ANNUAL REVIEWS

Annual Review of Statistics and Its Application

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, 3500 Hasselt, Belgium;
email: tapiwa.ganyani@uhasselt.be

[2]Centre for Health Economics Research and Modelling Infectious Diseases (CHERMID), Vaccine and Infectious Disease Institute, University of Antwerp, 2610 Antwerp, Belgium

Annu. Rev. Stat. Appl. 2021. 8:69–88

First published as a Review in Advance on November 16, 2020

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-061120-034438

Keywords

stochastic SIR model, stochasticity, simulation, estimation, epidemic modeling

Abstract

This article 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 data set and compare models assuming different kinds of stochasticity. We also give references for other publications where detailed information on this topic can be found.

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 (Hollingsworth 2009). 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 so that their parameters have a natural epidemiological or biological interpretation (Anderson & May 1992). 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 (Brauer & Castillo-Chávez 2001, Hollingsworth 2009). They provide conceptual results, such as reproduction numbers (average number of infections caused by an infectious individual), the herd immunity threshold (population fraction to be immunized to stop the infection from spreading in the entire population), the final size (number of individuals who ultimately become infected during the entire period of the epidemic), and the future course of the epidemic (Anderson & May 1992, Daley & Gani 2001, Chowell et al. 2009, Hens et al. 2012). 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 1918–1919: Other influenza pandemics, such as H1N1/09, H2N2, and H3N2, have been caused by descendants of the 1918 virus (Taubenberger & Morens 2006). 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 400,000 deaths as of June 9, 2020 (Dong et al. 2020).

In this article, for illustrative purposes, we consider one example of a compartmental epidemic model, the susceptible-infected-removed (SIR) model in a well-mixed population. This model is important in infectious disease modeling because all compartmental epidemic models can be thought of as extensions of the SIR model (Brauer & Castillo-Chávez 2001, Kypraios & Minin 2018). For example, in the context of Ebola virus disease, 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 (Ozcaglar et al. 2012). 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., Brauer & Castillo-Chávez 2001, Lekone & Finkenstädt 2006, Hens et al. 2012, Ozcaglar et al. 2012, King et al. 2015). 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, vector-borne transmission for diseases such as malaria, where mosquitoes are the vectors (Mandal et al. 2011).

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 way. They do not account for the fact that different individuals in the population have different nutritional, environmental, or genetic statuses and may therefore have different epidemiological behaviors, making infectious disease spread a random process (Anderson & May 1992, Daley & Gani 2001, Keeling & Rohani 2008). These models are 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, so deterministic models become unsuitable (Anderson & May 1992, Daley & Gani 2001, Keeling & Rohani 2008). 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 (Daley & Gani 2001). It is therefore well recognized that stochastic compartmental epidemic models are important when studying infectious disease dynamics (Bailey 1955, Bartlett 1957, Daley & Gani 2001). From an estimation viewpoint, stochastic models enable quantification of uncertainty for parameters estimated from disease outbreak data (Andersson & Britton 2000).

Simulation of stochastic models plays an important role in the modeling of infectious diseases using compartmental models. It allows understanding of the qualitative behavior of models (e.g., sensitivity of solutions to different assumptions; see Ganyani et al. 2018, 2020). It also enables the quantitative study of problems that are mathematically intractable—simulations are necessary for detailed models that are difficult or impossible to solve analytically (Bartlett 1961, Brauer & Castillo-Chávez 2001). Also, strategies to mitigate or delay the impact of future seasonal or pandemic outbreaks can be explored via simulation (see, e.g., Chao et al. 2010, Halloran et al. 2017, Kaminsky et al. 2019). Moreover, analysis of simulated data can be used to evaluate the performance of estimation algorithms as well as to investigate which model parameters or combinations of parameters are estimable from data (see, e.g., Lekone & Finkenstädt 2006; King et al. 2015, 2016). From an estimation point of view, the ability to simulate from models comes in handy for problems where the likelihood function is not tractable (King et al. 2016).

Statistical analysis of infectious disease data has been geared toward estimating model parameters from observed infectious disease data (Held et al. 2019); therefore, accurate estimation of model parameters from data is of the utmost importance. Parameter estimation methods for stochastic compartmental models usually rely on calculating the likelihood. Often, due to the interplay of a proposed model and the observed data, the likelihood is intractable, leading to estimation challenges (O'Neill 2010). With the advent of modern computing power, the problem of intractable likelihood has led to the continual development of statistical methodologies for parameter inference, such as Markov chain Monte Carlo (MCMC) methods, approximate Bayesian computation (ABC) methods, and sequential Monte Carlo (see, e.g., O'Neill 2010, King et al. 2016, McKinley et al. 2018, and references therein).

This review provides an overview of basic ideas on data simulation and statistical analysis for stochastic compartmental models. We 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 SUSCEPTIBLE-INFECTED-REMOVED MODEL

In a closed population with no births or deaths, a simplified version of the SIR model (Kermack & McKendrick 1927) assumes that a population is divided into three compartments (**Figure 1**): $S(t)$,

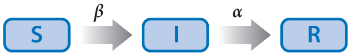

Figure 1

Flowchart for the SIR (susceptible-infected-removed) model. β represents the rate at which individuals come into effective contact per unit time, and α represents the rate at which infected individuals recover.

the number of susceptible individuals at time t; $I(t)$, the number of infectious individuals at time t who are capable of infecting susceptible individuals; and $R(t)$, the 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 infective or a small number of infectives, and 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.

