Journal of Theoretical Biology 549 (2022) 111210

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Estimating the basic reproduction number from noisy daily data

Marie-Hélène Descary [a,*], Sorana Froda [a]

[a] Université du Québec à Montréal, Département de mathématiques, Montréal H2X 3Y7, Québec, Canada

ARTICLE INFO

Article history:
Received 11 February 2022
Revised 20 May 2022
Accepted 27 June 2022
Available online 3 July 2022

Keywords:
Basic reproduction number
Estimation
GLM
Non homogeneous Poisson process
SIR and SEIR models

ABSTRACT

In this paper, we propose an easy to implement generalized linear models (GLM) methodology for estimating the basic reproduction number, R_0, a major epidemic parameter for assessing the transmissibility of an infection. Our approach rests on well known qualitative properties of the classical SIR and SEIR systems for large populations. Moreover, we assume that information at the individual network level is not available. In inference we consider non homogeneous Poisson observation processes and mainly concentrate on epidemics that spread through a completely susceptible population. Further, we examine the performance of the estimator under various scenarios of relevance in practice, like partially observed data. We perform a detailed simulation study and illustrate our approach on Covid-19 Canadian data sets. Finally, we present extensions of our methodology and discuss its merits and practical limitations, in particular the challenges in estimating R_0 when mitigation measures are applied.

1. Introduction

During the recent Covid-19 pandemic there was revived interest in finding reliable estimates of the so-called basic reproduction number R_0 that can, among others, give a conservative bound to the proportion of the population that needs to be immunized in order to stop the spread of an epidemic; according to Anderson and May (1992) for $R_0 > 1$ this proportion is $1 - 1/R_0$ in a homogeneous population. In this paper, we consider GLM type methods for estimating R_0 in an epidemic that spreads through a completely susceptible population, a good approximation for the recent Covid-19 pandemic where no previous immunity seemed to be present. A generalization is also proposed, for the case where part of the population is immune at the start of the epidemic. The main purpose is to suggest a core methodology that can be easily applied to data gathered by health agencies while pointing to its limitations.

In what follows, we refer to *compartmental models*, where it is supposed that individuals go through three main stages, once a small number of infectious cases are introduced in a large population: from susceptible (not diseased, not immune) to infected (diseased) to removed (cured and immune or dead). To these main stages one can add various refinements, in particular, one can consider intermediate phases (see Section 2).

In modeling and defining R_0 the most common approaches in the literature are either purely deterministic (Hethcote, 2000) or purely stochastic, in either continuous or discrete time (Kendall, 1956; Isham et al., 2005; Yan, 2008). In spite of the stochastic aspect of any outbreak, in large populations the deterministic models are still good approximations and therefore are widely used to give preliminary estimates of the main epidemic parameters and make predictions. Moreover, it has been proven (Kurtz, 1970; Kurtz, 1971) that, if the population size N tends to infinity, the paths of some stochastic models like the one in Kendall (1956) converge weakly to the solution of the SIR model (1) where the integer values of the stochastic process are translated into proportions. In order to perform data analyses, the deterministic modeling has been enriched by considering stochastic noise, given that we deal with count data. In particular, Poisson noise in the sense of Capaldi et al. (2012) was successfully applied to analyzing COVID-19 data, for instance in Kuniya (2020). Another proposal that relates deterministic epidemic systems to data is the one in Southall et al. (2020) where incidence is modeled as a non homogeneous Poisson process; in their simulation study, O'Driscoll et al. (2021) resorted to the same kind of Poisson model.

For the definition of R_0 in the deterministic context we refer to Heesterbeek and Dietz (1996). In the applied literature, R_0 is described as the average number of infections generated by a single infectious case, sometimes called *index case*, introduced in a completely susceptible population: given the exponential nature of outbreaks, this initial number of infections plays a major role in the *final size* (how many people get infected in the end) of the epidemic.

