.622	2716
	NB	0.494 (0.451, 0.519)	2.024 (1.849, 2.128)	0.022 (0.002, 0.064)	-	317.611	2840
Stochastic SIR*	Poisson	0.461 (0.399, 0.499)	1.889 (1.638, 2.045)	-	0.162 (0.015, 0.241)	393.726	6107
	NB	0.494 (0.439, 0.532)	2.026 (1.801, 2.181)	0.001 (0.0003, 0.071)	0.117 (0.003, 0.154)	317.204	7234

For each measurement model, the greater the level of stochasticity in the transmission model the wider the confidence intervals (Deterministic SIR $<$ Stochastic SIR $<$ Stochastic SIR*). This is expected when data suggest presence of overdispersion ($\theta \neq 0$ and $\sigma \neq 0$), see e.g. [19]. This indicates that deterministic models have a potential pitfall to underestimate uncertainty associated with parameter estimates and consequently key epidemiological parameters calculated therefrom. As such, stochastic models should be preferred over deterministic models as they offer improved accounting for variability in the data and improved quantification of uncertainty. Moreover, stochastic terms which account for environmental stochasticity can, to some extent, compensate for model mispecification resulting in even greater uncertainty [11].

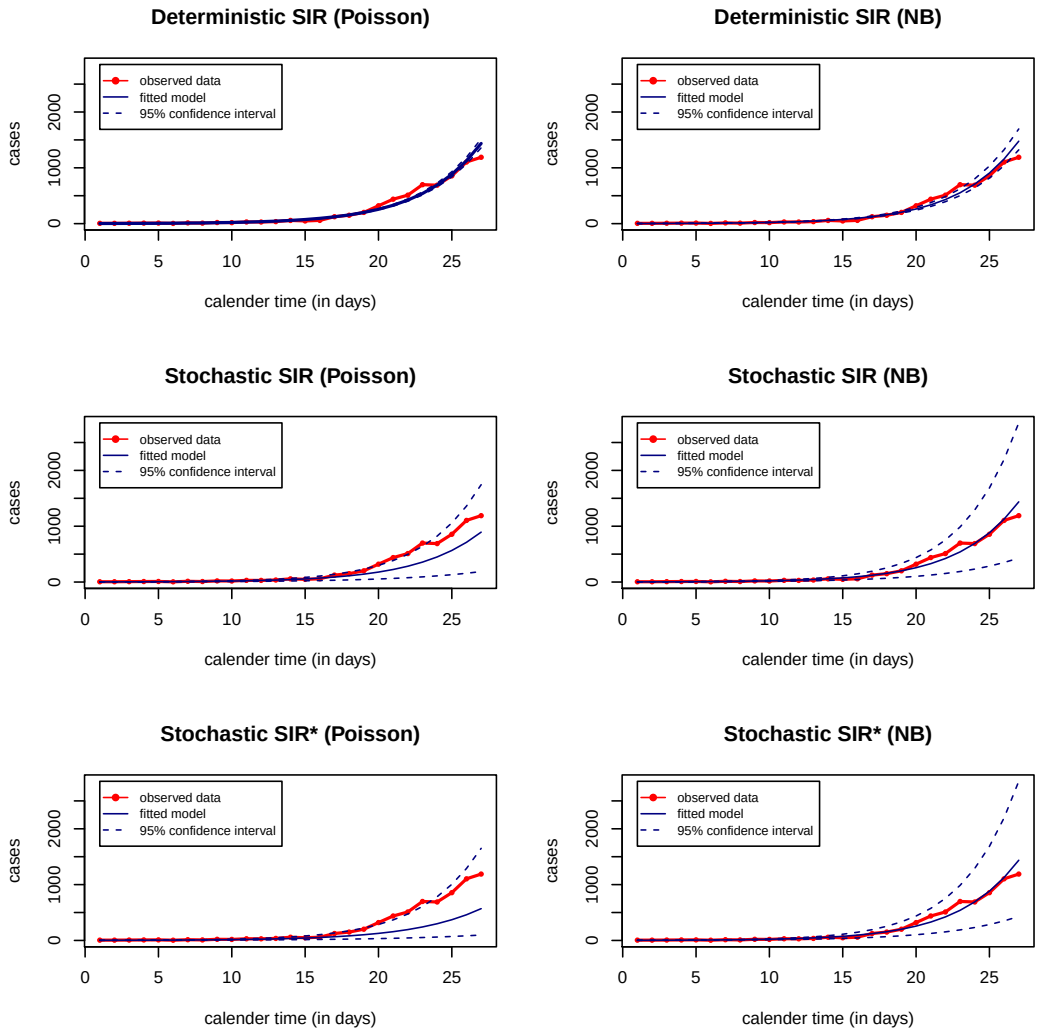

Figure 4: Comparison of observed data (red) and fitted models (navy blue).

7 Discussion and Conclusions

The goal of this paper was to give an overview of basic concepts behind simulation and statistical analysis of infectious disease outbreaks using stochastic compartmental models in well mixed populations. Data simulation is important for studying qualitative and quantitative features of compartmental models. It also facilitates likelihood estimation in cases where the likelihood has no closed form [24]. Moreover, simulated data can be used for testing the performance of estimation methods as well as to investigate the estimability of parameters from available data, see e.g. [18, 19]. In terms of estimation, we showed that, compared to deterministic models, stochastic models offer an opportunity to better quantify uncertainty of parameter estimates. We also showed that, when data suggest presence of overdispersion, incorporating overdispersion

in the transmission model or the measurement model can improve model fit as well as yield less optimistic parameter estimates, see e.g. [11, 19, 61].

The SIR model used through out this paper assumes a well mixed population. This assumption implies that an infective is equally likely to infect any other susceptible in the whole population and that all infectives have the same number of contacts. In reality, an individual only has contact with a small fraction of the whole population. Network models relax the 'well mixed' assumption by assigning to each individual in the population a finite set of contacts (links) [63]. In the network, individuals in the population are represented by vertices and links are represented by edges. Infection can be transmitted via an edge - for each edge between an infective and a susceptible, it is assumed that there is a probability that an infection will be transmitted. The network approach can be integrated into the class of compartmental models, see e.g. [3, 63]. Though network-based compartmental models can be formulated, simulation and analysis of the resulting models depends on