Assuming that individuals within a population mix completely and move randomly (homogeneous mixing), the total number of possible contacts between susceptibles and infectives is given by $S(t)I(t)$ (a principle called mass action) (Heesterbeek 2005). 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 of $\beta S(t)I(t)$. 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 (McCallum 2001). In continuous time, the SIR model is described by the following set of ordinary differential equations, referred to hereafter as Model 1:

$$\begin{aligned}\frac{\mathrm{d}S(t)}{\mathrm{d}t} &= -\frac{\beta S(t)I(t)}{M}\\ \frac{\mathrm{d}I(t)}{\mathrm{d}t} &= \frac{\beta S(t)I(t)}{M} - \alpha I(t)\\ \frac{\mathrm{d}R(t)}{\mathrm{d}t} &= \alpha I(t),\end{aligned} \qquad 1.$$

where $1/\alpha$ is the average infectious period and t represents calendar time. A common initial condition of Model 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 (Diekmann et al. 1990).

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 Model 1 can be formulated as Model 2:

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

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

3. THE GENERAL STOCHASTIC SUSCEPTIBLE-INFECTED-REMOVED MODEL

Most stochastic models for compartmental epidemic models are Markov processes (Keeling & Ross 2009). 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., Ross 2014). Let $M_I(t)$ and $M_R(t)$ be Poisson processes that 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 Bartlett (1960), the standard stochastic version of Model 1, also known as the general stochastic model, is defined as a bivariate continuous-time Markov process (Markov jump process) $\{(S(t), I(t)) \colon t \geq 0\}$. For $h > 0$, it is specified by the following infinitesimal increment probabilities, referred to hereafter as Model 3:

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

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 nonhomogeneous 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 modeling, the nature of stochasticity represented by model in 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 (McKinley et al. 2018).

4. SIMULATION

4.1. Kolmogorov Forward Equations

From the stochastic model (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 behavior of the stochastic system (see, e.g., Keeling & Ross 2009). For an SIR model the master equations are given by

$$\frac{\mathrm{d}p_{S,I}(t)}{\mathrm{d}t} = p_{S+1,I-1}(t)\left(\frac{\beta}{M}(S+1)(I-1)\right) + p_{S,I+1}(t)\left(\alpha(I+1)\right) - p_{S,I}(t)\left(\frac{\beta}{M}SI + \alpha I\right), \qquad 4.$$

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 Keeling & Ross 2009, Allen 2017 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 (Equation 4) (Allen 2017). For large populations stochastic simulation is handy. A commonly used stochastic simulation algorithm (SSA) for Markovian models is the Gillespie algorithm (Gillespie 1977). The algorithm simulates an exact stochastic version of the trajectory that would be obtained by solving the corresponding master equations.

4.2. The Gillespie Stochastic Simulation Algorithm

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 three-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 a 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 cs as there are reactions), Z is distributed as:

$$g_Z(z) = \left(\sum_{i=1}^{2} c_i\right) \exp\left(-z\sum_{i=1}^{2} c_i\right). \qquad 5.$$

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:

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

Using these distributions, the algorithm proceeds as follows:

(*a*) Set initial population numbers.
(*b*) Calculate c_1 and c_2.
(*c*) Simulate time to next event from Equation 5.
(*d*) Simulate which event occurs from Equation 6.
(*e*) Update the population sizes in line with the event that occurred.
(*f*) Update the time and return to step *b*.

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

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 (Gillespie 2001). Accuracy is achieved for small τ (leap condition) (Gillespie 2001, Pineda-Krch 2008, Wilkinson 2018).

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 follows 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:

(*a*) Set initial population numbers.
(*b*) Calculate c_1 and c_2.
(*c*) Simulate the number of events of each type from $\mathrm{Po}(c_1\tau)$ and $\mathrm{Po}(c_2\tau)$.
(*d*) Update the population sizes in line with the events that occurred.
(*e*) Update the time to $t + \tau$ and return to step *b*.

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 (Pineda-Krch 2008). 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 (Gillespie 2001, Wilkinson 2018).

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 (Fuchs 2013). An SDE of a continuous process Y is usually of the form

$$\mathrm{d}Y(t) = \mu\big(Y(t)\big)\mathrm{d}t + \Psi\big(Y(t)\big)\mathrm{d}W(t), \qquad 7.$$

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 (Equation 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 (Fuchs 2013, Wilkinson 2018). The SDEs for the SIR are given by

$$\begin{aligned}
\mathrm{d}S(t) &= -\big(\beta S(t)I(t)/M\big)\mathrm{d}t - \sqrt{\beta S(t)I(t)/M}\,\mathrm{d}W_1(t)\\
\mathrm{d}I(t) &= \big(\beta S(t)I(t)/M - \alpha I(t)\big)\mathrm{d}t + \sqrt{\beta S(t)I(t)/M}\,\mathrm{d}W_1(t) - \sqrt{\alpha I(t)}\,\mathrm{d}W_2(t)\\
\mathrm{d}R(t) &= \alpha I(t)\mathrm{d}t + \sqrt{\alpha I(t)}\,\mathrm{d}W_2(t),
\end{aligned} \qquad 8.$$

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 $\mathrm{d}W_i \sim N(0, \mathrm{d}t)$ [see derivations in Allen (2017)]. 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 (Fuchs 2013). The approximation proceeds on a regular grid t_0, $t_0 + \Delta$, $t_0 + 2\Delta, \ldots, t_n$ as

Table 1 Time required to simulate the stochastic SIR model (Model 3) using the Gillespie SSA and Poisson τ-leap ($\tau = 1/1{,}000$) approaches

M	Time to perform simulation (s)	
	Gillespie SSA	Gillespie τ-leap
10,000	33.17	32.60
100,000	66.39	67.14
500,000	1,405.01	65.99
750,000	3,218.06	77.70
1,000,000	5,456.66	64.13

Simulations are performed using the R package `GillespieSSA` on a 3.1 GHz PC. *M* is the population size, as defined in Section 2. Abbreviations: PC, personal computer; SIR, susceptible-infected-removed; SSA, stochastic simulation algorithm.

follows:

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

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

4.4. Comparison of Simulation Procedures

Table 1 shows the computational time required to simulate one realization of the stochastic SIR model for different population sizes using the Gillespie SSA and the Poisson τ-leap method ($\tau = 1/1{,}000$). Simulations are performed using the R package `GillespieSSA`. Computational time for SDE is not included for two reasons. First, the Euler-Maruyama method used to simulate SDEs is closely related to τ-leaping approaches; hence, it is expected that computational time for SDEs would be similar to that for τ-leaping approaches. Second, 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.

