Cecilia Viscardi, Alessio Lachi* and Michela Baccini

Discrete-time compartmental models with partially observed data: a comparison among frequentist and Bayesian approaches for addressing likelihood intractability

https://doi.org/10.1515/em-2024-0032
Received December 30, 2024; accepted August 6, 2025; published online September 26, 2025

Abstract: Compartmental models have emerged as useful tools in epidemiology due to their mechanistic nature. They provide insights into complex dynamic systems and allow predictions under different scenarios. However, despite their widespread use, there is still a gap in the literature, concerning their statistical formalization and a systematic discussion of the statistical methods suitable for both tasks of inference and forecasting. In this work, starting from the fundamental distinction between deterministic and stochastic compartmental models, we focus on how the formulation of the likelihood function becomes a necessary and challenging step in the transition from a deterministic to a stochastic framework. We then analyse the various difficulties encountered in evaluating the likelihood function associated with discrete-time stochastic models. We distinguish two reasons for the intractability of the likelihood function, the high dimension of missing data and the complexity of the model structure, and discuss suitable methods for addressing the problem both from a frequentist and Bayesian perspective. We overview likelihood-based methods and explore the use of likelihood-free approaches in this framework, namely approximate Bayesian computation algorithms and a method that combines model calibration with a parametric bootstrap procedure. We emphasize their ability to make inferences from data that are partially observed, or only observed in some aggregated form. To showcase their feasibility and reliability, we compare the likelihood-free and likelihood-based methods at work with a toy example of the Susceptible-Infected-Removed. Finally, we explore the relevance of likelihood-free methods in a real-world framework through an example of a complex compartmental model developed to study smoking dynamics in Tuscany (Italy).

Keywords: approximate Bayesian computation; calibration; compartmental models; incomplete data; intractable likelihood

Introduction

Compartmental models are a class of models used to understand and describe the dynamic evolution of a phenomenon of interest in a population. Due to their simple mechanistic nature, they are widely used for modelling infectious diseases. In particular, in the last few years, compartmental models have been experimented with a vast spreading due to the increasing interest in epidemiological analyses of the SARS-CoV-2 epidemic [1]. However, compartmental models are widely employed also in areas of epidemiology beyond infectious disease

***Corresponding author: Alessio Lachi**, Department of Medicine, Saint Camillus International University of Health and Medical Sciences, Rome, Italy, E-mail: alessio.lachi@unicamillus.org. https://orcid.org/0000-0002-0764-0866
Cecilia Viscardi, Department of Economics and Statistics, University of Salerno, Salerno, Italy, E-mail: cviscardi@unisa.it
Michela Baccini, Department of Statistics, Computer Science, and Applications "Giuseppe Parenti", University of Florence, Florence, Italy, E-mail: michela.baccini@unifi.it

modelling, such as in the study of smoking habits [2, 3], as well as in other fields like engineering, pharmacology, and the study of social phenomena [4–6].

Compartmental models assume that, at each point in time, the population is divided into non-overlapping groups, called compartments, which are homogeneous concerning some specific characteristics of the individuals in the population – e.g. the health status, smoking status etc. Starting from an initial population, the transitions between compartments are allowed and described by a system of ordinary differential equations (ODEs) that define the evolution of the size of each compartment over time. These ODEs are governed by a set of model parameters that tune the transition rates.

Compartmental models are very flexible and can be employed in a forward perspective, to simulate dynamics under different scenarios defined by specific values of transition parameters, or to understand and predict the evolution of a phenomenon by estimating the model parameters based on observed dynamics. Here we will refer to this second case as the backward procedure. Although compartmental models based on deterministic functions, such as ODEs, have proven useful for both purposes, their reliability builds upon "large population" assumption and assumptions about the size of each compartment [7]. When violating these conditions (e.g. at the beginning of a pandemic) and in situations where various unknown/unobservable factors generate uncertainty, it can be beneficial to account for the variability that characterizes complex phenomena. This is particularly relevant when these models are used, not just to explain reality, but to make long-term predictions and in decision-making contexts [8]. A stochastic extension of these models enables more reliable predictions of future scenarios by properly accounting for many sources of uncertainty. This extension is based on the formulation of a stochastic model, beyond the deterministic model, in which each transition of an individual from a given compartment to another follows probabilistic rules. The stochastic counterpart of a deterministic compartmental model requires the formulation of a likelihood function that derives from the introduced probabilistic rules. Unfortunately, this likelihood function is often intractable. In some cases, it can be complex to specify an analytical form for the likelihood, due to the complexity of the structure of dependencies among the involved variables, e.g., due to the large number of compartments or to challenging definitions of the allowed transitions and related probabilistic rules. In other cases, an analytical form of the likelihood is available but its point-wise evaluation is infeasible due to the presence of a large number of unobserved latent variables and missing data. In both cases, the inference on the model parameters requires suitable methods. Often, these methods may be computationally less efficient than those required for a deterministic model, therefore, stochastic models should not always be preferred. Note that a comparison between deterministic and stochastic models is beyond the scope of this work, which primarily deals with stochastic modeling and issues related to their likelihood function (for a comparison, we refer the reader to [8, 9], among others). Despite this, we also present the deterministic models, as they are useful for understanding how the stochastic component can be introduced and how the likelihood function is consequently formalized.

In the literature, existing works that discuss methods for making inferences on compartmental models focus only on very specific approaches. For example, McKinley et al. [10–12] discuss algorithms used to provide Bayesian parameter estimates in various epidemiological models. They mainly address the problem of continuous-time epidemic models and focus on the comparison between a popular data augmentation Markov chain algorithm and methods based on Monte Carlo approximations of the intractable likelihood. Tang et al. [13] provide a comprehensive review of frequentist and Bayesian methods. However, each of them is applied to a deterministic/stochastic model carefully tailored and adapted to fit the requirements of the specific method. Thus, there is still a lack of critical comparisons among the main frequentist and Bayesian methods, when applied to the same model (i.e. under the same probabilistic assumptions), particularly within the discrete-time framework.

This paper addresses this gap by presenting and evaluating different estimation strategies within both frequentist and Bayesian frameworks, focusing specifically on discrete-time compartmental models because epidemiological data are usually recorded and made available on a discrete-time scale (e.g., daily, weekly, monthly). However, when the time-scale is heterogeneous among data sources and across time, this mismatch can lead to problems in the evaluation of the likelihood function. Here, we show that this issue, often addressed by adopting continuous-time models, can also be handled within the discrete-time framework by using suitable methods. To

clarify the modeling process, we describe both the mathematical and statistical models underlying deterministic and stochastic compartmental models, emphasizing that the introduction of stochastic components gives rise to the model's (intractable) likelihood function. We then discuss typical causes of likelihood intractability and outline appropriate methodological solutions. A special focus is given to the reliability and flexibility of likelihood-free methods. In particular, within the frequentist framework, we deal with a method that combines model calibration and a parametric bootstrap procedure for uncertainty quantification, and, within the Bayesian framework, we deal with approximate Bayesian computation algorithms. We emphasize their ability not only to handle complex models for which the likelihood is unavailable, but also to make inferences when data are available only in aggregated form or are characterized by a large number of latent variables and missing values.

The paper is organized as follows: in Section "Discrete-time compartmental models" we provide a general description of compartmental models and the relation between deterministic and stochastic models. In Section "Working example: the SIR model" a focus is given on the Susceptible-Infected-Removed (SIR) model. The SIR model will be used as an illustrative example in the rest of the paper. Section "Frequentist approaches and Bayesian approaches" describe frequentist and Bayesian methods for conducting inference in three different situations: the case in which the likelihood function is tractable and complete data are available; the case in which the likelihood function would be available and tractable but missing or sparsely observed data makes its evaluation infeasible; the case in which the likelihood function is not tractable at all. The results of the SIR example are reported in Section "Results". Finally, as a real example of intractable likelihood, we consider a complex compartmental model designed to describe the evolution over time of smoking dynamics in the population of the Tuscany region (Italy). Section "Discussion and conclusions" discusses and compares the results of all the methods considered, providing final remarks and conclusions.

Discrete-time compartmental models

Compartmental models describe the evolution of a phenomenon in a population over time. At each point in time, the population is divided into compartments – i.e. groups of individuals homogeneous concerning some characteristics, such as health status. Starting from an initial condition, individuals can change their status and transit from a given compartment to another one. It follows that the size of each compartment changes over time. Compartmental models formalize a dynamic system relying on a system of ordinary differential equations (ODEs). This simple mathematical model describes the trajectory of the size of each compartment over time as a function of a set of parameters that govern the transition rates. The system of ODEs, for computational reasons, is often transformed into a system of difference equations defined in discrete time [14].

From deterministic to stochastic compartmental models

Deterministic models describe reality through mathematical functions, however realistic modelling often requires a stochastic extension. Stochastic models enable us to account for sampling variability and to quantify uncertainty both in the estimation and prediction phases. They integrate a systematic component – i.e. a mathematical function – and a stochastic component. This latter is introduced by establishing that the number of transitions that occur between two compartments are realizations of random variables that follow specific probability distributions.

Let us denote by $X(t) = \big(X_1(t), \dots, X_c(t), \dots, X_C(t)\big)$ the state of a system made of C compartments, where $X_c(t)$ denotes the size of the c-th compartment at time t. A deterministic model defines a function $f(\cdot;\, \theta)$ that expresses the change of the size of each compartment as a function of a set of parameters θ. In particular, the mathematical function can be a system of ODEs in continuous time: $\frac{d}{\mathrm{d}t}X(t) = f(X(t); \theta)$.

Given a vector of parameters, θ, that describes a specific scenario, the solution of the system of ODEs, intended as the dynamic that satisfies the ODEs, provides the evolution of the compartment sizes in a forward perspective. The system of ODEs is often difficult to solve analytically and involves a lot of calculations. A

practical solution is given by Euler's method [14]. This method considers a system of equations defined in discrete time, meaning that t assumes values in a subset of $\mathbb{I}$, the set of integer numbers. When considering discrete time, the deterministic model specifies the size of each compartment at time t as a function of the sizes at the previous point in time: $X(t) = X(t-\delta) + \Delta X(t)$, where $\Delta X(t) = f(X(t-\delta);\ \theta)$ denotes the variation caused by the transitions between compartments that occur in the time interval $[t-\delta, t)$. δ is an integer number that represents a time increment and is often assumed to be equal to 1. The smaller δ the better approximation from the continuous to the discrete process.

Deterministic models can also be used in a backward perspective, to understand the dynamic underlying the observed phenomenon by learning the transition parameters. In such a case, optimization strategies are implemented to find the optimal value θ^*, i.e. the value that minimizes a distance function between the observed data and the trajectories described by the solution of the system of equations with $\theta = \theta^*$.

Both from a forward and a backward perspective, deterministic models do not account for the occurrence of any deviation from the solution of the system. They are not intended to capture the randomness of the phenomena, although they are inherently stochastic at an individual level, meaning that each individual can experience (or not) a transition with a certain probability at any given time. Kurtz [15, 16] proved that the solution of a well-defined deterministic compartmental model is the infinite population limit of a stochastic system. However, when the size of the population is finite, variability must be modeled and evaluated to come up with conscious analyses and avoid misleading conclusions. Stochastic models are the proper tools to combine the mathematical function with a stochastic component. They define a stochastic function that expresses the evolution of the size of each compartment as a function of the parameters θ and a random noise. Again, from a forward perspective, the stochastic model can be employed to simulate the dynamic corresponding to a scenario described by the vector of parameters θ. However, the model will simulate different dynamics even considering the same vector θ, since they are the output of a stochastic generative process that involves a random noise. The probability of observing a specific dynamic $\mathbf{x} = \{x(0), \dots, x(T)\}$ depends on how the random components are integrated into the model. The likelihood function, $L(\theta \mid \mathbf{x})$, comes from the probability mass/density functions of $\mathbf{X} = \{X(0), \dots, X(T)\}$ evaluated at $\mathbf{x}$ and viewed as a function of the parameters. It quantifies how likely the scenario described by θ is, in the light of the observed data. The likelihood function plays a key role in the inference process, but in stochastic compartmental models it often results in being intractable.

Let us consider a simple way to come up with an easy-to-handle stochastic model: the introduction of additive random noises. Denoting by $\hat{\mathbf{x}}(\theta) = \{\hat{x}(0;\theta), \dots, \hat{x}(T;\theta)\}$ the solution of the system of equations evaluated at $t \in \{0, \dots, T\}$, we have that $X(t) = \hat{x}(t;\theta) + E(t)$ is a random variable the randomness of which comes from $E(t)$, the vector of random perturbation noises at time t. Observed data are $x(t) = \hat{x}(t;\theta) + \varepsilon(t)$, where $\varepsilon(t)$ denotes a realization of the random vector $E(t)$, for each t. When assuming Gaussian errors $E(t) \sim MVN(\boldsymbol{\mu} = \mathbf{0}, \Sigma)$ – where $\boldsymbol{\mu}$ and Σ are a vector and a matrix of sizes C and $C \times C$, respectively – random variables $X(t)$ have MVN distributions with mean $\hat{x}(t;\theta)$ and variance Σ. Under the strong assumption of independence among the errors over time, the likelihood function is tractable and can be expressed as follows:

$$L\big(\theta \mid \mathbf{x}\big) = \prod_{t=0}^{T} \frac{1}{(2\pi)^{C/2}|\Sigma|^{1/2}} \exp\Big(-\frac{1}{2}(x(t) - \hat{x}(t;\theta))'\Sigma^{-1}(x(t) - \hat{x}(t;\theta))\Big). \tag{1}$$

Even in this simple case, the point-wise evaluation of the likelihood is based on the computation of $\hat{\mathbf{x}}(\theta)$ that may require a numerical solution of the system of equations. This often makes the inferential process computationally demanding, because both frequentist and Bayesian methods require several evaluations of the likelihood function, e.g. during the implementation of a maximization procedure or Monte Carlo algorithms. This problem becomes even more serious depending on the way the random component is formulated – e.g. random noises that depend in turn on the parameter vector θ and/or cannot be assumed independent over time. For example, this happens when the waiting time (continuous or discrete) between events – i.e. transitions between compartments – is assumed to follow an exponential or geometric distribution [17] and the number of transitions at each point in time are modeled as dependent Poisson or Binomial random variables. In such cases, being based on more complex dependence structures, the joint probability/mass function of $\mathbf{X}$ cannot be factorized as

in (1) and is more often associated with an intractable likelihood function. These situations will be detailed in the next section resorting to a working example which clarifies also why evaluating this kind of likelihood can be, not only computationally demanding, but truly infeasible.