In this paper we consider the analysis of large populations counts as reported by health agencies. When analyzing such data,

[*] Corresponding author.
E-mail address: descary.marie-helene@uqam.ca (M.-H. Descary).

https://doi.org/10.1016/j.jtbi.2022.111210
0022-5193/© 2022 The Authors. Published by Elsevier Ltd.

an important class of methods is based on the so-called final size formula (Daley and Gani, 2001; Ma and Earn, 2006) that is valid at the end of the epidemic, i.e. once the number of infected is essentially reduced to 0; for the flu this would be the end of a given flu season. This formula relates the total number of infected to the final number of susceptibles, as described in the SIR model (1) but is valid in other instances as well; for a more extensive description see Chowell and Brauer (2009). The estimation method is described for example in Ma and Earn (2006) and has been applied to serological survey data as in Farrington et al. (2001), among others.

Our approach is centered on applying standard generalized linear models (GLM) methodology to data collected by health agencies by proposing to relate observed new cases to the number of passed infections until time t, as expressed in the cumulative number of dead and cured, i.e. the total number of removed until t. This relation is a consequence of a dynamic link between susceptibles and removed, a more flexible approach than the one expressed in the final size formula but maybe somewhat less general, as its application supposes a more restricted class of deterministic systems. Therefore, in the vein of O'Driscoll et al. (2021) or Southall et al. (2020), we take as starting point a deterministic model and introduce count data modeled by non homogeneous Poisson processes, with the novelty that we consider a pair of such processes, corresponding to cases and removals, and we emphasize how their rates are related (see Leduc (2011) and Section 2.3). Thus, in order to estimate R_0, we make use of well known qualitative properties of the underlying SIR or SEIR deterministic systems and introduce the GLM methodology. It is mentioned in Section 2.1 why the SEIR system makes sense in the context of the COVID-19 pandemic; indeed, the classical SEIR system and its various refinements have been extensively used in recent data analyses (see Kemp et al. (2021) and references therein).

Moreover, another major point of our paper is to address various important robustness issues that occur in practice. First, we take into account that data is never collected from time 0 (start of the epidemic). Further, it can happen that cases are assessed only partially and therefore we have to base our estimation on an unknown proportion of actual cases, as a second typical hurdle in this context. Indeed, the preoccupation with underreporting or reporting bias in epidemics, or more generally count data, is not new, and the literature is quite extensive. The issue has been addressed in various ways, depending on the estimation methodology, the type of disease or the available data: for a good overview we refer to the recent paper by Bracher and Held (2021) and references therein; see also Chowell et al. (2007) for specific issues that are common in practice. In this paper, we take a simple approach, described in Sections 2.2 and 2.3: we assume there is a fixed theoretical unknown proportion of cases and corresponding removals, given either the nature of the epidemic (presence of asymptomatic cases, e.g.) or the design study (follow-up of a subpopulation, like hospitalizations, e.g.) while the reported data is produced by binomial thinning. Finally, having a large number of initially immune reduces R_0 (Britton, 2010) and our method estimates correctly this reduced value. A related type of situation, where part of the population does not participate in the spread of the infection although not immune, occurs when mitigation or suppression measures are implemented for longer periods of time.

Finally, it is worth mentioning that, unlike some of the most popular estimation methods cited in O'Driscoll et al. (2021), our approach does not require information at the individual network level and makes no assumptions on additional parameters. On the other hand, our method supposes that both new cases and new removals (or at least new deaths) are regularly followed up (daily or weekly) but when such data are available they are assessed only approximately.

The paper is organized as follows. Section 2 gives the theoretical basis of our approach, namely both the deterministic and stochastic models and their properties, as well as how the GLM estimation methodology can be applied to a typical data set. In Section 3 we make an extensive simulation study and illustrate the method on some Covid-19 data sets. Finally, in Section 4 we discuss various practical issues.