For the Gillespie SSA, computational time is high and increases rapidly with increasing population size. In contrast, for τ-leaping, there is huge gain in computation speed; moreover, computational time does not increase for larger population sizes (Keeling & Rohani 2008).

Table 2 shows, for epidemics simulated using the Gillespie SSA, Poisson τ-leap ($\tau = 1/1{,}000$) and SDE ($\Delta = 1/1{,}000$), a descriptive summary of peak time (amount of time from epidemic onset

Table 2 Descriptive summary of peak time, peak incidence, epidemic duration, and final size for epidemics simulated using the Gillespie SSA, Poisson τ-leap ($\tau = 1/1{,}000$), and SDE ($\Delta = 1/1{,}000$)

Simulation procedure	2.5%, 50%, and 97.5% percentiles			
	Peak time (days)	Peak incidence	Epidemic duration (days)	Final size
Gillespie SSA	(43, 50, 64)	(3,054, 3,197, 3,358)	(114, 129, 150)	(72,500, 73,246, 73,875)
Gillespie τ-leap	(43, 51, 67)	(3,055, 3,193, 3,350)	(119, 137, 165)	(72,537, 73,224, 73,993)
SDE	(53, 54, 54)	(3,153, 3159, 3,165)	(144, 145, 147)	(73,223, 73,244, 73,268)
True values	54	3,159	145	73,244

Abbreviations: SDE, stochastic differential equation; SSA, stochastic simulation algorithm.

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 follows: $S(0) = M - 1$, $I(0) = 1$, and $R(0) = 0$ with $M = 100{,}000$; $R_0 = 1.8$ is chosen similar to R_0 for influenza (Coburn et al. 2009) with a recovery rate $\alpha = 1/4.1$ days (Chowell et al. 2007) corresponding to a transmission rate of $\beta = \alpha R_0$. For each method, 650 epidemics were simulated.

As expected, summary statistics for the Gillespie SSA and the τ-leap approaches are in the same ranges since the τ-leap algorithm approximates the Gillespie SSA well for small τ (Gillespie 2001, Pineda-Krch 2008). Summaries for these two are also fairly close to the true values; this is also expected since demographic stochasticity averages out in large populations. Summaries for SDE are very close to the true values because the noise term is very small, i.e., $\sim N(0, 1/1{,}000)$; 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 with 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 behavior 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. On the one hand, the recovery rate and 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 (Anderson & May 1992).

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 moments at which events occur are never observed—recorded data are an aggregation of events that occurred between consecutive reporting periods, with the exact time of occurrence for each event unknown. Also, it is often the case that data on at least one compartment are unobserved. Data incompleteness presents an estimation problem when performing likelihood-based inference as it leads to an intractable likelihood (Lekone & Finkenstädt 2006, O'Neill 2010). In this section we first introduce the partially observed Markov process (POMP) modeling framework (following King et al. 2018), 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 (Bretó 2018, Bretó et al. 2009). Thereafter, we provide an overview of the key aspects regarding likelihood-based inference.

5.1. Relating Models to Data: The Partially Observed Markov Process 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 be either discrete or continuous (Section 2). In this case, $\{X_t, t \geq t_0\}$ represents the

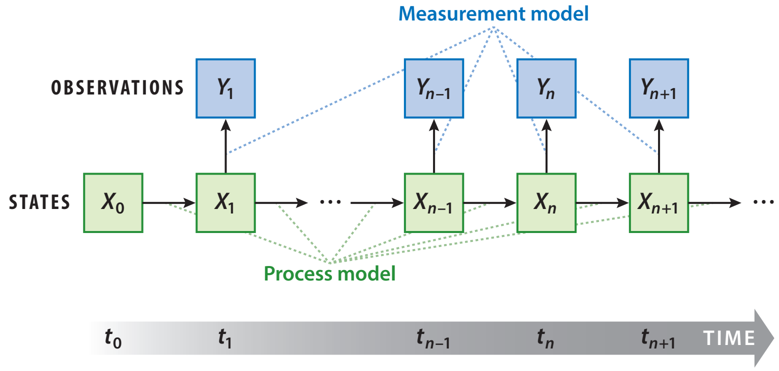

Figure 2

Illustration of a POMP model: Y_{t_n} denotes observed incidence data at discrete time points connected to the continuous state process X_t at that time point. Adapted with permission from King et al. (2018). Abbreviation: POMP, partially observed Markov process.

incidence trajectory obtained from the SIR model (process model) which can be either 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., $\mathrm{P}(X_n|X_{n-1}, \ldots, X_0) = \mathrm{P}(X_n|X_{n-1})$. 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 $\{t_1, \ldots, t_N\}$ 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 misdiagnosis (Coulson et al. 2004, Keeling & Rohani 2008).

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*a*** shows $\{X_t,\ t \geq t_0\}$, and **Figure 3*b*** shows observations Y_{t_n} simulated via a Poisson measurement model with conditional mean equal to X_n. A population of size $M = 5{,}000$ is assumed; initial conditions and parameters are specified as in Section 4.4. We use a Gillespie τ-leap algorithm ($\tau = 1/100$).