Working example: the SIR model

The definition and the evaluation of the likelihood function are of paramount importance in the inferential process, both from a Bayesian and a frequentist point of view. Unfortunately, in many compartmental models, a point-wise evaluation of the likelihood function is prohibitive. This may happen at least in two cases: I) when the analytical form is available but its evaluation is computationally demanding; II) when the model is so complex that it is infeasible to write down an analytical form for the associated likelihood function. We will refer to the second circumstance as *likelihood unavailability*. In what follows we will describe the formulation of the likelihood function and the reasons that make its evaluation infeasible using a simple working example. An example addressing a much more complex model is deferred to Section "A real-world example: the SHC model".

Let us consider the well-known Susceptible-Infected-Removed (SIR) model [18–20]. The SIR model subdivides the population into three compartments: susceptible individuals (S) are those who can potentially contract the disease when they come into contact with an infectious individual since they are not immunized; infectious individuals (I) are those who are currently infected and infective, thus they can transmit the disease to susceptible individuals; removed individuals (R) have been infected and have either recovered from the disease or have died. In this model, once an individual recovers, she/he is assumed to have immunity to the disease.

Figure 1 shows the transitions allowed by the model and the parameters governing the transition rates. We denote by γ the resolution rate defined as $1/\tau$, where τ is the average time spent by an infected individual in the compartment I. The instantaneous rate of transmission of the infection is denoted by β. From β and γ, we can compute the basic reproduction number – i.e. the expected number of secondary infections caused by a single infected individual at the beginning of the epidemic – as $R_0 = \beta/\gamma$. Here, we denote by $\theta = (\tau, R_0)$ the vector of parameters to be inferred, and by $S(t)$, $I(t)$ and $R(t)$ the sizes of the three compartments at time t.

The following system of differential equations describes the dynamic of the SIR model:

$$\begin{cases} \dfrac{dS(t)}{\mathrm{d}t} = -\beta \dfrac{I(t)}{S(0)} S(t) \\ \dfrac{dI(t)}{\mathrm{d}t} = \beta \dfrac{I(t)}{S(0)} S(t) - \gamma I(t) \\ \dfrac{\mathrm{d}R(t)}{\mathrm{d}t} = \gamma I(t). \end{cases} \tag{2}$$

Note that, in Equation (2), $\beta \frac{I(t)}{S(0)}$ is the infection rate at time t. It depends on $\frac{I(t)}{S(0)}$, which is the fraction of infectious individuals with whom a susceptible individual can come into contact at time t, and on β, which represents the instantaneous rate at which an infectious individual infects a susceptible one. The system of ODEs in Equation (2), may be replaced by the following system of equations in discrete time:

$$\begin{cases} S(t) = S(t-1) - \pi_{\mathrm{SI}}(t-1) S(t-1) \\ I(t) = I(t-1) + \pi_{\mathrm{SI}}(t-1) S(t-1) - \pi_{\mathrm{IR}} I(t-1) \\ R(t) = R(t-1) + \pi_{\mathrm{IR}} I(t-1), \end{cases} \tag{3}$$

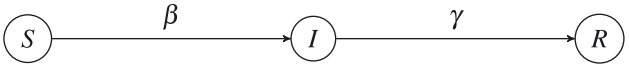

Figure 1: Graphical representation of the SIR model. Each node represents a compartment and edges indicate the allowed transitions between compartments.

where $\pi_{SI}(t-1) = 1 - \exp\left(-\beta \frac{I(t-1)}{S(0)}\right)$ and $\pi_{IR} = 1 - \exp(-\gamma)$ are the probability of being infected and the probability of recovering or dying during a unit time interval, respectively. These probabilities come from the assumption that the waiting time before experimenting with an event (infection or recovery/death) has an exponential distribution. In particular, π_{SI} and π_{IR} are the probabilities of waiting a time smaller than 1, i.e. the probability of exiting the compartment between $t-1$ and t.

A possible way to introduce stochasticity in the SIR model is assuming two independent Binomial distributions for the random variables that count new infections and resolutions, say $i(t)$ and $r(t)$: $i(t) \sim Binom(S(t), \pi_{SI}(t))$ and $r(t) \sim Binom(R(t), \pi_{IR})$ for each $t \in \{0, \dots, T\}$ (Figure 2). Note that, as shown by Allen et al. [21], the solution of the deterministic model can be seen as the expected value of the stochastic one. This is apparent looking at Equation (3), where the number of new infections and recoveries, at each time t, resemble expected values of two Binomial distributions.

Let us denote by $i_{t_1:t_2} = \big(i(t_1), \dots, i(t_2)\big)$ and $r_{t_1:t_2} = \big(r(t_1), \dots, r(t_2)\big)$ the random vectors of new infections and new resolutions from t_1 until t_2 (t_1 and $t_2 \in \{0, \dots T\}$, $t_2 \geq t_1$), and by $i^*_{t_1:t_2}$ and $r^*_{t_1:t_2}$ their realization. Under the binomial assumption, the likelihood function is:

$$\begin{aligned} L\big(\theta \mid i_{1:T} = i^*_{1:T}, r_{1:T} = r^*_{1:T}\big) &= \prod_{t=1}^{T} \Pr\big(i(t) = i^*(t) \mid i^*_{0:t-1}, r^*_{0:t-1}\big) \Pr\big(r(t) = r^*(t) \mid i^*_{0:t-1}, r^*_{0:t-1}\big) \\ &= \prod_{t=1}^{T} \binom{S(t)}{i^*(t)} \pi_{SI}(t)^{i^*(t)} [1 - \pi_{SI}(t)]^{S(t) - i^*(t)} \binom{I(t)}{r^*(t)} \pi_{IR}^{r^*(t)} [1 - \pi_{IR}]^{I(t) - r^*(t)}, \end{aligned} \tag{4}$$

where $S(t) = S(t-1) - i^*(t-1)$ and $I(t) = I(t-1) + i^*(t-1) - r^*(t-1)$. The initial condition of the system is assumed to be: $S(0) = \mathrm{n} - 1$, with n equal to the size of the population; $I(0) = 1$; $R(0) = 0$. This likelihood is analytically tractable, but different distributional assumptions about the process may complicate its form.

The model described so far is based on several strong assumptions:

- the population is closed to births and deaths (except those due to the studied infectious disease), to immigration and emigration, thus $S(t) + I(t) + R(t) = \mathrm{n}$ for each $t \in \{0, \dots, T\}$;
- the population is homogeneously mixed;

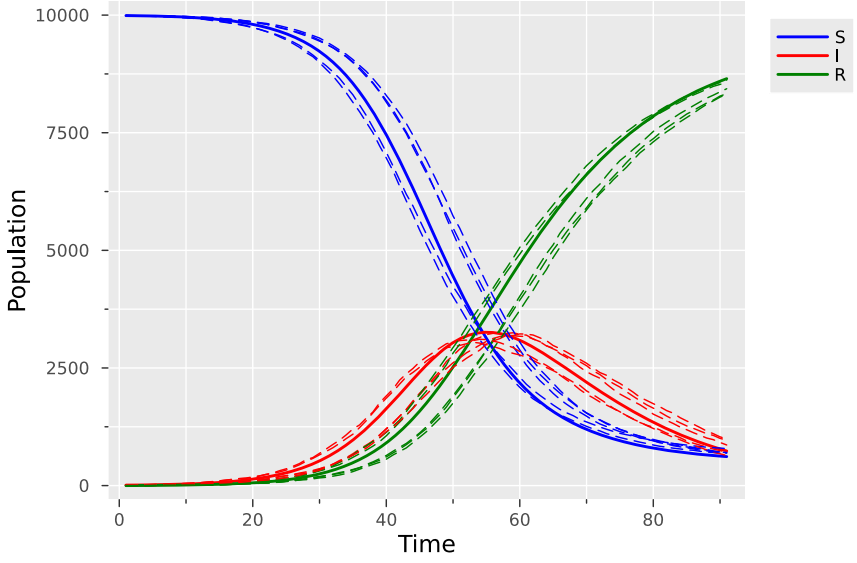

Figure 2: Evolution of compartment sizes in a deterministic SIR model (continuous line) and in four realizations from a stochastic SIR model (dashed line) for $T = 90$ days, setting $\theta = (\tau = 14, R_0 = 3)$, $S(0) = 9,990$, $I(0) = 10$, and $R(0) = 0$.

- all the individuals in the same compartment at the same time t have the same risk of leaving the compartment, regardless of the time already spent in it;
- individuals are infective from the onset of the infection until recovery or death;
- the reinfection rate is equal to 0;
- both the instantaneous infection rate and the resolution rate are constant over time.

To relax the above-mentioned assumptions, in the literature have been proposed several strategies, such as the introduction of further compartments to take into account the incubation period [22] or the availability of vaccines [23]. Some of them relax the homogeneity assumption by assuming that the spreading process of infectious disease occurs over a network structure. Some models relax the assumption of constant transition rates by introducing some dependencies from the calendar time, or other variables (e.g. age) [24, 25]. Most of these extensions complicate the likelihood function in Equation (4) making it intractable, or even unavailable.

However, even very simple compartmental models may be associated with an intractable likelihood function. This is often due to a problem of missing data. A typical example is when data drive only information about a specific compartment, while the sizes of all the others are missing. For instance, in the SIR model, we may observe only the daily number of new infections, $i^*_{1:T}$. This is a quite realistic situation that occurs when there is no notification of recovery. Indeed, in this framework, the vector $r^*_{1:T}$ is missing, $r_{1:T}$ represents a latent variable, and the evaluation of the probability of the data requires marginalization over it:

$$\begin{aligned} L\left(\theta \mid i_{1:T} = i^*_{1:T}\right) &= \sum_{r^*_{1:T} \in \mathscr{R}} \Pr\left(i_{1:T} = i^*_{1:T}, r_{1:T} = r^*_{1:T} \mid \theta\right) \\ &= \sum_{r^*_{1:T} \in \mathscr{R}} \prod_{t=1}^{T} \binom{S(t)}{i^*(t)} \pi_{\mathrm{SI}}(t)^{i^*(t)} [1 - \pi_{\mathrm{SI}}(t)]^{S(t) - i^*(t)} \binom{I(t)}{r^*(t)} \pi_{\mathrm{IR}}^{r^*(t)} [1 - \pi_{\mathrm{IR}}]^{I(t) - r^*(t)}, \end{aligned} \tag{5}$$

where $\mathscr{R}$ is the subset of $\mathbb{I}^T$ that contains all possible sequences $r^*_{1:T}$ that are compatible with the observed series of infections. The structure and the cardinality of $\mathscr{R}$ often make the point-wise evaluation of the likelihood and likelihood-based methods computationally intensive.

Another frequent situation is when data are recorded with different granularities (e.g. weekly or monthly) or when only aggregated data are available, such as the average weekly number of new infections. In all these cases, the evaluation of the likelihood requires additional marginalization over unobserved quantities (e.g., the daily number of new infections) and becomes even more complex.

Note that, in the light of what has been discussed so far, what makes a likelihood hard to compute – even when it is theoretically tractable – is primarily the assumptions about the structure of dependencies and the presence of latent variables or missing data, rather than the choice of the probabilistic model (such as Binomial, Poisson, Gaussian, etc.).

Estimation methods

In this section, we review some of the principal frequentist and Bayesian methods for inferring the parameters governing statistical compartmental models defined in discrete time. We distinguish the following circumstances:

- simple model: it is possible to derive an analytical expression for the likelihood function;
- complex model: the structure of dependence among the involved random variables is too complex to derive an analytical expression for the likelihood function, regardless of the type of data available;
- complete data: all the time series needed to evaluate the likelihood function (when analytically available) are fully observed, and the data granularity matches the time discretization used in the model formulation;
- incomplete data: not all the information required to evaluate the likelihood function (when analytically available) is accessible, more specifically we distinguish two cases:
  a) the time series referred to one or more compartments are completely missing;

Table 1: Overview of the cases and inference methods considered in this work. The following approaches were analyzed: maximum likelihood estimation (MLE), Monte Carlo (MC) sampling, expectation-maximization (EM), data augmentation with Markov chain Monte Carlo (DA-MCMC), and approximate Bayesian computation (ABC).

Model structure	Case	Likelihood tractability	Frequentist methods	Bayesian methods
Simple (likelihood available)	Complete data	✓	MLE	MC
	Incomplete data – missing series	×	EM	DA-MCMC
	Incomplete data – sparse/aggregated	×	Calibration	ABC
Complex (likelihood unavailable)	Complete data	×	Calibration	ABC
	Incomplete data – missing series	×	Calibration	ABC
	Incomplete data – sparse/aggregated	×	Calibration	ABC

b) the available time series are incomplete, either because the data granularity is coarser than the model discretization (sparse data), or because the data are observed in an aggregated form (e.g., average values over weeks or months).

Table 1 summarizes the different scenarios considered and the corresponding estimation methods. It is worth noting from the beginning that likelihood-free methods, such as calibration and approximate Bayesian computation (ABC), can also be used when the likelihood function is available, while they are the only option when the likelihood is unavailable. In Section "Working example: the SIR model" we compare the results provided by likelihood-based methods with those provided by likelihood-free methods at work on the SIR example with incomplete data.

Frequentist approaches

From a frequentist point of view, we are interested in providing point estimates and confidence intervals around the estimates, to account for sampling variability. Depending on the availability of the likelihood function and complete data, we suggest three different solutions to come up with point estimates. As regards confidence intervals, we describe a bootstrap procedure suitable for all the considered methods.