2. Theoretical basis

2.1. SIR and SEIR systems

In what follows, we consider two classical deterministic compartmental models for homogeneous populations: SIR (Susceptibles, Infected and Infectious, and Removed), introduced in M'Kendrick (1925) and Kermack and McKendrick (1927), and SEIR (Susceptibles, Exposed, Infectious, and Removed), as a special case of a more general model MSEIR given in Hethcote (2000).

The systems can be expressed in proportions and we start with the SIR system where we have $(x(t), y(t), z(t))$ standing for proportions at time t, namely: $x(t)$, susceptibles, $y(t)$, infected and infectious, $z(t)$, removed (dead or cured); at time 0, the initial value is $(x_0, y_0, z_0) \in [0,1]^3$ with $x_0 + y_0 \leqslant 1$, and the system is:

$$\begin{cases} x'(t) = -\beta x(t)y(t), \\ y'(t) = \beta x(t)y(t) - \gamma y(t) \;=\; \gamma y(t)\left\{\frac{\beta}{\gamma}x(t) - 1\right\}, \\ z'(t) = \gamma y(t), \end{cases} \tag{1}$$

with $\beta > 0, \gamma > 0$. We define $R_0 = \beta/\gamma$; as far as $x(t), z(t)$ are concerned, $x(t)$ is decreasing and $z(t)$ is increasing, while $-x'(t)$ and $z'(t)$ are unimodal.

Further, consider a population of fixed (during the time of the epidemic) size N, so that the population is divided into three disjoint classes (compartments). Then, the relationships in (1) can be expressed in terms of the compartments' sizes

$$(\tilde{x}(t), \tilde{y}(t), \tilde{z}(t)) = (Nx(t), Ny(t), Nz(t));$$

the initial value is $(\tilde{x}_0, \tilde{y}_0, \tilde{z}_0) = (Nx_0, Ny_0, Nz_0)$ and the equations relating the three classes are identical to those in (1) but of parameters $\beta_N = \beta/N, \gamma_N = \gamma$ with $\beta > 0, \gamma > 0$ as given in (1). In the literature, $\beta_N = \beta/N$ is sometimes designated as contact rate per susceptible (the name of the concept makes sense in the case where $\tilde{x}_0 \approx N$). Further, consider SEIR, where the infected are divided into two subclasses, namely $w(t)$, exposed, and $y(t)$, infectious, and thus the percentage of infected comes to the sum $w(t) + y(t)$. In other words, there is a notable delay (latency) between the moment of exposure and the moment of becoming infectious and infected does not always mean infectious. It seems appropriate to make this type of distinction in the case of an epidemic like Covid-19, but may be less important to make in the case of the flu, where the latency period is much shorter. In the SEIR system, the initial values are $(x_0, w_0, y_0, z_0) \in [0,1]^4, x_0 + y_0 + w_0 \leqslant 1$, and the first and last equations of the SEIR system (A.1) (see Appendix A) are the same as in the case of the SIR system (1); the behaviour of $x(t), z(t)$ is the same as in (1). As for the evolution of infected, their progress is divided into two stages: a susceptible gets infected but is exposed only (not yet infectious) and after a delay (latency time) the exposed becomes infectious. This latency period is described by the parameter σ, with $1/\sigma$ the mean latency time. As in the case of the SIR system, one can translate the equations in the SEIR system (A.1) in terms of sizes of each compartment, $(Nx(t), Nw(t), Ny(t), Nz(t))$; then the parameters are $\beta_N = \beta/N, \gamma_N = \gamma, \sigma_N = \sigma$ with $\beta > 0, \gamma > 0, \sigma > 0$ as in (A.1).

It is worth noting that in models like (1) or (A.1) the epidemic "reproduces itself" in proportion, no matter the size of the population, i.e. if we consider two populations where the outbreak starts from the same initial percentages (x_0, w_0, y_0), respectively (x_0, w_0, y_0, z_0), the timing of events such as the maximum number of new deaths or new removals is the same; also, the percentage of infected after a fixed number of days since the beginning of the epidemic is the same in any population (no matter its size) although, obviously, the absolute numbers differ according to population size.