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 (Bretó 2018). Examples of factors that contribute to environmental stochasticity include external unpredictable events (Keeling & Rohani 2008) (e.g., temperature, rainfall, or humidity) and individual variation (e.g., variability in contacts between infectives and susceptibles, superspreaders). These factors translate into fluctuations in the transmission parameter, a concept that can be related to overdispersion of the Poisson distribution (Bretó 2018, Bretó et al. 2009). As in generalized linear models, failure to account for overdispersion may lead to severe underestimation of standard errors (Bretó et al. 2009). Environmental stochasticity is typically modeled by randomizing the transmission rate with white noise. One way to incorporate it in the SIR model is to multiply the transmission rate

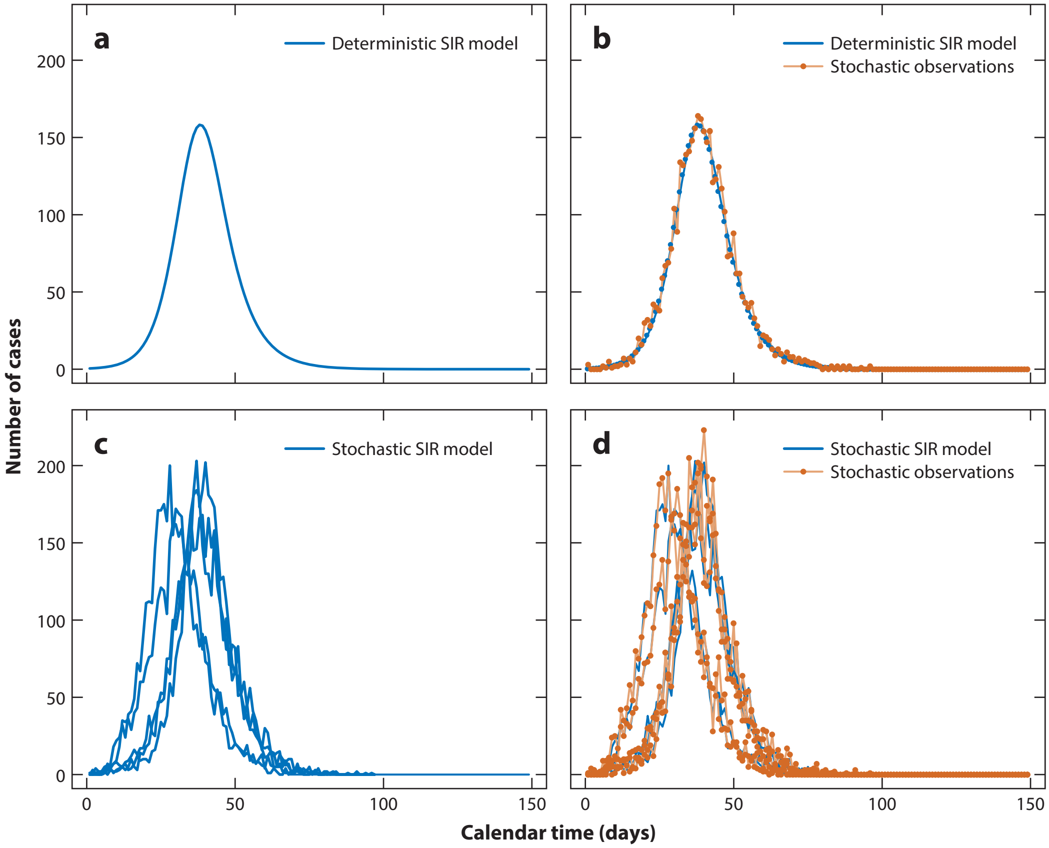

Figure 3

(*a*) Deterministic state process. (*b*) Observations simulated at discrete time points (*orange dots*) via a Poisson measurement model with mean given by values of the deterministic state process at the corresponding time points. (*c*) Five realizations of the stochastic state process. (*d*) For each realization of the stochastic state process, observations simulated at discrete time points (*orange dots*) via a Poisson measurement model with mean given by values of the stochastic state process at the corresponding time points. Abbreviation: SIR, susceptible-infected-removed.

by a Lévy white noise process ξ_t, which fluctuates around one. In Bretó et al. (2009), the process ξ_t is chosen to be $\frac{\mathrm{d}\Gamma(t)}{\mathrm{d}t}$ where marginally, $\Gamma(t+h) - \Gamma(t) \sim \mathrm{Gamma}(h/\sigma^2, \sigma^2)$. 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 that accounts for overdispersion. Failure to do so, or failure to include 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., Nadeem et al. 2016). As such, accounting for overdispersion associated with the measurement process can facilitate distinguishing observational stochasticity from all other sources of variability (Fujiwara & Takada 2001, Nadeem et al. 2016, Bretó 2018). 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 distribution, but other choices are possible, e.g., the Poisson-lognormal model or the 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, including consistency and asymptotic efficiency (see, e.g., Mood et al. 1950). 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 King et al. (2018), 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:

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

where $x_{0:N} = (x_0, x_1, \ldots, x_N)$, $y_{1:N} = (y_1, y_2, \ldots, y_N)$, Θ is the parameter vector, $f(x_n | x_{n-1}; \Theta)$ is the transition density, $f(y_n | x_n; \Theta)$ is the measurement density, and $f(x_0; \Theta)$ 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 (Bretó et al. 2009, King et al. 2018). Assuming that the initial values of the model are given, $\{X_t\}$ is nonrandom and the likelihood function (Equation 11) is given by

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

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 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:

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

The likelihood function (Equation 12) is a high-dimensional integral that 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 MCMC is popular for parameter estimation in situations of an intractable likelihood (Van Dyk & Meng 2001). The rationale behind data augmentation is to make the likelihood tractable through introducing parameters that represent the unobserved data; the joint posterior distribution of parameters and augmented data is then explored via MCMC sampling (see, e.g., Gibson & Renshaw 1998, 2001; O'Neill & Roberts 1999; Lekone & Finkenstädt 2006). 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 that 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 are both methodologically and computationally challenging (Fasiolo et al. 2016, McKinley et al. 2018). Considering a discrete state process whose time step is equal to the reporting interval simplifies the problem at the expense of accuracy (Lekone & Finkenstädt 2006, Li et al. 2018).