the network specification, i.e., different networks will lead to different qualitative behavior of the model trajectory as well as different quantitative features of the model (peak time, peak incidence, duration of epidemic and final size), see e.g. [64]. Instead of assuming that the population being modelled is homogeneous and describing a disease system only with variables representing the state of the whole system, agent based models (ABMs), also known as individual-based models, capture interactions and behaviour at the individual level by representing how individuals and the environmental variables that affect them, vary over space, time, or other dimensions. In ABMs individuals are explicitly assumed to be unique and autonomous. Individuals are usually assumed to be different from each other with respect to characteristics as age, gender, health status or behaviour. Autonomy means that individuals act independently of each other and pursue their own objectives. Epidemiological ABMs mainly consist of four components, namely, disease, society, transportation, and the environment. All four components are to be modelled when formulating an ABM (typically jointly). Modelling the disease involves describing how the infectious disease is transmitted between individuals and how the disease progresses in an infected agent. Modelling the society involves simulating the population. Modelling transportation concerns how the individuals will move within the environment. Modelling the environment involves creating the space in which the individuals will interact. An advantage of ABMs is that they allow for more flexibility and a large amount of freedom in the model structure; a disadvantage is that they are computationally intensive and may require long running time, see e.g. [65, 66].

On fitting models to data we side-stepped parameter identifiability which is an important issue from an estimation point of view. Compartmental model parameters are typically difficult to identify due to inherent

model nonlinearities - if parameters are not well identified, epidemiological and biological questions which these models seek to address may not be addressable because parameter estimates will be unrealiable [67]. As such, a fundamental prerequisite to parameter estimation is investigation of structural and practical identifiability. The former concerns studying which parameters are functionally related in a way that they cannot be uniquely determined under noise-free conditions. On the other hand, the latter concerns studying how well parameters can be determined by the quantity and quality of data together with the estimation method used - few data are typically available during the early stages of a disease outbreak; issues that may affect the quality of data include underreporting and reporting delays, see e.g. [25]. A parameter that is structurally identifiable can be practically nonidentifiable, on the other hand, if a parameter is structurally nonidentifiable it is practically nonidentifiable as well. Structural identifiability can be recognized when large parameter variations yield small changes in model output; practical identifiability can be recognized when confidence intervals are not too wide, see [68, 67, 69] and references therein.