Although there are versions of (1) or (A.1) that include demographics (i.e. take into account births and deaths that occur during the epidemic) our research concerns only diseases of epidemic type, like seasonal influenza (or a specific wave of Covid-19), that reach eventual extinction in a relatively short time (i.e. $y(t_F) \approx 0$ at some final time t_F); thus, demographics can be ignored in such cases.

To summarize, the positive parameters in the SIR and SEIR systems have the following interpretation (all are supposed unknown in our estimation method):

- β/N: contact (infectivity) rate per susceptible (or per individual);
- $1/\gamma$: mean infectious time (how long one is infectious, on average);
- $1/\sigma$: mean latency time;
- $R_0 = \beta/\gamma$: basic reproduction number.

2.2. Additional properties of the deterministic systems

At the core of our estimation method are some straightforward analytic properties of the deterministic systems (1) and (A.1). Therefore, in this section, we set forth these results, while in Section 2.3 we introduce the probabilistic model and related features of the observed data on which we base our inference. In view of this stochastic development we use statistical terminology to express some deterministic properties. In the case of the SIR system, Bailey (1955) and Daley and Gani (2001) mention an equation that relates $\tilde{x}(t)$ to $\tilde{z}(t)$, and applies in the case of the SEIR system (A.1) as well. This relationship is obtained as follows: consider the first and last equations in (1) and (A.1) translated into compartment sizes, namely

$$\begin{aligned} &\tilde{x}'(t) = -\tfrac{\beta}{N}\tilde{x}(t)\tilde{y}(t) \Longleftrightarrow [\log\{\tilde{x}(t)\}]' = -\tfrac{\beta}{N}\tilde{y}(t);\\ &\tilde{z}'(t) = \gamma\tilde{y}(t), \end{aligned} \tag{2}$$

that imply $[\log\{\tilde{x}(t)\}]' = -(R_0/N)\tilde{z}'(t)$. By integrating from time $s_0 \geqslant 0$ we have the following implicit relationship (system integral):

$$\log\tilde{x}(t) = \log\tilde{x}(s_0) - \frac{R_0}{N}\{\tilde{z}(t) - \tilde{z}(s_0)\}. \tag{3}$$

We start by dealing with the special case $s_0 = 0$ and $\tilde{z}_0 = 0$, while the case $\tilde{z}_0 > 0$ is left to Appendix A.

In practice, one observes new cases that correspond to a reduction in susceptibles; in the deterministic model they are typified by $\tilde{x}(s) - \tilde{x}(t)$ with $s < t$; note that $\tilde{x}(s) > \tilde{x}(t)$. Thus it makes sense to consider the behaviour of the differences $\tilde{x}(s) - \tilde{x}(t), s < t$:

$$\begin{aligned} \tilde{x}(s) - \tilde{x}(t) &= \tilde{x}_0\left[\exp\left\{-\tfrac{R_0}{N}\tilde{z}(s)\right\} - \exp\left\{-\tfrac{R_0}{N}\tilde{z}(t)\right\}\right]\\ &= \tilde{x}_0\exp\left\{-\tfrac{R_0}{N}\tilde{z}(s)\right\}\left(1 - \exp\left[-\tfrac{R_0}{N}\{\tilde{z}(t) - \tilde{z}(s)\}\right]\right). \end{aligned} \tag{4}$$

Further, we approximate the logarithm of a difference $\tilde{x}(s) - \tilde{x}(t), s < t$, by applying $1 - e^{-a} \approx a$ and subsequently letting $\tilde{x}_0 \approx N$:

$$\begin{aligned} \log\{\tilde{x}(s) - \tilde{x}(t)\} &\approx \log(\tilde{x}_0) - \tfrac{R_0}{N}\tilde{z}(s) + \log(R_0) - \log(N) + \log\{\tilde{z}(t) - \tilde{z}(s)\}\\ &\approx \log(R_0) - \tfrac{R_0}{N}\tilde{z}(s) + \log\{\tilde{z}(t) - \tilde{z}(s)\}. \end{aligned} \tag{5}$$