Statistical methods for POMP models fall under two main classes: 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. In contrast, information reduction methods perform inference without having to calculate the likelihood of the observed data—one example is ABC.

5.3.3. Sequential Monte Carlo.

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

$$\begin{aligned}\mathcal{L}(\Theta) &= \int_{x_{0:N}} \prod_{n=1}^{N} f\left(y_n|x_n; \Theta\right) f\left(x_0; \Theta\right) f\left(x_n|x_{n-1}; \Theta\right) \mathrm{d}x_{0:N} \\ &= \int_{x_{0:N}} \prod_{n=1}^{N} f\left(y_n|x_n; \Theta\right) f\left(x_{0:N}; \theta\right) \mathrm{d}x_{0:N} \\ &= \mathbb{E}\left[\prod_{n=1}^{N} f\left(y_n|X_n; \Theta\right)\right], \end{aligned} \qquad 13.$$

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

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

where $\{X_{0:N}^j, j = 1, 2, \ldots, J\}$ is a Monte Carlo sample of size J drawn from $f(x_{0:N}; \theta)$. 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 taking the average (King et al. 2018). 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 toward values which are consistent with observed measurements. The likelihood function (Equation 12) can be rewritten as:

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

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

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

where X_n^j is drawn from $f(x_n|y_{1:n-1}; \Theta)$. On the basis of the Monte Carlo likelihood estimate, standard optimization methods (e.g., the 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 (Eberhard et al. 1999). Stochastic optimization methods, e.g., iterated filtering (Ionides et al. 2006), 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 $\mathrm{Var}(\Theta_t|\Theta_{t-1}) = \phi^2$. As the sequential iterations progress, the intensity of the random walk is successively reduced (limit $\phi \rightarrow 0$), and the algorithm converges toward the maximum likelihood estimate; more details are provided by Ionides et al. (2006). Iterated filtering has been successfully tested on a variety of complex epidemiological models, some of which are computationally intractable for available Bayesian methods (Fasiolo et al. 2016). This method is less computationally intensive than its comparable alternatives and has the potential to yield more precise results (Fasiolo et al. 2016, Bretó 2018). The method can be easily implemented in the R package `pomp` (King et al. 2016).

5.3.4. Approximate Bayesian computation.

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

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

where $\mathcal{P}(\Theta)$ is the prior distribution of the model parameters. The procedure proceeds as follows (see, e.g., Csilléry et al. 2010, King et al. 2016, Kypraios et al. 2017, McKinley et al. 2018, Beaumont 2019). 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 data $y_{1:N}^j$ 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 $\mathcal{P}(\Theta|y_n)$. The choice of summary statistics and distance measures is a subject of active research (for the main issues and challenges, see Csilléry et al. 2010, King et al. 2016, Kypraios et al. 2017, McKinley et al. 2018, Beaumont 2019). One ABC approach is implemented in the R package `pomp` (King et al. 2016).

ABC and sequential Monte Carlo are examples of estimation approaches from a range of approaches that have been developed to perform parameter inference when the likelihood is intractable. Other examples include particle MCMC, synthetic likelihood, nonlinear forecasting, and trajectory matching. These methods are described and their applications are demonstrated by King et al. (2016), Kypraios et al. (2017), and McKinley et al. (2018); their weakness and strengths are studied by Fasiolo et al. (2016) and King et al. (2016).

6. REAL DATA EXAMPLE: 1918 INFLUENZA PANDEMIC IN SAN FRANCISCO

The city of San Francisco (in Northern California, United States) 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 (Chowell et al. 2007). We use this data set to compare estimates obtained from different transmission models incorporating varying levels of stochasticity. Models considered are (*a*) a deterministic SIR model, (*b*) a stochastic SIR model, and (*c*) a 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 data set. 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 that 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 (Chowell et al. 2007). Where applicable, we also estimate the dispersion parameter of the negative binomial 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 misspecification in the transmission model (Stocks et al. 2020). The AIC is calculated as

$$\mathrm{AIC} = -2 \log L(\Theta) + 2p, \qquad 19.$$

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 (Claeskens & Hjort 2008). 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 data set. In terms of goodness of fit, for all three transmission models, a negative binomial measurement model provides a better fit than a Poisson measurement model. All the three models assuming a negative binomial 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 six models.