In closing, stochastic modelling of infectious diseases, an area which encompasses simulation and analysis, is a subject of enormous research. An important caveat of this paper is that it is by no means exhaustive, nevertheless, it is envisaged that it provides a useful background on the essentials of simulation and analysis of stochastic compartmental models.

Acknowledgements

The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government – department EWI.

References

[1] T D Hollingsworth. “Controlling infectious disease outbreaks: Lessons from mathematical modelling”. In: *Journal of Public Health Policy* 30.3 (2009), pp. 328–341.

[2] R M Anderson and R M May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1992.

[3] F Brauer and C Castillo-Chávez. *Mathematical Models in Population Biology and Epidemiology*. Springer New York, 2001.

[4] D J Daley and J Gani. *Epidemic modelling: an introduction*. Vol. 15. Cambridge University Press, 2001.

[5] G Chowell et al., eds. *Mathematical and Statistical Estimation Approaches in Epidemiology*. Springer Netherlands, 2009.

[6] N Hens et al. *Modeling Infectious Disease Parameters Based on Serological and Social Contact Data*. Springer New York, 2012.

[7] J K Taubenberger and D M Morens. "1918 Influenza: the mother of all pandemics". In: *Revista Biomedica* 17.1 (2006), pp. 69–79.

[8] E Dong, H Du, and L Gardner. "An interactive web-based dashboard to track COVID-19 in real time". In: *The Lancet infectious diseases* 20.5 (2020), pp. 533–534.

[9] T Kypraios and V N Minin. "Introduction to the Special Section on Inference for Infectious Disease Dynamics". In: *Statistical Science* 33.1 (2018), pp. 1–3.

[10] C Ozcaglar et al. "Epidemiological models of Mycobacterium tuberculosis complex infections". In: *Mathematical Biosciences* 236.2 (2012), pp. 77–96.

[11] A A King et al. "Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola". In: *Proceedings of the Royal Society B: Biological Sciences* 282.1806 (2015), p. 20150347.

[12] P E Lekone and B F Finkenstädt. "Statistical Inference in a Stochastic Epidemic SEIR Model with Control Intervention: Ebola as a Case Study". In: *Biometrics* 62.4 (2006), pp. 1170–1177.

[13] S Mandal, R Sarkar, and S Sinha. "Mathematical models of malaria - a review". In: *Malaria Journal* 10.1 (2011), p. 202.

[14] M J Keeling and P Rohani. *Modeling infectious diseases in humans and animals*. Princeton University Press, 2008.

[15] N T J Bailey. "Some Problems in the Statistical Analysis of Epidemic Data". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 17.1 (1955), pp. 35–58.

[16] M S Bartlett. "Measles Periodicity and Community Size". In: *Journal of the Royal Statistical Society. Series A (General)* 120.1 (1957), pp. 48–70.

[17] H Andersson and T Britton. *Stochastic epidemic models and their statistical analysis*. Vol. 151. Springer Science & Business Media, 2000.

[18] T Ganyani et al. "Assessing inference of the basic reproduction number in an SIR model incorporating a growth-scaling parameter". In: *Statistics in Medicine* 37.29 (2018), pp. 4490–4506.

[19] T Ganyani, C Faes, and N Hens. "Inference of the generalized-growth model via maximum likelihood estimation: A reflection on the impact of overdispersion". In: *Journal of Theoretical Biology* 484 (2020), p. 110029.

[20] M S Bartlett. "Monte Carlo studies in ecology and epidemiology". In: *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*. Vol. 4. University of California Press London, UK. 1961, pp. 39–55.

[21] D L Chao et al. "FluTE, a publicly available stochastic influenza epidemic simulation model". In: *PLoS computational biology* 6.1 (2010).