Maximum likelihood estimation

When the likelihood function has a tractable analytical form as that in Equation (4) and complete data are available, one can infer the parameters governing transition rates via numerical maximization of the likelihood function using algorithms such as those described by Nelder et al. [26]. In the literature, there are several works dealing with the asymptotic theory of these estimators in compartmental models. In some of them, the asymptotic behavior is achieved by considering an increasing number of observations in a given time window – this is possible when working in continuous time – while others consider an increasing time window with $T \to \infty$. We do not deepen into the technicalities of the properties of such estimators and refer the readers to the review by Tang et al. [13]. This review, in discussing different maximum likelihood (ML) estimators, also reports methods for the estimation of their variance, resorting to composite likelihood strategies or martingale methods. Here, we suggest to implement a flexible bootstrap approach to compute confidence intervals in all the considered cases.

Expectation-maximization algorithm

The expectation-maximization (EM) algorithm [27] is a computational method for finding ML estimates when the likelihood function is available but intractable, and imputing hidden/missing data simplifies its evaluation.

This is the case of simple compartmental models that exhibit a tractable likelihood function when complete data are available, but the incompleteness of the data makes its evaluation infeasible.

Let us denote by $\mathbf{x}$ the observed data, by $\mathbf{z}$ some missing data, and by $\mathbf{y} = (\mathbf{x}, \mathbf{z})$ the complete data. The evaluation of the complete likelihood, $L(\theta \mid \mathbf{y})$, is straightforward but the evaluation of the incomplete likelihood, $L(\theta \mid \mathbf{x})$, is not. The EM algorithm iterates two steps:

- the E-step computes the expected value of the complete log-likelihood, $\ell(\theta \mid \mathbf{x}, \mathbf{z})$, w.r.t. missing variables $\mathbf{Z}$;
- the M-step maximizes this expected value w.r.t. to θ.

When the expected value is difficult to evaluate, its analytical evaluation can be replaced by a Monte Carlo estimate based on a sample of size m from the distribution of the latent variables $\mathrm{p}(\mathbf{z} \mid \mathbf{x}, \theta^{(s-1)})$, as proposed by Levine and Casella [28].

Algorithm 1 Expectation-maximization

Inizialize $\theta^{(0)}$ as random starting value
Set e
for s in 1: S **do**
  Assign $\theta^{(s)} = \arg\max_{\theta} E_{\mathcal{Z}p(\mathbf{z}|\mathbf{x},\theta^{(s-1)})}[\ell(\theta|\mathbf{x},\mathbf{z})] \approx \arg\max_{\theta} \frac{1}{m}\sum_{i=1}^{m} \ell(\theta|\mathbf{x},\mathbf{z}^{(i)})$
  if $\frac{\theta^{(s)}-\theta^{(s-1)}}{\theta^{(s-1)}} \leq e$ **then**
    Break
  end if
end for

Algorithm 1 summarizes the EM algorithm implemented in our working example where missing data $\mathbf{z}$ correspond to the series $r^*_{1:T}$, and observed data $\mathbf{x}$ are $i^*_{1:T}$ (see Section "Working example: the SIR model").

The main problem of the EM algorithm is that it is highly dependent on the starting points and does not ensure convergence when the expected value of the log-likelihood is too complex to optimize.

Calibration

In some cases, compartmental models are so complex that there is no analytical form for the associated likelihood function. In such cases, a possible solution to provide point estimates is a calibration procedure that consists of searching for the optimal parameter values that lead to an evolution of the system as close as possible to the observed one [29]. The procedure requires only the availability of a mathematical model that allows for producing forward simulations of the compartment sizes. Note that, hereafter, the term "calibration" will refer specifically to this process of minimizing a loss function, whereas in the literature, the term may have a broader meaning that includes several estimation methods [30].

Let us denote by $\mathbf{x} = \big(x(0), \dots, x(T)\big)$ the observed dynamic of the system, and by $\hat{\mathbf{x}}(\theta) = \big(\hat{x}(0;\theta), \dots, \hat{x}(T;\theta)\big)$ the dynamic simulated from the deterministic model when it takes the vector of parameter θ as an input. Given a discrepancy function $\rho(\cdot,\cdot)$, the calibration procedure optimizes the objective function, i.e. minimizes over θ the discrepancy between observed and simulated data:

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \rho\big(\mathbf{x}, \hat{\mathbf{x}}(\theta)\big). \tag{6}$$

The optimization is performed numerically and the results achieved by minimization algorithms often depend on the values at which they are initialized. To avoid the problem of getting stuck in local minima, we select the initial values through a preliminary optimization over a multidimensional grid [29, 31].

This method is very flexible since it completely disregards the likelihood formulation and relies only on the mathematical part of the model. This fact enables its use regardless of the reason why the likelihood is challenging to handle (e.g. complex model or high dimensional latent variables). This flexibility can also be leveraged to handle cases where even the observed dynamics are only partially recorded. In such cases, we

can define a function $s\colon \mathbb{R}^T \to \mathbb{R}^K$ with $K < T$ that maps the complete series $\mathbf{x}$ to a lower dimensional vector containing only the observed partial information. Examples are new infections recorded weekly, or the average number of new infections during a given interval of time. The calibration procedure is still applicable by setting $\rho\big(s(\mathbf{x}), s\big(\hat{\mathbf{x}}(\theta)\big)\big)$ as objective function.

Bootstrap procedure

In all of the three cases considered above, we can quantify sampling variability around the point estimate of θ and provide confidence intervals, by using a bootstrap procedure [32]. Specifically, we resort to a parametric bootstrap as proposed by Chowell [29]. This choice is motivated by the fact that, in the described framework, we observe a single univariate or multivariate time series and nonparametric bootstrap procedures are not suitable. The adopted parametric bootstrap procedure relies on the fact that, under the assumption that the specified stochastic model is valid, once a point estimate for θ is provided, the observed data represent just one possible realization of the variables $X(\hat{\theta})$. Thus, to account for sampling variability, we can generate B alternative realizations from the estimated distribution. Each of these hypothetical datasets would lead to a different estimate $\hat{\theta}^b$, and all such estimates form a bootstrap distribution, from which sampling variability can be quantified and confidence intervals can be derived (e.g., using the percentile method). This approach can be summarized as follows:

For $b \in \{1, \dots, B\}$

1. sample one dynamic $\mathbf{x}^b$ from the stochastic model with input $\hat{\theta}$, i.e. the MLE or the optimal parameters retrieved via numerical maximization/EM/calibration;
2. obtain an estimate $\hat{\theta}^b$ of θ, by alternatively:
   a) computing the ML estimate, using $\mathbf{x}^b$ as observed data;
   b) implementing the EM algorithm, using as observed data only the dynamics that are not missing, as derived from $\mathbf{x}^b$;
   c) calibrating the model searching for $\hat{\theta}^b$ minimizing the discrepancy function between simulated and estimated dynamics $\rho\big(\mathbf{x}^b, \hat{\mathbf{x}}(\theta)\big)$, or $\rho\big(s\big(\mathbf{x}^b\big), s\big(\hat{\mathbf{x}}(\theta)\big)\big)$.

We use the percentiles of the bootstrap samples $\hat{\theta}^1, \dots, \hat{\theta}^B$ to compute confidence intervals.

Bayesian approaches

In the statistical formulation of a compartmental model, observed data $\mathbf{x}$ are realizations of a random variable $\mathbf{X}$. In Bayesian statistics, the set of parameters that governs the probability distribution of $\mathbf{X}$ is, in turn, modeled as a random variable $\theta \in \Theta$ with a prior distribution, here denoted by $\pi(\cdot)$. Thus, given the observed data $\mathbf{x}$, the object of interest for the inference is the posterior distribution derived through Bayes's formula:

$$\pi(\theta \mid \mathbf{x}) = \frac{\pi(\theta) L(\theta \mid \mathbf{x})}{\int_{\Theta} \pi(\theta) L(\theta \mid \mathbf{x}) d\theta}, \tag{7}$$

where the denominator is the marginal likelihood, which is a normalizing constant. Often, the computation of this normalizing constant is infeasible and requires a numerical approximation.

When the model involves high dimensional latent variables, $\mathbf{Z}$, they should be integrated out to derive the posterior distribution:

$$\pi(\theta \mid \mathbf{x}) = \frac{\pi(\theta) \int_{\mathcal{Z}} L(\theta \mid \mathbf{x}, \mathbf{z}) \mathrm{d}\mathbf{z}}{\int_{\Theta} \int_{\mathcal{Z}} \pi(\theta) L(\theta \mid \mathbf{x}, \mathbf{z}) d\mathbf{z}\, d\theta}. \tag{8}$$

This is also the case of the SIR model with incomplete data, which requires the solution of several high-dimensional summations, as those in Equation (5). In the literature, several methods for addressing these problems and conducting Bayesian inference via simulations are available (see the comprehensive discussion of the use of these methods in the epidemiological framework by McKinley et al. [12]).

Monte Carlo methods

When the likelihood function is tractable and we observe complete data, the only hurdle is the computation of the normalizing constant in Equation (7). A possible strategy to overcome this problem is resorting to Monte Carlo methods, a class of algorithms aimed at solving inferential or optimization problems through stochastic simulations [33]. In particular, here we consider importance sampling (IS) [34], although other solutions are available (e.g., accept-reject and Markov chain Monte Carlo methods) – see the survey of these methods by Robert and Changye [35]. The main idea of this algorithm is to get samples of the model parameters from an easy-to-sample distribution, and then convert them into a sample from the posterior distribution by assigning an importance weight to each parameter proposal. The algorithm is summarized in Algorithm 2, where $q(\cdot)$ denotes the easy-to-sample proposal distribution.

Algorithm 2 Importance sampling

1: **Draw** $\theta^{(1:S)}$ i.i.d. from $q(\cdot)$
2: **Assign** to each $\theta^{(s)}$ an importance weight $\omega^{(s)} = \frac{\pi(\theta^{(s)})L(\theta^{(s)}|\mathbf{x})}{q(\theta^{(s)})}$
3: **Compute** normalised weights $\tilde{\omega}^{(s)} = \frac{\omega^{(s)}}{\sum_{i=1}^{S}\omega^{(i)}}$

The output is a weighted sample $(\theta^{(1)}, \tilde{\omega}^{(1)}), \ldots, (\theta^{(S)}, \tilde{\omega}^{(S)})$ drawn as a single batch. It can be used to estimate posterior quantities through weighted averages or by introducing a resampling step that uses normalized weights as probabilities. The final sample can be considered as an i.i.d. sample from the exact posterior distribution but the variability of the posterior estimates depends on the variability of the importance weights. This latter depends in turn on the choice of the proposal distribution that should be as close as possible to the target.

Data augmentation Markov chain Monte Carlo methods

MC methods and Markov chain Monte Carlo (MCMC) methods usually rely on point-wise evaluations of the likelihood function. In the presence of missing data, the evaluation of the likelihood function also requires the solution of high dimensional integrals/summations, as shown in Equation (8). To avoid their computation, a possible solution is to provide a sample from a posterior distribution defined on the augmented space $\Theta \times \mathscr{Z}$: $\pi(\theta, \mathbf{z} \mid \mathbf{x})$.

To this aim, data augmentation Markov chain Monte Carlo methods (DA-MCMC) can be implemented [36, 37]. Usually, these algorithms rely on a Gibbs sampling scheme and require the ability to get samples from the full conditional distributions $\pi(\theta \mid \mathbf{x}, \mathbf{z})$ and $\mathrm{p}(\mathbf{z} \mid \mathbf{x}, \theta)$, or the collapsed distribution $\mathrm{p}(\mathbf{z} \mid \mathbf{x})$. Sometimes, in compartmental models, their definition is not straightforward. Different strategies may be the implementation of a Metropolis-within-Gibbs algorithm, in which samples from unavailable full conditional distributions are obtained through Metropolis steps in the Gibbs sampling scheme, or the implementation of a Metropolis-Hastings (MH) algorithm to get samples directly on the joint space $\Theta \times \mathscr{Z}$, as displayed in Algorithm 3 (details on the above mentioned MCMC algorithms are in the book by Robert and Casella [38]).

Algorithm 3 Metropolis-Hastings

Inizialize $\theta^{(0)}$, $\mathbf{z}^{(0)}$
for s in 1: S **do**
  Propose missing data $\mathbf{z}^* \sim q_z(\cdot|\mathbf{z}^{(s-1)})$
  Propose $\theta^* \sim q_\theta(\cdot|\theta^{(s-1)})$
  Compute $\alpha = \min\left\{1, \frac{\pi(\theta^*)L(\theta^*|\mathbf{z}^*,\mathbf{x})q_z(\mathbf{z}^{(s-1)}|\mathbf{z}^*)q_\theta(\theta^{(s-1)}|\theta^*)}{\pi(\theta^{(s-1)})L(\theta^{(s-1)}|\mathbf{z}^{(s-1)},\mathbf{x})q_z(\mathbf{z}^*|\mathbf{z}^{(s-1)})q_\theta(\theta^*|\theta^{(s-1)})}\right\}$
  Sample $u \sim U(0,1)$
  if $u < \alpha$ **then**
    Set $\theta^{(s)} = \theta^*$ and $\mathbf{z}^{(s)} = \mathbf{z}^*$
  else
    set $\theta^{(s)} = \theta^{(s-1)}$ and $\mathbf{z}^{(s)} = \mathbf{z}^{(s-1)}$
  end if
end for

The output of the algorithm is a realization of a Markov chain, $[(\theta^{(0)}, \mathbf{z}^{(0)}), \dots, (\theta^{(S)}, \mathbf{z}^{(S)})]$. The algorithm satisfies the detailed balance condition (see Section *DA-MCMC and detailed balance condition* in Supplementary Material) thus the limiting distribution of the chain is $\pi(\theta, \mathbf{z} \mid \mathbf{y})$. This means that, as always in MCMC methods, samples are only asymptotically distributed according to the joint posterior, and checks for the convergence of the chain are needed.