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 model incorporating environmental stochasticity), which is expected when the data suggest the presence of overdispersion ($\theta \neq 0$ and $\sigma \neq 0$) (see, e.g., Ganyani et al. 2020). This indicates that deterministic models have the potential pitfall of underestimating uncertainty associated with parameter estimates and, consequently, the key epidemiological parameters calculated using them. As such,

Table 3 Parameter estimates, 95% confidence intervals, AIC values, and computation time

Transmission model	Estimate (95% confidence interval)				AIC	Time (s)
	β	R_0	θ	σ		
Deterministic SIR	0.494 (0.492, 0.496)	2.026 (2.019, 2.034)	NA	NA	442.380	250
	0.495 (0.491, 0.500)	2.029 (2.011, 2.048)	0.034 (0.011, 0.063)	NA	313.147	259
Stochastic SIR	0.477 (0.422, 0.501)	1.955 (1.730, 2.053)	NA	NA	337.622	2,716
	0.494 (0.451, 0.519)	2.024 (1.849, 2.128)	0.022 (0.002, 0.064)	NA	317.611	2,840
Stochastic SIR*	0.461 (0.399, 0.499)	1.889 (1.638, 2.045)	NA	0.162 (0.015, 0.241)	393.726	6,107
	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	7,234

The models were fitted on a 3.1 GHz PC with four cores. For all three transmission models, a Poisson NB measurement model was used. θ denotes the dispersion parameter of the 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. Abbreviations: AIC, Akaike information criterion; NA, not applicable; NB, negative binomial; PC, personal computer; SIR, susceptible-infected-removed.

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 that account for environmental stochasticity can, to some extent, compensate for model misspecification resulting in even greater uncertainty (King et al. 2015).

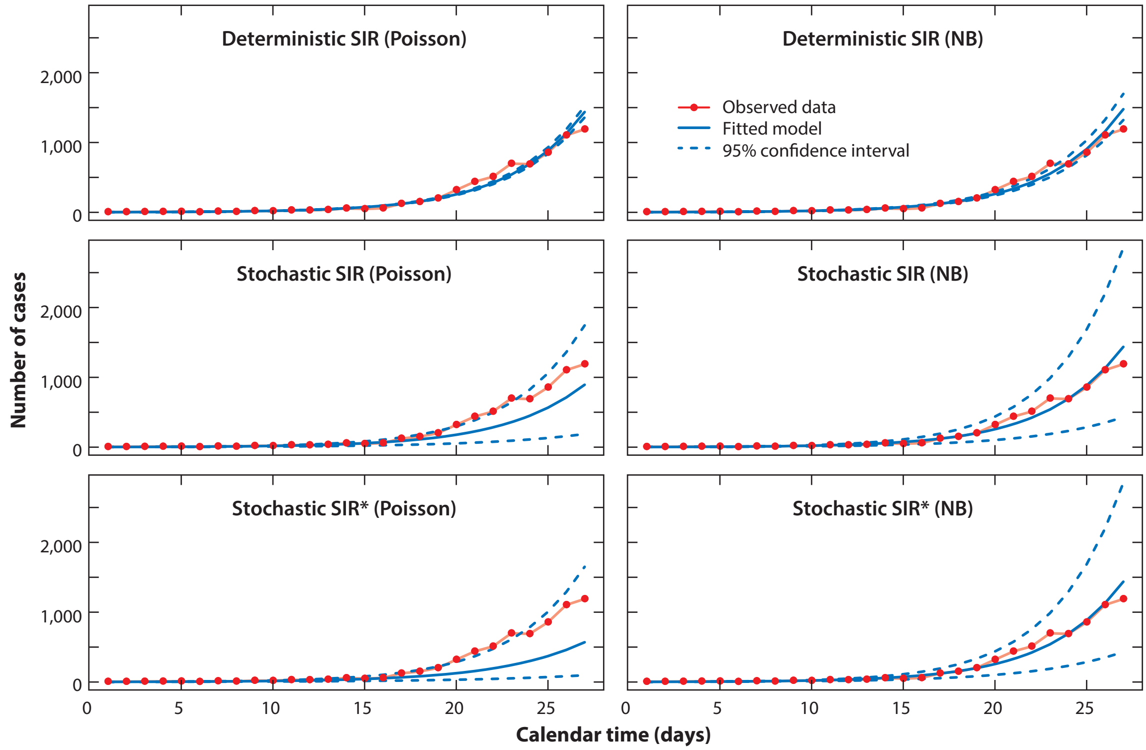

Figure 4

Comparison of observed data (*red*) and fitted models (*blue*). Abbreviations: NB, negative binomial; SIR, susceptible-infected-removed; stochastic SIR*, stochastic SIR model incorporating environmental stochasticity.

7. DISCUSSION AND CONCLUSIONS

The goal of this article is 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 (King et al. 2016). Moreover, simulated data can be used for testing the performance of estimation methods as well as investigating the estimability of parameters from available data (see, e.g., Ganyani et al. 2018, 2020). In terms of estimation, we show that, compared with deterministic models, stochastic models offer an opportunity to better quantify uncertainty of parameter estimates. We also show that, when data suggest the presence of overdispersion, incorporating overdispersion in the transmission model or the measurement model can improve model fit and yield less optimistic parameter estimates (see, e.g., King et al. 2015, Ganyani et al. 2020, Stocks et al. 2020).

The SIR model used throughout this article assumes a well-mixed population. This assumption implies that an infective is equally likely to infect any 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) (Newman 2002). 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., Brauer & Castillo-Chávez 2001, Newman 2002). Though network-based compartmental models can be formulated, simulation and analysis of the resulting models depend 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., Keeling & Eames 2005). Instead of assuming that the population being modeled 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 behavior 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 such as age, gender, health status, or behavior. Autonomy means that individuals act independently of each other and pursue their own objectives. Epidemiological ABMs mainly consist of four components: disease, society, transportation, and the environment. All four components are to be modeled when formulating an ABM (typically jointly). Modeling the disease involves describing how the infectious disease is transmitted between individuals and how the disease progresses in an infected agent. Modeling the society involves simulating the population. Modeling transportation concerns how the individuals will move within the environment. Modeling 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., Railsback & Grimm 2011, Hunter et al. 2017).