[22] M E Halloran et al. "Simulations for designing and interpreting intervention trials in infectious diseases". In: *BMC medicine* 15.1 (2017), pp. 1–8.

[23] J Kaminsky et al. "Perfect counterfactuals for epidemic simulations". In: *Philosophical Transactions of the Royal Society B* 374.1776 (2019), p. 20180279.

[24] A A King, D Nguyen, and E L Ionides. "Statistical Inference for Partially Observed Markov Processes via the R Package pomp". In: *Journal of Statistical Software* 69.12 (2016).

[25] L Held et al. *Handbook of infectious disease data analysis*. Chapman and Hall/CRC, 2019.

[26] P D O'Neill. "Introduction and snapshot review: Relating infectious disease transmission models to data". In: *Statistics in Medicine* 29.20 (2010), pp. 2069–2077.

[27] T J McKinley et al. "Approximate Bayesian Computation and Simulation-Based Inference for Complex Stochastic Epidemic Models". In: *Statistical Science* 33.1 (2018), pp. 4–18.

[28] M Kermark and A McKendrick. "Contributions to the mathematical theory of epidemics. Part I". In: *Proceedings of the Royal Society, Series A* 115.5 (1927), pp. 700–721.

[29] H Heesterbeek. "The law of mass-action in epidemiology: a historical perspective". In: *Ecological paradigms lost: routes of theory change* (2005), pp. 81–104.

[30] H McCallum. "How should pathogen transmission be modelled?" In: *Trends in Ecology and Evolution* 16.6 (2001), pp. 295–300.

[31] O Diekmann, J A P Heesterbeek, and J A J Metz. "On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations". In: *Journal of Mathematical Biology* 28.4 (1990), pp. 365–382.

[32] M J Keeling and J V Ross. "Efficient methods for studying stochastic disease and population dynamics". In: *Theoretical Population Biology* 75.2-3 (2009), pp. 133–141.

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

[34] A T Bharucha-Reid and M S Bartlett. "Stochastic Population Models in Ecology and Epidemiology". In: *Biometrics* 18.2 (1962), p. 253.

[35] L Allen. "A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis". In: *Infectious Disease Modelling* 2.2 (2017), pp. 128–142.

[36] D T Gillespie. "Exact stochastic simulation of coupled chemical reactions". In: *The Journal of Physical Chemistry* 81.25 (1977), pp. 2340–2361.

[37] D T Gillespie. "Approximate accelerated stochastic simulation of chemically reacting systems". In: *The Journal of Chemical Physics* 115.4 (2001), pp. 1716–1733.

[38] D J Wilkinson. *Stochastic modelling for systems biology*. CRC press, 2018.

[39] M Pineda-Krch. "GillespieSSA: implementing the stochastic simulation algorithm in R". In: *Journal of Statistical Software* 25.12 (2008), pp. 1–18.

[40] C Fuchs. *Inference for diffusion processes: with applications in life sciences*. Springer Science & Business Media, 2013.

[41] B J Coburn, B G Wagner, and S Blower. "Modeling influenza epidemics and pandemics: insights into the future of swine flu (H1N1)". In: *BMC medicine* 7.1 (2009), pp. 1–8.

[42] C Gerardo, N Hiroshi, and L M A Bettencourt. "Comparative estimation of the reproduction number for pandemic influenza from daily case notification data". In: *Journal of the Royal Society Interface* 4.12 (2007), pp. 155–166.

[43] A A King. *Introduction to partially observed Markov processes*. `https://kingaa.github.io/short-course/intro/intro.html`. 2018.

[44] C Bretó et al. "Time series analysis via mechanistic models". In: *The Annals of Applied Statistics* 3.1 (2009), pp. 319–348.

[45] C Bretó. "Modeling and inference for infectious disease dynamics: a likelihood-based approach". In: *Statistical science: a review journal of the Institute of Mathematical Statistics* 33.1 (2018), p. 57.

[46] T Coulson, P Rohani, and M Pascual. "Skeletons, noise and population growth: the end of an old debate?" In: *Trends in Ecology and Evolution* 19.7 (2004), pp. 359–364.