The sampling scheme in Algorithm 3 is very general as it requires only the ability to evaluate the prior probability and the complete likelihood, and samples from the target distribution $\pi(\theta \mid \mathbf{x})$ can be easily retrieved by disregarding the sequence $(\mathbf{z}^{(0)}, \dots, \mathbf{z}^{(S)})$ from the final output. Problems may be related to the autocorrelation of the chain and the choice of good proposal distributions $q_\theta(\cdot)$ and $q_z(\cdot)$. To accelerate the convergence and improve the efficiency of the algorithm, many alternatives are available in the literature. Examples are algorithms that define adaptive proposal distributions [39] or the Hamiltonian version of the MH algorithm [40], which is efficient in the case of smooth density functions [41].

In the case of compartmental models, particular attention should be paid to the definition of the proposal distribution for the missing data $q_z(\cdot)$. A possible strategy is suggested by O'Neill et al.[17] (see Section "Results"). To overcome this problem, in some specific cases, one can also resort to particle marginal Metropolis-Hastings samplers which avoid the definition of a proposal distribution $q_z(\cdot)$ and approximate p($\mathbf{z} \mid \mathbf{y}$) through a sequential Monte Carlo algorithm [42].

Approximate Bayesian computation

All the Bayesian methods described so far require at least an analytical form for the complete likelihood function. In the case of complex models, the likelihood function may be unavailable and its analytical form cannot be written down. Furthermore, even in the case of a simple model with incomplete data, when the size of the latent variables is very large, DA-MCMC is computationally demanding. In such cases, likelihood-free methods, such as approximate Bayesian computation (ABC), result in being convenient due to their flexibility.

ABC is a broad class of methods allowing Bayesian inference on parameters governing complex models with intractable likelihood functions. The original intuition can be traced back to an explanation of the Bayes' Theorem provided by Rubin in the 80s [43], but primal ABC algorithms have been formalized by Tavaré et al. [44] and Pritchard et al. [45]. ABC methods dispense with exact likelihood computation and only require the ability to simulate pseudo-data by sampling observations from the assumed statistical model employing a computer program that reproduces the stochastic data generative process, usually called a "simulator", here denoted by $\Pr(\cdot \mid \cdot)$. The underlying idea of ABC methods is to convert samples from the prior distribution into samples from the posterior through three simple steps:

1. Draw S parameter proposals from the prior distribution $\pi(\cdot)$;
2. Give each parameter $\theta^{(s)}$ as an input to the simulator to sample pseudo-data $\hat{\mathbf{x}}(\theta^{(s)})$;

3. Retain only parameter proposals $\theta^{(s)}$ such that $\hat{\mathbf{x}}(\theta^{(s)}) = \mathbf{x}$.

The output is a sample from the exact posterior distribution. However, in practice, several sources of approximation are introduced by replacing the equality constraint at the third step with a "closeness" constraint: parameter proposals are retained when $\rho(\hat{\mathbf{x}}(\theta^{(s)}); \mathbf{x}) \leq e$, where $\rho(\cdot;\cdot)$ is a distance function and e is positive tolerance threshold. Note that the acceptance rate depends also on the cardinality of the space on which observed data live. Thus, a common practice is that of reducing data dimensionality and comparing observed and simulated data through summary statistics $s(\cdot)$ that retain only key information for inference. It follows that the output becomes an i.i.d. sample from an approximate posterior distribution, the closeness of which to the true one depends on the choice of $\rho(\cdot;\cdot)$, the selected summaries $s(\cdot)$ and the magnitude of e. Note that the use of summary statistics, in addition to the purposes of dimensionality reduction, offers a natural way to address the issue of partially observed data, in a similar fashion to what was explained for the calibration procedure in Section "Calibration".

In the literature, several advanced sampling schemes have been proposed (we refer the reader to the handbook by Sisson et al. [46] for a comprehensive description of the method and to Kypraios et al. [47] for an introduction to ABC for stochastic epidemic models). Most of them are sequential methods based on a decreasing sequence $e_1 \geq e_2 \geq \ldots \geq e_S$ of thresholds, rather than a fixed tuning parameter. Usually, these algorithms rely on the definition of a sequence of tempered target distributions based on the sequence of thresholds, and get samples from each of them using an importance sampling step. Examples are the population Monte Carlo ABC presented by Beaumont et al. [48] and some adaptive versions inspired by it (see Algorithm 4). Here, we relied on one of the strategies proposed by Lenormand et al. [49], where new thresholds are automatically selected during the execution of the algorithm in a way that ensures a decreasing level of approximation from the given iteration to the next one.

Algorithm 4 Adaptive population Monte Carlo ABC

Initialize e_1
for j in 1: M **do**
  Simulate $\theta_j^{(1)} \sim \pi(\cdot)$ and $\hat{\mathbf{x}}(\theta_j^{(1)}) \sim \Pr(\cdot \mid \theta_j^{(1)})$ until $\rho(\hat{\mathbf{x}}(\theta_j^{(1)}); \mathbf{x}) < e_1$
  Set $\omega_j^{(1)} = \frac{1}{M}$
end for
Select e_2 using an adaptive strategy.
for s in 2: S **do**
  Set Σ_s to twice the empirical covariance matrix of $\theta_1^{(s-1)}, \ldots, \theta_M^{(s-1)}$
  for j in 1: M **do**
    Pick θ_j^* from $(\theta_1^{(s-1)}, \ldots, \theta_M^{(s-1)})$ with probabilities $(\omega_1^{(s-1)}, \ldots, \omega_M^{(s-1)})$
    Generate $\theta_j^{(s)} | \theta_j^* \sim MVN\left(\theta_j^*, \Sigma_s\right)$ and $\hat{\mathbf{x}}(\theta_j^{(s)}) \sim \Pr(\cdot \mid \theta_j^{(s)})$
    Set $\omega_j^{(s)} \propto \frac{\pi(\theta_j^{(s)})}{\sum_{m=1}^{M} \omega_m^{(s-1)} \phi\left\{\tau_s^{-1}(\theta_m^{(s)} - \theta_m^{(s-1)})\right\}} \mathbb{1}\left\{\rho(\hat{\mathbf{x}}(\theta_j^{(s)}); \mathbf{x}) < e_s\right\}$
      where ϕ represents the density of the Standard Normal distribution
    end for
    Select e_{s+1} using an adaptive strategy.
end for

Note that the ABC procedure is very similar in spirit to the calibration. However, ABC algorithms use the statistical model and take into account two sources of uncertainty: the uncertainty on the parameter space and sampling variability.

Results

In this section, we show the results of the presented frequentist and Bayesian methods on the working example presented in Section "Working example: the SIR model". Furthermore, we show the results of likelihood-free methods in the case of a complex model with unavailable likelihood.

Working example: the SIR model

We applied all the methods described in Section "Estimation methods" on simulated data obtained from the stochastic version of the SIR model defined in Equation (3), after setting $T = 90$, $\tau^{\text{true}} = 14$, $R_0^{\text{true}} = 3$, $S(0) = 9,990$, $I(0) = 10$, and $R(0) = 0$. In particular, we simulated the two-time series $i^*_{1:T}$ and $r^*_{1:T}$, the knowledge of which is sufficient to reconstruct compartment sizes $S^*_{1:T}$, $I^*_{1:T}$ and $R^*_{1:T}$, given $S(0)$, $I(0)$ and $R(0)$. We focused on the estimate of R_0 and τ considering the initial size of the compartments as known. Regarding Bayesian inference, it has been conducted using uniform prior distributions: $\tau \sim U[7, 21]$ and $R_0 \sim U(0, 6]$. We considered both the case of complete and incomplete data.

Complete data

Let us consider the case in which we observe the entire evolution of the compartment sizes, equivalent to observing both $i^*_{1:T}$ and $r^*_{1:T}$. We can compute MLE for R_0 and τ and their posterior distributions via IS, respectively from a frequentist and a Bayesian point of view.

Regarding MLE, the maximization of the complete likelihood in Equation (4) has been performed using Nelder's procedure implemented in the Optim package of JULIA software [26, 50]. Then, the bootstrap procedure described in Section "Frequentist approaches" has been implemented with B =1,000. More precisely, to be consistent with the likelihood function in Equation (4), at each bootstrap iteration b, we draw samples from two Binomial distributions: $\left(S^b(t-1) - S^b(t)\right) \sim Binom\left(S^b(t-1), 1 - \exp(-\frac{\hat{R}_0}{\hat{\tau}} \frac{I^b(t-1)}{S(0)})\right)$ and $\left(R^b(t) - R^b(t-1)\right) \sim Binom\left(R^b(t-1), 1 - \exp(-\frac{1}{\hat{\tau}})\right)$, with $\hat{R}_0$ and $\hat{\tau}$ the maximum likelihood estimates of the model parameters. As regards the IS implementation, we used the joint prior distribution of the parameters as proposal distribution.

Table 2 reports the results in terms of ML estimates with 90 % bootstrap intervals, and maximum a posteriori (MAP) estimates with 90 % highest posterior density (HPD) intervals. These results, as well as Figure 3, show that both methods provide point estimates very close to the true parameters.

Incomplete data

Let us consider the case in which we observe only the series $i^*_{1:T}$, meaning that information about newly resolved infections is not available. We can resort to EM and DA-MCMC algorithms, respectively in the frequentist and Bayesian setting. In our implementation, computing the expected value in the E-step would involve a complex summation over $\mathscr{R}$, thus at each iteration s we use an MC estimate based on m =1,000 simulations from $p(r_{1:T} \mid i^*_{1:T}, \theta^{(s-1)})$ (see *SIR model: sampling missing data and pseudo-data* in Supplementary Material). As regards the

Table 2: Frequentist versus Bayesian inference on the parameters of the SIR model, in the case of complete data. Results are reported in terms of point estimate and 90 % confidence intervals in the frequentist case, and in terms of maximum a posteriori (MAP) and 90 % highest posterior density (HPD) intervals in the Bayesian case.

	Frequentist	Bayesian
Complete data	**Maximum likelihood**	**Importance sampling**
τ	14.00 (13.73–14.25)	13.96 (13.75–14.25)
R_0	3.00 (2.92–3.08)	3.01 (2.93–3.08)

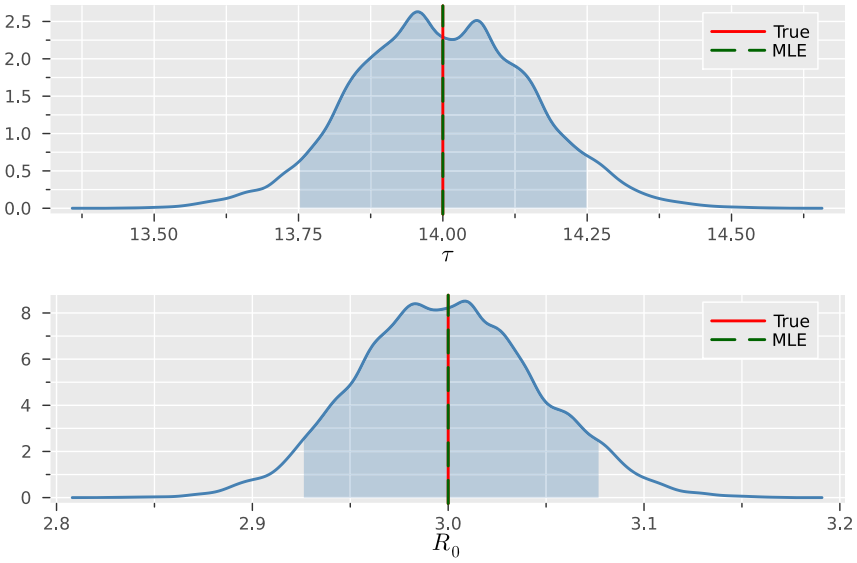

Figure 3: Marginal posterior distributions of SIR model parameters obtained through importance sampling (IS), with shaded areas representing the 90 % highest posterior density (HPD) intervals. Vertical lines indicate the true parameter value, along with its maximum likelihood estimate (MLE).

DA-MCMC proposal distributions, we resorted to Gaussian multivariate proposal distributions for $\theta = (\tau, R_0)$, and to a mixture of proposal distributions for $r_{1:T}$. In particular, we sample from the mixture of the three proposal distributions described by O'Neill and Roberts [17]: at each iteration s of the DA-MCMC algorithm, we select at random one of the following small perturbations:

1. add a resolution: select at random t and propose a series in which $r^{(s)}(t)$ is set to $r^{(s-1)}(t) + 1$;
2. subtract a resolution: select at random t and propose a series in which $r^{(s)}(t)$ is set to $r^{(s-1)}(t) - 1$;
3. move a resolution: select at random (t_1, t_2) and propose a series in which $r^{(s)}(t_1)$ is set to $r^{(s-1)}(t_1) - 1$ and $r^{(s)}(t_2)$ is set to $r^{(s-1)}(t_2) + 1$.

We evaluate the probability of this new proposal and use it in the computation of the MH acceptance ratio.

Note that the longer is the observed series, the higher is the cardinality of the space of the latent variables. This makes proposal distributions based on small perturbations inefficient and the chain strongly autocorrelated. In such a case, likelihood-free algorithms should be implemented to obtain more efficient posterior estimates.

This illustrative example provides the opportunity to compare the implementation of likelihood-based and likelihood-free methods in the case of incomplete data, as both of these approaches are feasible. Hence, we tested also calibration and ABC algorithms for inferring model parameters when only $i^*_{1:T}$ is observed. In both the algorithms, we compared observed and simulated trajectories through the Euclidean distance:

$$\rho(\hat{\mathbf{x}}(\theta), \mathbf{x}) = \frac{1}{T} \sum_{t=1}^{T} \sqrt{\left(i^*(t) - \hat{i}(t;\theta)\right)^2}, \tag{9}$$

where $\hat{i}(1;\theta), \ldots, \hat{i}(T;\theta)$ denotes the series simulated through the deterministic or the stochastic SIR model, in the calibration and ABC procedure respectively, when the vector θ is given as an input (for a description of the procedure for simulating data from the statistical model see *SIR model: sampling missing data and pseudo-data* in Supplementary Material).

The calibration procedure is performed using the optimization strategy implemented in the JULIA package Optim [50]. To avoid the problem of getting stuck in local minima, we performed several optimizations using different starting points, then we selected the solution that minimized the distance function in Equation (9) [29,

31]. To consider sampling variability and compute confidence intervals, we implement the parametric bootstrap procedure described in Section "Frequentist approaches". Here, at each bootstrap iteration, the optimization algorithm has been initialized at random starting values.