On fitting models to data, we sidestepped 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 that these models seek to address may not be addressable because parameter estimates will be unreliable (Raue et al. 2009). As such, a fundamental prerequisite for parameter estimation is investigation of structural and practical identifiability. The former concerns studying which parameters are functionally related in such a way that they cannot be uniquely determined under noise-free conditions. 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., Held et al. 2019). On the one hand, 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 Raue et al. 2009, Chowell 2017, Tuncer & Le 2018, and references therein).

In closing, stochastic modeling of infectious diseases, an area which encompasses simulation and analysis, is a subject of enormous research. An important caveat of this article is that it is by no means exhaustive; nevertheless, we attempt to provide a useful background on the essentials of simulation and analysis of stochastic compartmental models.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

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.

LITERATURE CITED

Allen LJS. 2017. A primer on stochastic epidemic models: formulation, numerical simulation, and analysis. *Infect. Dis. Model.* 2:128–42

Anderson RM, May RM. 1992. *Infectious Diseases of Humans: Dynamics and Control*. Oxford, UK: Oxford Univ. Press

Andersson H, Britton T. 2000. *Stochastic Epidemic Models and Their Statistical Analysis*. New York: Springer

Bailey NTJ. 1955. Some problems in the statistical analysis of epidemic data. *J. R. Stat. Soc. Ser. B* 17:35–58

Bartlett MS. 1957. Measles periodicity and community size. *J. R. Stat. Soc. Ser. A* 120:48–70

Bartlett MS. 1960. *Stochastic Population Models in Ecology and Epidemiology*. London: Methuen

Bartlett MS. 1961. Monte Carlo studies in ecology and epidemiology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, ed. J Neyman, pp. 39–55. Berkeley: Univ. Calif. Press

Beaumont MA. 2019. Approximate Bayesian computation. *Annu. Rev. Stat. Appl.* 6:379–403

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

Bretó C. 2018. Modeling and inference for infectious disease dynamics: a likelihood-based approach. *Stat. Sci. Rev. J. Inst. Math. Stat.* 33:57–69

Bretó C, He D, Ionides EL, King AA. 2009. Time series analysis via mechanistic models. *Ann. Appl. Stat.* 3:319–48

Chao DL, Halloran ME, Obenchain VJ, Longini IM Jr. 2010. FluTE, a publicly available stochastic influenza epidemic simulation model. *PLOS Comput. Biol.* 6(1):e1000656

Chowell G. 2017. Fitting dynamic models to epidemic outbreaks with quantified uncertainty: a primer for parameter uncertainty, identifiability, and forecasts. *Infect. Dis. Model.* 2:379–98

Chowell G, Hyman JM, Bettencourt LMA, Castillo-Chavez C, eds. 2009. *Mathematical and Statistical Estimation Approaches in Epidemiology*. New York: Springer

Chowell G, Nishiura H, Bettencourt LM. 2007. Comparative estimation of the reproduction number for pandemic influenza from daily case notification data. *J. R. Soc. Interface* 4:155–66

Claeskens G, Hjort NL. 2008. *Model Selection and Model Averaging*. Cambridge, UK: Cambridge Univ. Press

Coburn BJ, Wagner BG, Blower S. 2009. Modeling influenza epidemics and pandemics: insights into the future of swine flu (H1N1). *BMC Med.* 7:30

Coulson T, Rohani P, Pascual M. 2004. Skeletons, noise and population growth: the end of an old debate? *Trends Ecol. Evol.* 19:359–64

Csilléry K, Blum MG, Gaggiotti OE, François O. 2010. Approximate Bayesian computation (ABC) in practice. *Trends Ecol. Evol.* 25:410–18

Daley DJ, Gani J. 2001. *Epidemic Modelling: An Introduction*. Cambridge, UK: Cambridge Univ. Press

Diekmann O, Heesterbeek JAP, Metz JAJ. 1990. On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. *J. Math. Biol.* 28:365–82

Dong E, Du H, Gardner L. 2020. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* 20:533–34

Eberhard P, Schiehlen W, Bestle D. 1999. Some advantages of stochastic methods in multicriteria optimization of multibody systems. *Arch. Appl. Mech. Ing. Arch.* 69:543–54

Fasiolo M, Pya N, Wood SN. 2016. A comparison of inferential methods for highly nonlinear state space models in ecology and epidemiology. *Stat. Sci.* 31:96–118

Fuchs C. 2013. *Inference for Diffusion Processes: With Applications in Life Sciences*. New York: Springer

Fujiwara M, Takada T. 2001. Environmental stochasticity. *eLS*. **https://doi.org/10.1002/9780470015902.a0021220.pub2**

Ganyani T, Faes C, Chowell G, Hens N. 2018. Assessing inference of the basic reproduction number in an SIR model incorporating a growth-scaling parameter. *Stat. Med.* 37:4490–506

Ganyani T, Faes C, Hens N. 2020. Inference of the generalized-growth model via maximum likelihood estimation: a reflection on the impact of overdispersion. *J. Theor. Biol.* 484:110029

Gibson GJ, Renshaw E. 1998. Estimating parameters in stochastic compartmental models using Markov chain methods. *Math. Med. Biol.* 15:19–40

Gibson GJ, Renshaw E. 2001. Likelihood estimation for stochastic compartmental models using Markov chain methods. *Stat. Comput.* 11:347–58

Gillespie DT. 1977. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* 81:2340–61

Gillespie DT. 2001. Approximate accelerated stochastic simulation of chemically reacting systems. *J. Chem. Phys.* 115:1716–33

Halloran ME, Auranen K, Baird S, Basta NE, Bellan SE, et al. 2017. Simulations for designing and interpreting intervention trials in infectious diseases. *BMC Med.* 15:223