[47] K Nadeem et al. "Integrating population dynamics models and distance sampling data: a spatial hierarchical state-space approach". In: *Ecology* 97.7 (2016), pp. 1735–1745.

[48] M Fujiwara and T Takada. "Environmental stochasticity". In: *Encyclopedia of Life Sciences (ELS)* (2001), pp. 1–8.

[49] A M Mood, F A Graybill, and D C Boes. *Introduction to the Theory of Statistics*. McGraw-hill, 1950.

[50] D A Van Dyk and X Meng. "The art of data augmentation". In: *Journal of Computational and Graphical Statistics* 10.1 (2001), pp. 1–50.

[51] G J Gibson and E Renshaw. "Likelihood estimation for stochastic compartmental models using Markov chain methods". In: *Statistics and Computing* 11.4 (2001), pp. 347–358.

[52] G J Gibson and E Renshaw. "Estimating parameters in stochastic compartmental models using Markov chain methods". In: *Mathematical Medicine and Biology: A Journal of the IMA* 15.1 (1998), pp. 19–40.

[53] P D O'Neill and G O Roberts. "Bayesian inference for partially observed stochastic epidemics". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 162.1 (1999), pp. 121–129.

[54] M Fasiolo, N Pya, and S N Wood. "A Comparison of Inferential Methods for Highly Nonlinear State Space Models in Ecology and Epidemiology". In: *Statistical Science* 31.1 (2016), pp. 96–118.

[55] M Li, J Dushoff, and B M Bolker. "Fitting mechanistic epidemic models to data: A comparison of simple Markov chain Monte Carlo approaches". In: *Statistical methods in medical research* 27.7 (2018), pp. 1956–1967.

[56] P Eberhard, W Schiehlen, and D Bestle. "Some advantages of stochastic methods in multicriteria optimization of multibody systems". In: *Archive of Applied Mechanics (Ingenieur Archiv)* 69.8 (1999), pp. 543–554.

[57] E L Ionides, C Bretó, and A A King. "Inference for nonlinear dynamical systems". In: *Proceedings of the National Academy of Sciences* 103.49 (2006), pp. 18438–18443.

[58] T Kypraios, P Neal, and D Prangle. "A tutorial introduction to Bayesian inference for stochastic epidemic models using Approximate Bayesian Computation". In: *Mathematical Biosciences* 287 (2017), pp. 42–53.

[59] M A Beaumont. "Approximate bayesian computation". In: *Annual review of statistics and its application* 6 (2019), pp. 379–403.

[60] K Csilléry et al. "Approximate Bayesian computation (ABC) in practice". In: *Trends in ecology and evolution* 25.7 (2010), pp. 410–418.

[61] T Stocks, T Britton, and M Höhle. "Model selection and parameter estimation for dynamic epidemic models via iterated filtering: application to rotavirus in Germany". In: *Biostatistics* 21.3 (2020), pp. 400–416.

[62] G Claeskens and N L Hjort. *Model selection and model averaging*. Cambridge University Press, 2008.

[63] M E J Newman. "Spread of epidemic disease on networks". In: *Physical review E* 66.1 (2002), p. 016128.

[64] M J Keeling and K T D Eames. "Networks and epidemic models". In: *Journal of the Royal Society Interface* 2.4 (2005), pp. 295–307.

[65] S F Railsback and V Grimm. *Agent-Based and Individual-Based Modeling: A Practical Introduction*. Princeton University Press, 2011.

[66] E Hunter, B Mac Namee, and J D Kelleher. "A taxonomy for agent-based models in human infectious disease epidemiology". In: *Journal of Artificial Societies and Social Simulation* 20.3 (2017).

[67] A Raue et al. "Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood". In: *Bioinformatics* 25.15 (2009), pp. 1923–1929.

[68] N Tuncer and T T Le. "Structural and practical identifiability analysis of outbreak models". In: *Mathematical Biosciences* 299 (2018), pp. 1–18.

[69] G Chowell. "Fitting dynamic models to epidemic outbreaks with quantified uncertainty: a primer for parameter uncertainty, identifiability, and forecasts". In: *Infectious Disease Modelling* 2.3 (2017), pp. 379–398.