As regards the ABC procedure, we implemented the Algorithm 4. At each iteration, we compared the posterior distributions approximated by ABC with those provided by the DA-MCMC after the assessment of the convergence of the chain (see Section *Bayesian algorithms: IS, DA-MCMC, ABC* in Supplementary Material for further details). Looking at the Kullback-Leibler divergence and the Hellinger distance between these distributions, it turned out that after 80 iterations they are quite stable.

Figure 4 and Table 3 show that all the results of the implemented methods are coherent with R_0^{true} and τ^{true}. From a frequentist perspective, point estimates provided by EM and calibration are very close to each other and to true values. It is worth noting that likelihood-based methods are challenging when the shape of the likelihood function makes the optimization difficult. Figure 5 shows the contour plot of the log-likelihood function (a) and one of its expected values (b), respectively in the case of complete and incomplete data. It is apparent that the function in panel (b) assumes approximately the same value whatever is τ when R_0 is close to its true value. Thus, the optimization algorithm solution strongly depends on the starting values. To overcome this problem and get reliable point estimates, we run each M-step several times with different starting values. This makes the EM algorithm inefficient and the bootstrap procedure infeasible, that is the reason why we do not provide confidence intervals in Table 3.

From a Bayesian perspective, we can see that posterior distributions concentrate around the true values of the parameters and that ABC can retrieve an approximate posterior distribution close to the true posterior computed via DA-MCMC (see Section *B. algorithms: IS, DA-MCMC, ABC* in Supplementary Material for further details). We can consider the negligible approximation introduced by likelihood-free methods as the price to pay for having more general and flexible methods that avoid the definition of the likelihood function, its evaluation, and the imputation of missing data. Further details about the implementation of the algorithms are in Section *SIR model: further details on the algorithm implementations* in Supplementary Material.

As a general comment, we can conclude that both from a frequentist and Bayesian point of view, likelihood-free methods have some advantages. The bootstrap procedure appears more feasible and fast using the calibration rather than the EM algorithm, while the ABC algorithm overcomes problems related to the

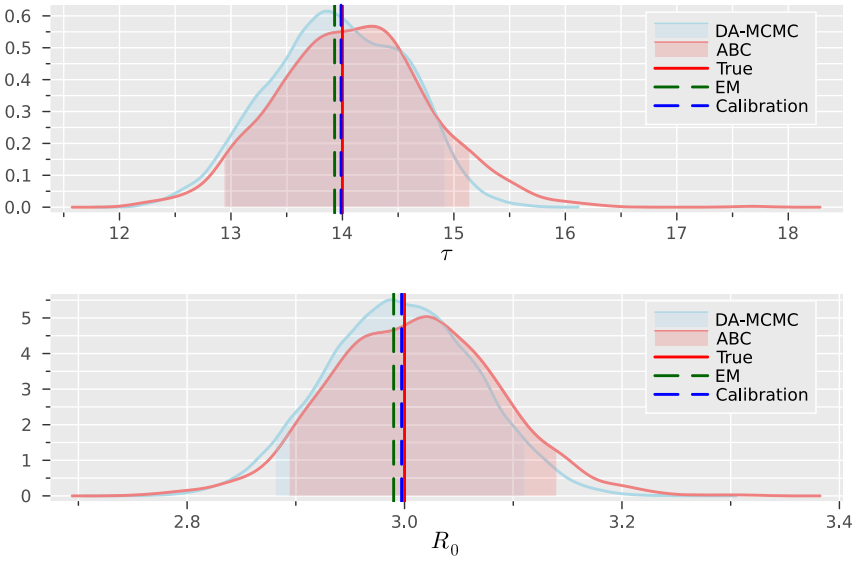

Figure 4: Marginal posterior distributions of SIR model parameters obtained through data augmentation Markov chain Monte Carlo methods (DA-MCMC) and approximate Bayesian computation (ABC), with shaded areas representing the 90 % highest posterior density (HPD) intervals. Vertical lines indicate the true parameter value, along with its estimates from expectation-maximization (EM) and calibration methods.

Table 3: Frequentist versus Bayesian inference on the parameters of the SIR model in the case of incomplete data. Results are reported in terms of point estimate and 90 % confidence intervals in the frequentist case and in terms of maximum a posteriori (MAP) and 90 % highest posterior density (HPD) intervals in the Bayesian case.

	Frequentist	**Bayesian**
Likelihood-based	**Expectation-maximization**	**Data augmentation Markov chain Monte Carlo methods**
τ	13.93	14.08 (13.18–15.09)
R_0	2.99	3.00 (2.90–3.12)
Likelihood-free	**Calibration**	**Approximate Bayesian computation**
τ	14.01 (13.61–14.40)	13.97 (12.93–15.15)
R_0	3.00 (2.95–3.06)	3.02 (2.88–3.14)

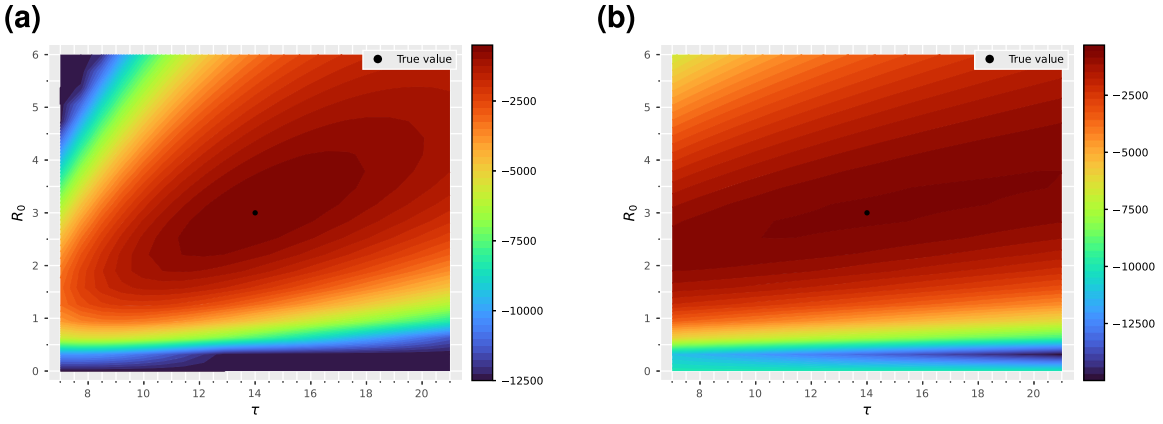

Figure 5: Contour plots of the log-likelihood for the SIR model parameters in the case of complete data (a), and of its expected value with respect to the distribution of missing variables in the case of incomplete data (b).

autocorrelation of the Markov chain and took only 9 min to get posterior quantities very close to those based on the DA-MCMC, the running time of which was equal to 11 min. Furthermore, likelihood-free methods allow a straightforward evaluation of the predictive distributions through forward simulations, appropriately accounting for all sources of uncertainty. Figure 6 shows the evolution of the compartment sizes over time until $T = 90$, estimated via calibration and ABC. Solid lines are day-by-day punctual estimates of the compartment sizes – i.e. the trajectories computed using MLE parameters (a) and day-by-day mean of the posterior predictive distribution (b). Their closeness to the observed data (dotted lines) suggests a proper fit of the model when the inference is performed via likelihood-free approaches. Confidence and credible bands are retrieved by calculating day-by-day the bootstrap quantiles and the highest posterior density intervals of the predictive distributions, respectively.

Note that, in the case of complete data, the frequentist and Bayesian approaches provide similar results also in terms of uncertainty – see confidence and credible intervals in Table 2. On the contrary, when the inference is based on incomplete data, Bayesian credibility intervals of the two model parameters are wider than the bootstrap confidence intervals (see Table 3), possibly because in the Bayesian approach the imputation of missing data occurs from a distribution that incorporates both the sampling variability and the uncertainty on the parameter space.

The use of summary statistics

In this example, we can consider at least two scenarios where replacing the observed/observable series $i^*_{1:T}$ with some low-dimensional statistics might be required or useful.

Case 1 Sparsely observed data. The series $i^*_{1:T}$ is not fully observed, and only the weekly number of new infections is available – i.e. $s(i^*_{1:T}) = \left(\sum_{t=1}^{7} i^*_t, \sum_{t=8}^{14} i^*_t, \dots, \sum_{t=T-6}^{T} i^*_t \right)$. In this case, DA-MCMC and EM methods would require imputing a too large number of missing data points, becoming infeasible.

Case 2 High dimensional observed data. The time series $i^*_{1:T}$ is very long and ABC and calibration methods become inefficient. In this case, summary statistics can be used to reduce dimensionality and summarize the information about θ driven by $i^*_{1:T}$. Here, we considered the weekly moving averages and the autocorrelation at lags that are multiples of 7.

Table 4 and Figure 7 demonstrate that likelihood-free methods can achieve results comparable to those obtained using the complete daily infection data, even when relying only on partial information.

Sensitivity to the choice of the distance function

To verify the robustness of the methods to variations in the distance function, both in ABC and calibration, the analyses were repeated using the chi-squared metric (Case 1) as distance function instead of the Euclidean distance. Table 5 and Figure 8 show that the results appear to be robust to changes in the distance function.

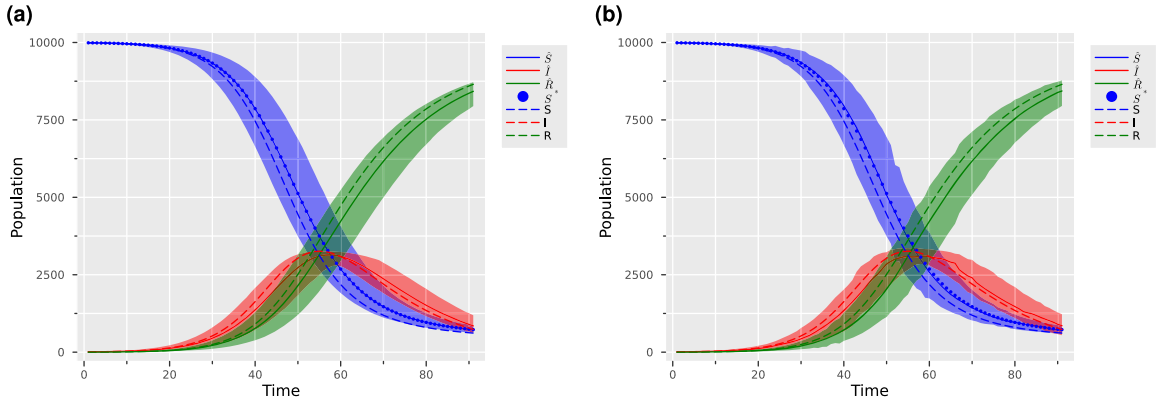

Figure 6: Trajectories of the SIR model estimated by calibration (a) and approximate Bayesian computation (ABC) (b). Results are reported in terms of point estimates (solid lines) and 90 % confidence bands (a) and in terms of posterior means (solid lines) and 90 % highest posterior density (HPD) bands (b). Dashed lines represent the trajectories of the deterministic model used to simulate the observed data. Observed data (dots) are reported as the trajectory $S^*_{1:T}$ since the observed series $i^*_{1:T}$ can be completely derived from it.

Table 4: Frequentist versus Bayesian inference on the parameters of the SIR model. Case 0 is based on the daily series of new infections, Case 1 is the case of sparsely observed data, and Case 2 is the case of high dimensional observed data. Results are reported in terms of point estimates and 90 % confidence intervals in the frequentist case, and in terms of maximum a posteriori estimates and 90 % highest posterior density intervals in the Bayesian case.

Calibration	**Case 0**	**Case 1**	**Case 2**
τ	14.01 (13.65–14.39)	14.00 (13.64–14.45)	14.03 (13.67–14.42)
R_0	3.00 (2.95–3.07)	3.00 (2.95–3.07)	3.00 (2.96–3.07)
ABC	**Case 0**	**Case 1**	**Case 2**
τ	14.05 (12.89–15.10)	14.15 (13.22–15.12)	14.10 (13.20–15.30)
R_0	3.00 (2.88–3.13)	3.00 (2.90–3.12)	3.00 (2.88–3.12)

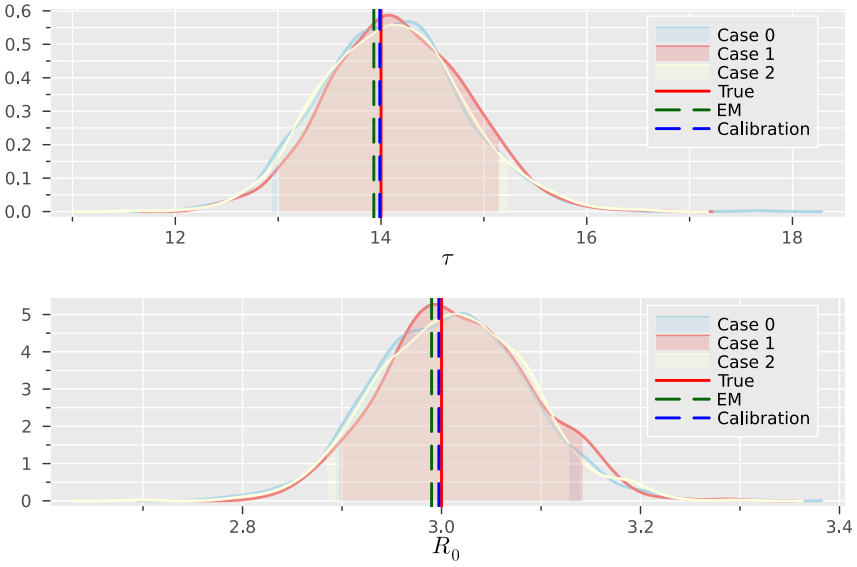

Figure 7: Marginal posterior distributions of SIR model parameters obtained. Case 0 is based on the daily series of new infections, case 1 is the case of sparsely observed data, and case 2 is the case of high dimensional observed data. Shaded areas represent the 90 % highest posterior density (HPD) intervals. Vertical lines indicate the true parameter value, along with its estimates from expectation-maximization (EM) and calibration methods.