Heesterbeek H. 2005. The law of mass-action in epidemiology: a historical perspective. In *Ecological Paradigms Lost: Routes of Theory Change*, ed. K Cuddington, B Beisner, pp. 81–104. Amsterdam: Elsevier

Held L, Hens N, O'Neill PD, Wallinga J. 2019. *Handbook of Infectious Disease Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC

Hens N, Shkedy Z, Aerts M, Faes C, Damme PV, Beutels P. 2012. *Modeling Infectious Disease Parameters Based on Serological and Social Contact Data*. New York: Springer

Hollingsworth TD. 2009. Controlling infectious disease outbreaks: lessons from mathematical modelling. *J. Public Health Policy* 30:328–41

Hunter E, Mac Namee B, Kelleher JD. 2017. A taxonomy for agent-based models in human infectious disease epidemiology. *J. Artif. Soc. Soc. Simul.* 20(3):2

Ionides EL, Bretó C, King AA. 2006. Inference for nonlinear dynamical systems. *PNAS* 103:18438–43

Kaminsky J, Keegan LT, Metcalf CJE, Lessler J. 2019. Perfect counterfactuals for epidemic simulations. *Philos. Trans. R. Soc. B* 374:20180279

Keeling MJ, Eames KT. 2005. Networks and epidemic models. *J. R. Soc. Interface* 2:295–307

Keeling MJ, Rohani P. 2008. *Modeling Infectious Diseases in Humans and Animals*. Princeton, NJ: Princeton Univ. Press

Keeling MJ, Ross JV. 2009. Efficient methods for studying stochastic disease and population dynamics. *Theor. Popul. Biol.* 75:133–41

Kermack WO, McKendrick AG. 1927. Contributions to the mathematical theory of epidemics. Part I. *Proc. R. Soc. Ser. A* 115:700–21

King AA, de Cellès MD, Magpantay FMG, Rohani P. 2015. Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. *Proc. R. Soc. Ser. B* 282:20150347

King AA, Ionides EL, Asfaw K. 2018. *Simulation-based inference for epidemiological dynamics*. Presented at Summer Institute in Statistics and Modeling in Infectious Diseases. **https://kingaa.github.io/sbied/**

King AA, Nguyen D, Ionides EL. 2016. Statistical inference for partially observed Markov processes via the R package pomp. *J. Stat. Softw.* 69:12

Kypraios T, Minin VN. 2018. Introduction to the Special Section on Inference for Infectious Disease Dynamics. *Stat. Sci.* 33:1–3

Kypraios T, Neal P, Prangle D. 2017. A tutorial introduction to Bayesian inference for stochastic epidemic models using approximate Bayesian computation. *Math. Biosci.* 287:42–53

Lekone PE, Finkenstädt BF. 2006. Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. *Biometrics* 62:1170–77

Li M, Dushoff J, Bolker BM. 2018. Fitting mechanistic epidemic models to data: a comparison of simple Markov chain Monte Carlo approaches. *Stat. Methods Med. Res.* 27:1956–67

Mandal S, Sarkar R, Sinha S. 2011. Mathematical models of malaria—a review. *Malaria J.* 10:202

McCallum H. 2001. How should pathogen transmission be modelled? *Trends Ecol. Evol.* 16:295–300

McKinley TJ, Vernon I, Andrianakis I, McCreesh N, Oakley JE, et al. 2018. Approximate Bayesian computation and simulation-based inference for complex stochastic epidemic models. *Stat. Sci.* 33:4–18

Mood AM, Graybill FA, Boes DC. 1950. *Introduction to the Theory of Statistics*. New York: McGraw–Hill

Nadeem K, Moore JE, Zhang Y, Chipman H. 2016. Integrating population dynamics models and distance sampling data: a spatial hierarchical state-space approach. *Ecology* 97:1735–45

Newman ME. 2002. Spread of epidemic disease on networks. *Phys. Rev. E* 66:016128

O'Neill PD. 2010. Introduction and snapshot review: relating infectious disease transmission models to data. *Stat. Med.* 29:2069–77

O'Neill PD, Roberts GO. 1999. Bayesian inference for partially observed stochastic epidemics. *J. R. Stat. Soc. Ser. A* 162:121–29

Ozcaglar C, Shabbeer A, Vandenberg SL, Yener B, Bennett KP. 2012. Epidemiological models of *Mycobacterium tuberculosis* complex infections. *Math. Biosci.* 236:77–96

Pineda-Krch M. 2008. GillespieSSA: implementing the stochastic simulation algorithm in R. *J. Stat. Softw.* 25:12

Railsback SF, Grimm V. 2011. *Agent-Based and Individual-Based Modeling: A Practical Introduction*. Princeton, NJ: Princeton Univ. Press

Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, et al. 2009. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* 25:1923–29

Ross SM. 2014. *Introduction to Probability Models*. New York: Academic

Stocks T, Britton T, Höhle M. 2020. Model selection and parameter estimation for dynamic epidemic models via iterated filtering: application to rotavirus in Germany. *Biostatistics* 23(3):400–16

Taubenberger JK, Morens DM. 2006. 1918 influenza: the mother of all pandemics. *Rev. Biomed.* 17:69–79

Tuncer N, Le TT. 2018. Structural and practical identifiability analysis of outbreak models. *Math. Biosci.* 299:1–18

Van Dyk DA, Meng XL. 2001. The art of data augmentation. *J. Comput. Graph. Stat.* 10:1–50

Wilkinson DJ. 2018. *Stochastic Modelling for Systems Biology*. Boca Raton, FL: Chapman and Hall/CRC

Annual Review of Statistics and Its Application

Volume 8, 2021

Contents

Errata

An online log of corrections to *Annual Review of Statistics and Its Application* articles may be found at http://www.annualreviews.org/errata/statistics