Table 5: Frequentist versus Bayesian inference on the parameters of the SIR model. Case 0 is based on the Euclidean distance and Case 1 is based on the chi-squared metric. Results are reported in terms of point estimates and 90 % confidence intervals in the frequentist case, and in terms of maximum a posteriori (MAP) estimates and 90 % highest posterior density (HPD) intervals in the Bayesian case.

Calibration	**Case 0**	**Case 1**
τ	14.01 (13.65–14.39)	14.01 (13.67–14.45)
R_0	3.00 (2.95–3.07)	3.00 (2.96–3.07)
ABC	**Case 0**	**Case 1**
τ	14.05 (12.89–15.10)	14.17 (12.98–15.05)
R	3.00 (2.88–3.13)	2.97 (2.90–3.11)

A real-world example: the SHC model

In this section, we consider the Smoking Habits Compartmental (SHC) model, developed by Lachi and colleagues [51], as an example of a complex compartmental model where the analytical form of the likelihood function is unavailable. The SHC model has been designed to describe the evolution of smoking habits in a population over the years. Here it is implemented to estimate smoking dynamics in Tuscany, a region of Central Italy, from 1993 to 2019, and forecasts them until 2043.

The model assumes that, at each point in time, the alive population is divided into the following non-overlapping compartments: never (N), current (C), and former (F) smokers. The compartments C and F are further divided into sub-compartments denoted by C_i and F_i, where $i \in \{l, m, h\}$ indicates the level of smoking intensity, corresponding to low (< 10 cigarettes/day), medium (≥ 10 and < 20 cigarettes/day), and high (≥ 20 cigarettes/day) smoking intensity, respectively. From each compartment, subjects can transit to a deceased compartment denoted by the letter D and a subscript corresponding to the compartment of origin. New births in the year t, denoted by $\nu(t)$ (that for of simplicity was assumed to be constant over time), increase the size of the compartment N. Transitions of the individuals from a given compartment to another one are governed by the probabilities of starting smoking (γ_i), stopping smoking (ε_i), and relapsing into smoking after having stopped (η_i). Death happens with different probabilities for never (δ_N), current (δ_{C_i}), and former (δ_{F_i}) smokers

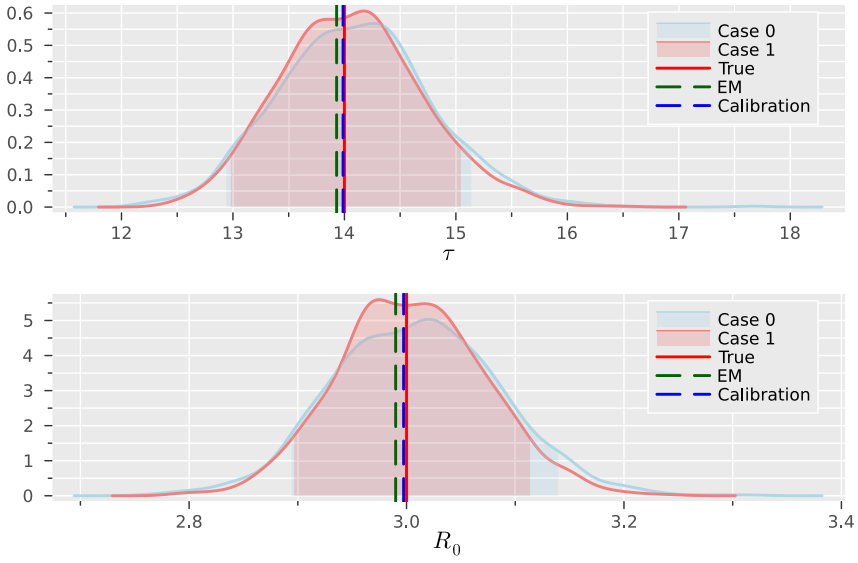

Figure 8: Marginal posterior distributions of SIR model parameters. Case 0 is based on the euclidean distance and Case 1 is based on the chi-squared metric. Shaded areas represent the 90 % highest posterior density (HPD) intervals. Vertical lines indicate the true parameter value, along with its estimates from expectation-maximization (EM) and calibration methods.

belonging to the smoking level category i. In Figure 9 a simplified version of the SHC model, which does not consider subjects' age, is depicted. Considering discrete time on an annual scale, $t \in \{1, \dots, T\}$, and introducing separate compartments for each age, $a \in \{1, \dots, 100\}$, as well as stratification by years since smoking cessation (c) for former smokers, the following system of difference equations arises:

$$
\begin{cases}
N(t;a) = \nu(t) & \text{if } a = 0 \\
N(t;a) = N(t-1;a-1)\big(1-\delta_N(a-1)\big)\big(1-\gamma(a-1)\big) & \text{if } a > 0 \\
C_i(t;a) = 0 & \text{if } a = 0 \\
C_i(t;a) = C_i(t-1;a-1)\big(1-\delta_{C_i}(a-1)\big)\big(1-\varepsilon(a-1)\big) + & \\
\qquad N(t-1;a-1)\big(1-\delta_N(a-1)\big)\pi_{C_i}\gamma(a-1) + & \\
\qquad \sum_{c>0} F_i(t-1;a-1,c-1)\big(1-\delta_{F_i}(a-1,c-1)\big)\eta(c-1) & \text{if } a > 0 \\
F_i(t;a,c) = 0 & \text{if } a = 0, c \geq 0 \\
F_i(t;a,c) = C_i(t-1;a-1)\big(1-\delta_{C_i}(a-1)\big)\varepsilon(a-1) & \text{if } a > 0, c = 0 \\
F_i(t;a,c) = F_i(t-1;a-1,c-1)\big(1-\delta_{F_i}(a-1,c-1)\big)\big(1-\eta(c-1)\big) & \text{if } a > 0, c > 0 \\
D_N(t;a) = 0 & \text{if } a = 0 \\
D_N(t;a) = D_N(t-1;a) + N(t-1;a-1)\delta_N(a-1) & \text{if } a > 0 \\
D_{C_i}(t;a) = 0 & \text{if } a = 0 \\
D_{C_i}(t;a) = D_{C_i}(t-1;a) + C_i(t-1;a-1)\delta_{C_i}(a-1) & \text{if } a > 0 \\
D_{F_i}(t;a,c) = 0 & \text{if } a = 0, c \geq 0 \\
D_{F_i}(t;a,c) = 0 & \text{if } a > 0, c = 0 \\
D_{F_i}(t;a,c) = D_{F_i}(t-1;a,c) + F_i(t-1;a-1,c-1)\delta_{F_i}(a-1,c-1) & \text{if } a > 0, c > 0
\end{cases}
\tag{10}
$$

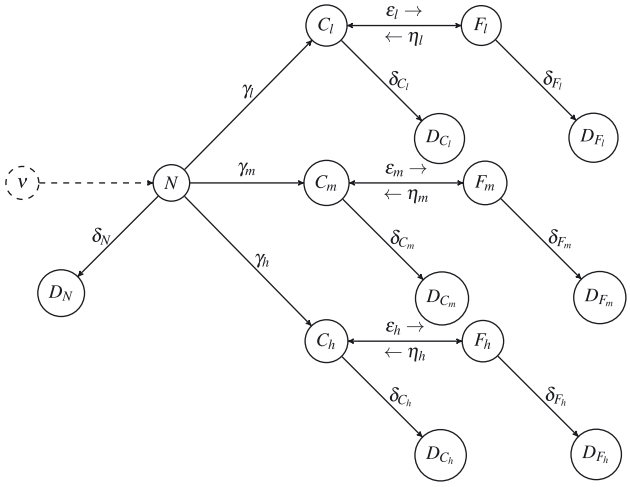

Figure 9: Smoking habits compartmental model (SHC) in its simplest form [51].

where $\boldsymbol{\pi} = (\pi_{C_l}, \pi_{C_m}, \pi_{C_h})$ denotes the distribution of the level of smoking intensity among the new current smokers. Note that the probability of starting smoking $\gamma(a)$, as well as the risks of dying $\delta_N(a)$ and $\delta_{C_i}(a)$, depend on the age a, while the probability of relapsing into smoke $\eta(c)$ depends on the years from smoking cessation c and the risk of dying for former smokers $\delta_{F_i}(a, c)$ depends both on a and c. From these dependencies follows that also the number of compartments in the model depends on the values of a and c.

More specifically, $\gamma(a)$ and $\varepsilon(a)$ are modeled through natural cubic regression splines of age with 2 equidistant internal knots, having parameters $\boldsymbol{\psi} = (\psi_0, \psi_1, \psi_2, \psi_3)$ and $\boldsymbol{\phi} = (\phi_0, \phi_1, \phi_2, \phi_3)$, respectively. Regarding the probability of relapsing into smoke $\eta(c)$, it was assumed to vary with time since cessation, according to a negative exponential function governed by positive parameters $\boldsymbol{\omega} = (\omega_0, \omega_1)$. Details on these functions are provided in Section *Transition probabilities* in Supplementary Material.

Mortality risks are $\delta_{C_i}(a) = RR_{C_i} \times \delta_N(a)$ and $\delta_{F_i}(a, c) = RR_{F_i}(c) \times \delta_N(a)$, with RR_{C_i} and $RR_{F_i}(c)$ the relative risks of dying for current smokers belonging to the smoking level category i and for people who stopped smoking since c years belonging to the same smoking level category, versus never smokers.

The distribution of the level of smoking intensity among the new current smokers, $\boldsymbol{\pi}$, is assumed as fixed at values obtained from the National Institute of Statistics (ISTAT) Multipurpose Surveys "Aspect of Daily Life" (www.istat.it/it/archivio/91926). Observed data are assumed to be the prevalence of current, never, and former smokers in the observed age classes a^* (14–17, 18–19, 20–24, 25–34, 35–44, 45–54, 55–59, 60–64, 65–74, 75+), for the years in the interval 1993–2019, denoted by $p_C^{\text{obs}}(t; a^*)$, $p_N^{\text{obs}}(t; a^*)$, $p_F^{\text{obs}}(t; a^*)$. Thus, the goal of the inference is to estimate, using two separate models by gender, the vector of unknown model parameters $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\phi}, \boldsymbol{\omega})$, given observed data and fixed quantities. Note that in our model formulation, the age-specific risk of dying for never-smokers is unknown, but in this analysis, we treat it as an ancillary quantity. Specifically, we preliminary computed $\delta_N(a)$ as described in a previous work [51] .[1]

The system of equations (10) formalizes the assumed deterministic model, but to perform both Bayesian and frequentist analyses, we need to formalize the likelihood function that relates unknown quantities, $\boldsymbol{\theta}$, to the observed prevalence. To this end, we must specify the stochastic generative process associated with the SHC model. For the sake of a straightforward description of the statistical model, let us denote by X the number of individuals who transit from a generic compartment to another one. We assume $X \sim Binom(n_x, q_x)$, where n_x is the number of individuals allowed to transit and q_x is the probability of that transition, whatever is the

1 If $a \in a^*$, we calculated

$$\delta_N(a) = \frac{1}{T} \sum_t \frac{\delta_{pop}(t; a)}{p_N^{\text{obs}}(t; a^*) + RR_C\, p_C^{\text{obs}}(t; a^*) + RR_F\, p_F^{\text{obs}}(t; a^*)},$$

where $\delta_{\text{pop}}(t;\ a)$ is the mortality rate in the population of age a at time t, obtained from ISTAT (www.istat.it), while RR_C and RR_F are the relative risks of dying for current smokers and former smokers versus never smokers obtained from the literature [52].

compartment. As an example, consider the number of smokers of age a with a low intensity that quit smoking at time t: n_x is the number of current smokers with low intensity and age a that do not die during the year t, and q_x is equal to $\epsilon(a)$. The same reasoning applies to the number of individuals relapsing smoking and the number of deaths. While the number of individuals who start smoking at age a is distributed according to a Multinomial distribution with the vector of probabilities $(\gamma_l(a), \gamma_m(a), \gamma_h(a))$.

Despite the simplicity of the assumed distributions, it is apparent that the analytical form of the likelihood function is infeasible to write down due to the complexity of the structure of dependence among variables: the equations governing the SHC model (10) are complex and involve a high number of compartments, the transition probabilities are complex functions of age and time from cessation. Only the prevalence $p_C^{\text{obs}}(t; a^*)$, $p_N^{\text{obs}}(t; a^*)$, $p_F^{\text{obs}}(t; a^*)$ are observed and the number of transitions that occur at each t, as well as the size of all compartments at each point in time, represent latent variables. However, the simulation of the stochastic data generative process is straightforward, thus likelihood-free methods such as calibration or ABC are the best solution to infer $\boldsymbol{\theta}$.

SHC model results

Let $p(t; a^*, \boldsymbol{\theta}) = \big(p_C(t; a^*, \boldsymbol{\theta}), p_N(t; a^*, \boldsymbol{\theta}), p_F(t; a^*, \boldsymbol{\theta})\big)$ be the vector of the prevalence of never, current and former smokers belonging to the class of age a^* at time t, predicted by the model in Equation (10), given a specific value of the parameters $\boldsymbol{\theta}$. Within the frequentist framework, we calibrate the model by searching for the value of $\boldsymbol{\theta}$ that minimized the Hellinger distance between the predicted trajectories and the observed ones, defined as follows [53]:

$$Obj(\boldsymbol{\theta}) = \frac{1}{T \times A^* \times \sqrt{2}} \sum_{t,a^*} \sqrt{\sum_{k \in \{C,N,F\}} \left(\sqrt{p_k(t; a^*, \boldsymbol{\theta})} - \sqrt{p_k^{obs}(t; a^*)} \right)^2}. \tag{11}$$

We quantify sampling variability around point estimates by using the parametric bootstrap procedure described in Section "Frequentist approaches". As in the previous example, at each bootstrap iteration, we initialized the optimization algorithm at random starting values. Within the Bayesian framework, we implemented the ABC Algorithm 4 by using the same distance function in Equation (11), and uniform prior distributions on the parameters. In particular, we specified uniform priors $U[-10, 10]$ on the spline parameters and $U(0, 10]$ as prior distribution on the exponential function parameters. Details on the implementation of the two algorithms are reported in Section *Further details of the algorithm implementations* in Supplementary Material.

Figure 10 presents the results of the inference on $\boldsymbol{\theta}$ obtained via calibration and ABC for males and females. The intervals obtained via ABC are wider compared to those obtained through calibration. Overall, the point estimates are similar for all parameters except ω_0, which shows a significant difference between the two procedures. However, the estimate obtained through calibration lies within the credible intervals derived from ABC. Figure 11 displays the estimates of the transition probabilities ($\gamma(a)$, $\varepsilon(a)$, and $\eta(c)$), obtained by calibration

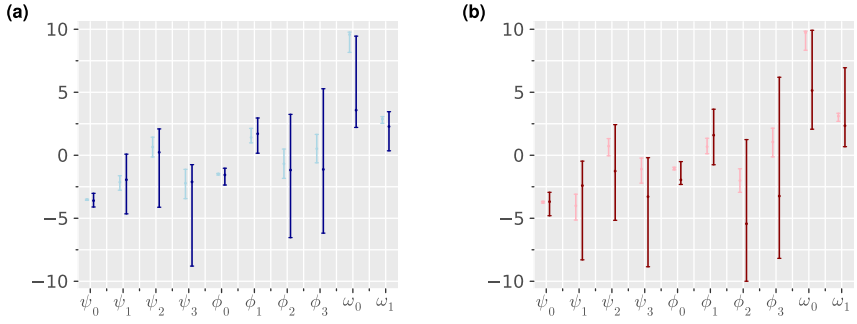

Figure 10: Estimates of SHC model parameters obtained through calibration (light color) and approximate Bayesian computation (ABC) (dark color) for males (a) and females (b). Results are reported in terms of point estimate and 90 % confidence intervals in the case of calibration and in terms of maximum a posteriori (MAP) and 90 % highest posterior density (HPD) intervals in the case of ABC.

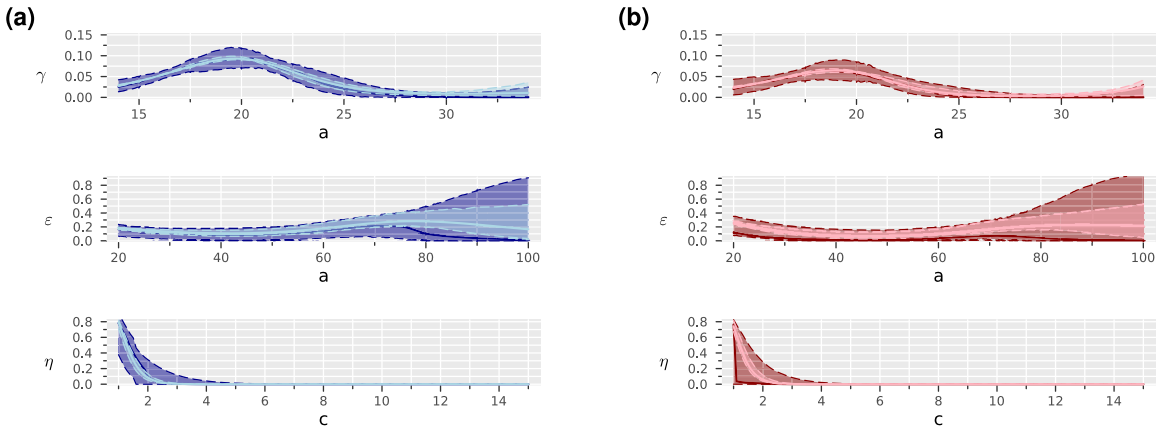

Figure 11: Probabilities of starting ($\gamma(a)$), quitting ($\varepsilon(a)$), and relapsing into smoke ($\eta(c)$) of the smoking habits compartmental (SHC) model estimated via calibration (light color) and approximate Bayesian computation (ABC) (dark color) for males (a) and females (b). Results are reported in terms of point estimate and 90 % confidence bands in the case of calibration and in terms of posterior means and 90 % highest posterior density (HPD) bands in the case of ABC.

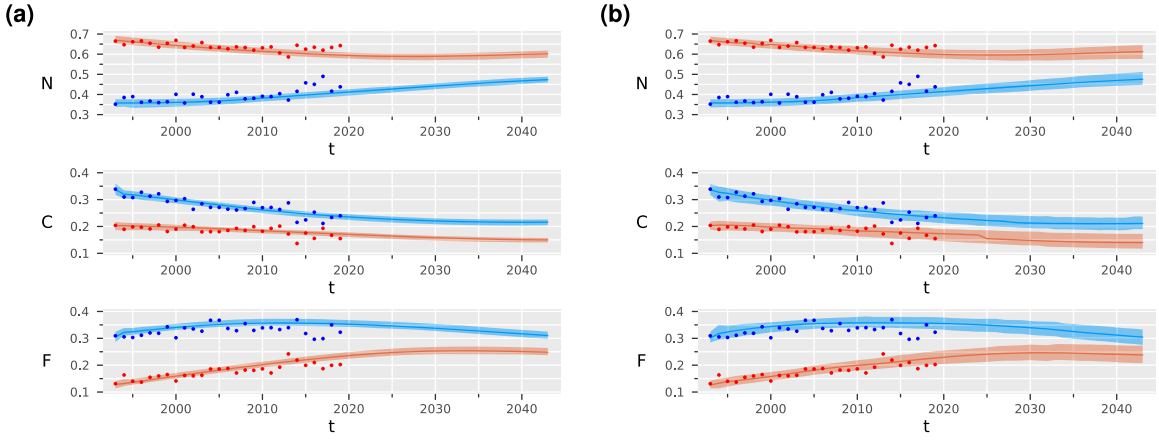

Figure 12: Prevalence of never (N), current (C), and former (F) smokers of the smoking habits compartmental (SHC) model estimated through calibration (a) and approximate Bayesian computation (ABC) (b), for males (blue) and females (pink), with projections up to 2043. Results are reported in terms of point estimate and 90 % confidence bands (a) and in terms of posterior means and 90 % highest posterior density (HPD) bands (b) the estimates are compared with the observed prevalences.

and ABC for both genders. Point-wise confidence bands are obtained by evaluating point-by-point the quantiles of the distributions resulting from computing $\gamma(a)$, $\varepsilon(a)$, and $\eta(c)$ using the bootstrap parameter samples. Point-wise credible bands correspond to the HPD intervals computed point-by-point using samples from the posterior distributions of $\boldsymbol{\theta}$. Solid lines correspond to the values obtained by considering the MLE of θ or the posterior means evaluated point-by-point. Credible bands are wider than confidence bands, and point estimates are almost always similar between the two methods. The curves indicate that males are more likely to start and quit smoking than females and that the probability of starting smoking has a peak around 19 and 20 years of age. The probability of stopping smoking increases after 50 years of age, while the probability of smoking relapse becomes negligible after 2–3 years since cessation. In Figure 12, the estimated prevalences for never, current, and former smokers are reported together with the observed ones. The model fit appears to be adequate, with the predicted values close to the observed ones. The forecasts suggest that the smoking prevalence will decrease in the coming years.

As in the SIR model, the uncertainty around the quantities of interest is greater in the Bayesian framework than in the frequentist one, due to the incorporation of the uncertainty around parameter values in the

predictive distribution of the latent variables. From a comparison between the two methods, it is apparent that ABC avoids problems related to the optimization procedures, such as the dependence on starting values, that could make the estimate unstable. Moreover, it allows the computation of point estimates and the evaluation of all the sources of uncertainty in a single procedure. However, in our implementation, ABC took a longer time with respect to calibration and bootstrap (for further details see *Further details of the algorithm implementations* in Supplementary Material).

Discussion and conclusions

In this work, we described and compared several Bayesian and frequentist methods that can be used for tasks of inference and prediction in compartmental models. This paper aims to discuss compartmental models from a statistical point-of-view, to fill a gap between the statistical literature and the state-of-the-art in applied fields, and to orient practitioners in the formulation of a proper statistical model and the choice of adequate estimation methods. To the best of our knowledge, few review articles on this topic have been published. Tang et al.[13] provided a comprehensive review of frequentist and Bayesian methods. They focused on models for SARS-CoV-2 epidemic data, but most of the considered methods are applied to a deterministic/stochastic model carefully tailored and adapted to fit the requirements of the specific method. For example, Bayesian methods are applied to a state-space formulation of the SIR model, where the likelihood associated with the model differs from that considered in frequentist MLE methods.

Here, we aim to describe a possible way to introduce randomness in a given deterministic model in a realistic manner, formulate the likelihood function, and provide a description and a comparison among estimation methods. Particular attention has been paid to problems that one may encounter in the evaluation of the likelihood function. We recognized the main reasons for its intractability: the case of high dimensional latent variables, sparse or partially observed data, and the case of complex compartmental model. This classification allowed us to identify proper statistical methods for each of the considered cases, and to compare their performance.

We tested the methods at work both on a toy example based on simulated data and a real-world example. From the simulation study, it turned out that all the considered methods can provide point estimates, or posterior distributions, consistent with the "truth". Frequentist methods are all based on optimization strategies that often suffer from problems of strong dependence on starting values and require repeating the procedure several times, each one initialized at different points. This increases the computational cost and sometimes makes infeasible the bootstrap procedure, needed to quantify sampling variability and compute confidence intervals. In this regard, we would like to stress that in the present work, we applied the bootstrap procedure described by Chowell[29], but, to the best of our knowledge, theoretical results on the coverage of the resulting bootstrap intervals are still missing in the statistical literature. As regards Bayesian methods, problems are mainly related to the choice of proposal distributions and to the computational cost of imputing missing data living in high dimensional spaces. DA-MCMC methods often suffer from problems of slow convergence and strong chain autocorrelation.

In this work, a special focus has been placed on the flexibility and potentiality of likelihood-free approaches: calibration and approximate Bayesian computation. They can be implemented whatever the reason for the intractability of the likelihood function, indeed they only require the ability to produce pseudo-data from the deterministic/stochastic model and, in the Bayesian framework, to sample from the prior distribution and evaluate point-wise its density.

In our SIR example, both likelihood-based and likelihood-free methods were feasible. This enabled a comparison between the results. In the Bayesian framework, DA-MCMC uses an analytical form of the likelihood function and iterates a Markov chain whose limiting distribution is the exact posterior distribution. At the same time, ABC methods rely on a simulation-based approach that introduces some sources of approximation. However, from our simulation study, it turned out that likelihood-free methods were able to provide good point estimates and that the approximation of the posterior could be considered negligible, in comparison

with DA-MCMC results. It is also worth noting that ABC methods produce i.i.d. samples from the approximate posterior distribution, thus they are highly and straightforwardly parallelizable compared to DA-MCMC algorithms. Finally, the computational cost of likelihood-free methods is less dependent on the cardinality of the latent variable space since they avoid the imputation of missing data and problems related to the mixing of the chain.

A comparison between ABC and MCMC techniques for Bayesian inference in ODE models was also provided by Alahmadi et al. [54] building upon the work by Toni et al. [55]. There, the authors observe that even if ABC provides some computational advantages, they are minimal compared with MCMC methods. Moreover, they highlighted that the ABC methods that are usually implemented in this framework fail in quantifying the uncertainty since they do not include adequately the sampling variability in the generative model. However, their comparisons are different from ours. First of all, they consider an observation model that simply adds a Gaussian random error to the numerical solution of the ODE system, as in Equation (1). A similar approach is also taken by Toni et al. [55] who illustrate the use of ABC on the SIR model. In such a case, under the strong assumption of independence among errors over time, MCMC methods take advantage of an easy-to-evaluate likelihood function, with the computational burden stemming solely from the resolution of the ODE system. This approach simplifies the MCMC strategy which is a standard MH algorithm, since it does not require missing data imputation. However, as we have noted, more realistic models for the observable quantities are required, such as the one described in Section "Working example: the SIR model", which typically hinder the straightforward implementation of the MH algorithm. The second difference lies in the fact that our ABC algorithm employs a simulator that replicates a data-generative process with the same underlying likelihood function as the one used in the corresponding MCMC method. In contrast, Alahmadi et al. [54] demonstrate that ABC fails to capture variability when pseudo-datasets are generated by solving the underlying system of ODEs – i.e., the likelihood is a point mass concentrated at the solution of the ODE system. In their case, the simulator does not account for sampling variability but is just a complex mathematical function of the parameters – see the discussion about model misspecification in ABC methods by Alahmadi et al.[54].

The real-world example fully highlights the great potential of likelihood-free methods, that can retrieve estimates and forecasts even when dealing with very complex models that prevent the use of whatever likelihood-based method. From a comparison between calibration and ABC, we concluded that the results are coherent with each other.

As a general comment, we can note that the calibration and ABC are very similar in spirit. The main difference between the two methods is that, in the point estimation phase, the calibration uses the deterministic model, while ABC resorts only to the statistical model. The presented ABC strategy allows us to consider two sources of variability in a single procedure: the uncertainty over the parameter space described by prior distributions and the sampling variability reproduced by the simulator. Instead, calibration must be combined with an adequate bootstrap procedure to quantify sampling variability. Another strength of ABC methods, compared with calibration, is that they do not rely on optimization strategies, thus avoiding problems of dependence on the starting values.

ABC is in some sense related to the generalized likelihood uncertainty estimation (GLUE) approach, a very common technique in the hydrological literature that represents one of the first attempts to overcome standard calibration procedures by providing an uncertainty assessment [56]. The connection between the two methods has been discussed by Nott et al.[57], who showed that the GLUE approach can be interpreted as a particular ABC algorithm, known as importance sampling ABC. However, there are some relevant differences between the two approaches. First of all, in GLUE techniques the use of a (uniform) prior distribution over the parameter space must be intended as a way to introduce uncertainty in the estimation rather than a formalization of prior belief in a fully Bayesian spirit. Moreover, GLUE procedures often produce pseudo-data from the deterministic model in a similar way to what was discussed by Alahmadi et al.[54]. Thus, the output of the algorithm is a sample from a distribution that should not be interpreted as a posterior distribution in a strict Bayesian sense, since it includes only the variability induced by the prior distribution. That distribution may be seen as a redefinition of the prior distribution over the subset of parameter values that are coherent with the observed data. It follows that, in the limit of the ABC threshold going to 0, this distribution converges to a point of mass over parameter

values that lead to pseudo-data equal to the observed dataset. Accordingly, GLUE/ABC results converge to those of a calibration procedure and the GLUE/ABC can be interpreted as a "stochastic search" of the optimal parameters. This conclusion suggests that, in a complex model with a high number of parameters, such as the SHC model, GLUE/ABC could help to find global minima by overcoming the difficulties related to optimization algorithms. However, the efficiency in ABC strongly depends on the computational cost of the simulator and the adequacy of the proposal distributions.

The present work has several limitations. First of all, we restricted ourselves to discrete-time models, but models for continuous-time Markov processes can be implemented in this field. In such cases, the formulation of the likelihood function is usually based on the assumption that the waiting time between two events is exponentially distributed. Several algorithms for performing Bayesian inference under this assumption have been proposed [58]. Other possible approaches are based on state space models (e.g., those described by Wang et al.[59]) for which posterior distributions can be inferred by resorting to particle Markov chain Monte Carlo methods [60]. Mckinley et al. [10, 11] compare DA-MCMC with SMC and MCMC methods based on simulation-based approximations of the likelihood function in the continuous-time framework and deal with data as censored – i.e. available only at discrete time intervals. One of the reasons that motivates their choice is that sometimes data can be recorded at different time granularity (e.g. daily, weekly, monthly). We addressed this point by implementing likelihood-free inference procedures based on different summaries of the data, both in the frequentist and Bayesian frameworks. It is also worth noticing that considering data as censored data from a continuous-time process becomes infeasible when, in addition to having only a few observations at specific points in time, the observations corresponding to two or more of the considered compartments are completely missing.

Another limitation of the present work is that we considered simple and easy-to-reproduce sampling methods in the implementation of the ABC method, as well as DA-MCMC method. The state-of-the-art includes more sophisticated gradient-based algorithms such as Metropolis adjusted Langevin algorithm (MALA) [61] or Hamiltonian Monte Carlo methods (HMC). Moreover, likelihood-free algorithms that exploit some approximations of the generative model, or can cleverly orient the simulation procedure, have been proposed [62]. Many of them rely on neural networks or other machine learning approaches such as normalizing flows [63, 64], among others. Finally, this work does not claim to discuss exhaustively all the statistical methods that can be potentially used in the described scenarios, such as methods that overcome the intractability of the likelihood function through the formulation of surrogate models – e.g. Gaussian processes (see [30] for a comprehensive discussion), indirect inference [65] or variational Bayes [66]. Further work should be done to investigate the performance of these methods and to compare them to those considered in the present paper.

Research ethics: Not applicable.
Informed consent: Not applicable.
Author contributions: C.V. and M.B. conceptualized the project; A.L. wrote the codes and conducted the statistical analysis under the statistical supervision of C.V. and the epidemiological supervision of M.B.; C.V. and A.L. wrote the first version of the paper; M.B. provided critical feedback. All the authors read and approved the final version of the paper.
Use of Large Language Models, AI and Machine Learning Tools: None declared.
Conflict of interest: The authors state no conflict of interest.
Research funding: The authors acknowledge the financial support provided by the Attributable Cancer Burden in Tuscany (ACAB) project funded by Regione Toscana and the "Dipartimenti Eccellenti 2018–2022" ministerial funds.
Data availability: Not applicable.

References

1. Broemeling LD. Bayesian analysis of infectious diseases: COVID-19 and beyond. Chapman & Hall/CRC biostatistics series. Boca Raton London New York: CRC Press, Taylor & Francis Group; 2021.
2. Levy DT, Friend K. A simulation model of policies directed at treating tobacco use and dependence. Med Decis Mak 2002;22:6—17.
3. Carreras G, Gallus S, Iannucci L, Gorini G. Estimating the probabilities of making a smoking quit attempt in Italy: stall in smoking cessation levels, 1986-2009. BMC Public Health 2012;12:183.
4. Jourdan N, Neveux T, Potier O, Kanniche M, Wicks J, Nopens I, et al. Compartmental modelling in chemical engineering: a critical review. Chem Eng Sci 2019;210:115196.
5. Booth V, Rinzel J, Kiehn O. Compartmental model of vertebrate motoneurons for Ca^{2+} -dependent spiking and plateau potentials under pharmacological treatment. J Neurophysiol 1997;78:3371—85.
6. Cao L, Zhao H, Wang X, An X. Competitive information propagation considering local-global prevalence on multi-layer interconnected networks. Front Phys 2023;11:1293177.
7. Brauer F, den Driessche PV, Wu J, Allen LJS. Mathematical epidemiology. Berlin: Springer; 2008, 1945.
8. Flaig J, Houy N. Epidemic control using stochastic and deterministic transmission models: performance comparison with and without parameter uncertainties. medRxiv 2022;2022—11. https://doi.org/10.1101/2022.11.12.22282246.
9. Champagne C, Cazelles B. Comparison of stochastic and deterministic frameworks in dengue modelling. Math Biosci 2019;310:1—12.
10. McKinley T, Cook AR, Deardon R. Inference in epidemic models without likelihoods. Int J Biostat 2009;5.
11. McKinley TJ, Ross JV, Deardon R, Cook AR. Simulation-based Bayesian inference for epidemic models. Comput Stat Data Anal 2014;71:434—47.
12. McKinley TJ, Vernon I, Andrianakis I, McCreesh N, Oakley JE, Nsubuga RN, et al. Approximate Bayesian computation and simulation-based inference for complex stochastic epidemic models. Stat Sci 2018;33:4—18.
13. Tang L, Zhou Y, Wang L, Purkayastha S, Zhang L, He J, et al. A review of multi-compartment infectious disease models. Int Stat Rev 2020;88:462—513.
14. Butcher JC. Numerical methods for ordinary differential equations, 3rd ed. Chichester, West Sussex, United Kingdom: Wiley; 2016.
15. Kurtz TG. Solutions of ordinary differential equations as limits of pure jump Markov processes. J Appl Probab 1970;7:49—58.
16. Kurtz TG. Limit theorems for sequences of jump markov processes approximating ordinary differential processes. J Appl Probab 1971;8:344—56.
17. O'Neill PD, Roberts GO. Bayesian inference for partially observed stochastic epidemics. J Roy Stat Soc 1999;162:121—9.
18. Mckendrick AG. Applications of mathematics to medical problems. Proc Edinb Math Soc 1925;44:98—130.
19. Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics. Proc R Soc Lond — Ser A Contain Pap a Math Phys Character 1927;115:700—21.
20. Anderson R. The Kermack-McKendrick epidemic threshold theorem. Bull Math Biol 1991;53:3—32.
21. Allen LJS, Burgin AM. Comparison of deterministic and stochastic SIS and SIR models in discrete time. Math Biosci 2000;163:1—33.
22. Peng L, Yang W, Zhang D, Zhuge C, Hong L. Epidemic analysis of COVID-19 in China by dynamical modeling. medRxiv 2020. https://doi.org/10.1101/2020.02.16.20023465.
23. Schlickeiser R, Kröger M. Analytical modeling of the temporal evolution of epidemics outbreaks accounting for vaccinations. Physics 2021;3:386—426.
24. Canto B, Coll C, Sanchez E. Estimation of parameters in a structured SIR model. Adv Differ Equ 2017;2017:1—13.
25. Baccini M, Cereda G, Viscardi C. The first wave of the SARS-CoV-2 epidemic in tuscany (Italy): a SI2R2D compartmental model with uncertainty evaluation. PLOS One 2021;16:e0250029.
26. Nelder JA, Mead R. A simplex method for function minimization. Comput J 1965;7:308—13.
27. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the *EM* algorithm. J Roy Stat Soc B 1977;39:1—22.
28. Levine RA, Casella G. Implementations of the monte carlo EM algorithm. J Comput Graph Stat 2001;10:422—39.
29. Chowell G. Fitting dynamic models to epidemic outbreaks with quantified uncertainty: a primer for parameter uncertainty, identifiability, and forecasts. Infect Dis Model 2017;2:379—98.
30. Baker E, Barbillon P, Fadikar A, Gramacy RB, Herbei R, Higdon D, et al. Analyzing stochastic computer models: a review with opportunities. Stat Sci 2022;37:64—89.
31. Zucchini W, MacDonald IL, Langrock R. Hidden Markov models for time series: an introduction using R. Second edition, first issued in paperback ed. No. 150 in Monographs on statistics and applied probability. Boca Raton London New York: CRC Press, Taylor & Francis Group; 2021.
32. Efron B, Tibshirani R. An introduction to the bootstrap. Nachdr. ed. No. 57 in Monographs on statistics and applied probability. Boca Raton, Fla: Chapman & Hall; 1998.
33. Metropolis N, Ulam S. The Monte Carlo method. J Am Stat Assoc 1949;44:335—41.
34. Marshall AW. The use of multi-stage sampling schemes in Monte Carlo computations. Rand Corporation 1954.
35. Robert CP, Changye W. All. Markov Chain Monte Carlo Methods, Survey with Some Frequent Misunderstandings. John Wiley & Sons, Ltd; 2021:1—28 pp.

36. Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation. J Am Stat Assoc 1987;82:528—40.
37. Liu JS. The collapsed gibbs sampler in Bayesian computations with applications to a gene regulation problem. J Am Stat Assoc 1994;89:958—66.
38. Robert C, Casella G. Monte Carlo statistical methods. New York: Springer Science & Business Media; 2013.
39. Haario H, Saksman E, Saksman J. An adaptive Metropolis algorithm. Bernoulli 2001;7:223—42.
40. Neal RM. MCMC using Hamiltonian dynamics. In: Handbook of Markov Chain Monte Carlo. Oxfordshire: Taylor & Francis Group; 2011.
41. Afshar HM, Holenstein J. Reflection, refraction, and Hamiltonian Monte Carlo. In: Advances in Neural Information Processing Systems. Red Hook, NY: Curran Associates, Inc; 2015, 28.
42. Andrieu C, Doucet A, Holenstein R. Particle Markov chain Monte Carlo methods. J Roy Stat Soc B Stat Methodol 2010;72:269—342.
43. Rubin DB. Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann Statistics 1984;12.
44. Tavaré S, Balding DJ, Griffiths RC, Donnelly P. Inferring coalescence times from DNA sequence data. Genetics 1997;145:505—18.
45. Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol Biol Evol 1999;16:1791—8.
46. Sisson SA, Fan Y, Beaumont MA, editors. Handbook of approximate Bayesian computation. Boca Raton: CRC Press, Taylor & Francis Group; 2019.
47. Kypraios T, Neal P, Prangle D. A tutorial introduction to Bayesian inference for stochastic epidemic models using approximate Bayesian computation. Math Biosci 2017;287:42—53.
48. Beaumont MA, Cornuet JM, Marin JM, Robert CP. Adaptive approximate Bayesian computation. Biometrika 2009;96:983—90.
49. Lenormand M, Jabot F, Deffuant G. Adaptive approximate Bayesian computation for complex models. Comput Stat 2013;28:2777—96.
50. Mogensen PK, Riseth AN. Optim: a mathematical optimization package for Julia. J Open Source Softw 2018;3:615.
51. Lachi A, Viscardi C, Cereda G, Carreras G, Baccini M. A compartmental model for smoking dynamics in Italy: a pipeline for inference, validation, and forecasting under hypothetical scenarios. BMC Med Res Methodol 2024;24.
52. Thun MJ, Carter BD, Feskanich D, Freedman ND, Prentice R, Lopez AD, et al. 50-Year trends in smoking-related mortality in the United States. N Engl J Med 2013;368:351—64.
53. Hellinger E. Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. J für die Reine Angewandte Math 1909;1909:210—71.
54. Alahmadi AA, Flegg JA, Cochrane DG, Drovandi CC, Keith JM. A comparison of approximate versus exact techniques for Bayesian parameter inference in nonlinear ordinary differential equation models. R Soc Open Sci 2020;7:191315.
55. Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MPH. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J R Soc Interface 2009;6:187—202.
56. Beven K, Binley A. The future of distributed models: model calibration and uncertainty prediction. Hydrol Process 1992;6:279—98.
57. Nott DJ, Marshall L, Brown J. Generalized likelihood uncertainty estimation (GLUE) and approximate Bayesian computation: what's the connection? technical note. Water Resour Res 2012;48.
58. Pooley CM, Bishop SC, Marion G. Using model-based proposals for fast parameter inference on discrete state space, continuous-time Markov processes. J R Soc Interface 2015;12:20150225.
59. Wang L, Zhou Y, He J, Zhu B, Wang F, Tang L, et al. An epidemiological forecast model and software assessing interventions on the COVID-19 epidemic in China. J Data Sci 2020;18:409—32.
60. Akira E, Edwin VL, Marc B. Introduction to particle Markov-chain Monte Carlo for disease dynamics modellers. Epidemics 2019;29:100363.
61. Roberts GO, Stramer O. Langevin diffusions and Metropolis-Hastings algorithms. Methodol Comput Appl Probab 2002;4:337—57.
62. Cranmer K, Brehmer J, Louppe G. The Frontier of simulation-based inference. Proc Natl Acad Sci 2020;117:30055—62.
63. Papamakarios G, Murray I. Fast ε-free inference of simulation models with Bayesian conditional density estimation. Adv Neural Inf Process Syst 2016;29.
64. Prangle D, Viscardi C. Distilling importance sampling for likelihood free inference. J Comput Graph Stat 2023;32:1461—71.
65. Gourieroux C, Monfort A, Renault E. Indirect inference. J Appl Econom 1993;8:S85—118.
66. Tran MN, Nott D, Kohn R. Variational Bayes. Amsterdam: Wiley StatsRef: Statistics Reference Online; 2022:1—9 pp.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/em-2024-0032).