The remainder is negligible because increments $\{\tilde{z}(t) - \tilde{z}(s)\}$ are typically very small (if s, t are close) when compared with N. This last formula is at the core of our estimation method to be described in the next section. Indeed, in order to perform the estimation of R_0, we can view Eq. (5) as the "regression"

$$\log\{\tilde{x}(s) - \tilde{x}(t)\} = a + b\tilde{z}(s) + \text{offset}, \tag{6}$$

where we regress $\log\{\tilde{x}(s) - \tilde{x}(t)\}$ on $\tilde{z}(s)$ and the offset is $\log\{\tilde{z}(t) - \tilde{z}(s)\}$.

Often, in practice (in large populations), one follows up only some proportion p of the data, either by design (we consider only hospitalizations for instance) or because of field realities (there are many asymptomatic cases) and therefore we study how the above relationship (5) is transformed if we relate $p\{\tilde{x}(s) - \tilde{x}(t)\}$ to $p\tilde{z}(s)$ while taking into account $p\{\tilde{z}(s) - \tilde{z}(t)\}$. Of course, if p is assumed known there is no issue; otherwise, we deal with incomplete information. The transformation is given below:

$$\begin{aligned} \log[p\{\tilde{x}(s) - \tilde{x}(t)\}] &= \log\{\tilde{x}(s) - \tilde{x}(t)\} + \log(p)\\ &\approx \log(R_0) - \tfrac{R_0}{N}\tilde{z}(s) + \log\{\tilde{z}(t) - \tilde{z}(s)\} + \log(p)\\ &= \log(R_0) - \tfrac{R_0}{pN}\{p\tilde{z}(s)\} + \log\{p\tilde{z}(t) - p\tilde{z}(s)\}, \end{aligned} \tag{7}$$

which can be viewed as the "regression"

$$\log[p\{\tilde{x}(s) - \tilde{x}(t)\}] = a + b_p\{p\tilde{z}(s)\} + \text{offset}, \tag{8}$$

where the intercept $a = \log(R_0)$ does not change with p, while the slope $b_p = -R_0/(pN) > -R_0/N$ depends on p. As a by-product we obtain a relationship that could be used to estimate p:

$$p = -\frac{R_0}{b_pN} = -\frac{\exp\{\log(R_0)\}}{b_pN} = -\frac{\exp(a)}{b_pN}. \tag{9}$$

Finally, another theoretical issue that is addressed in estimation is the fact that typically an epidemic is not observed from day one, but from some time $s_0 > 0$. Then, $\tilde{x}(s_0) = p^*\tilde{x}_0$ where $p^* \in (0, 1]$ is unknown and (3) implies $\tilde{x}(t) = \tilde{x}(s_0)\exp[-R_0/N\{\tilde{z}(t) - \tilde{z}(s_0)\}]$, which gives the differences

$$\begin{aligned} \tilde{x}(s) - \tilde{x}(t) &= \tilde{x}(s_0)\\ &\times \exp\left\{\frac{R_0}{N}\tilde{z}(s_0)\right\}\exp\left\{-\frac{R_0}{N}\tilde{z}(s)\right\}\left(1 - \exp\left[-\frac{R_0}{N}\{\tilde{z}(t) - \tilde{z}(s)\}\right]\right). \end{aligned} \tag{10}$$

By taking logarithms, we obtain:

$$\begin{aligned} \log\{\tilde{x}(s) - \tilde{x}(t)\} &\approx \log\tilde{x}(s_0) - \tfrac{R_0}{N}\{\tilde{z}(s) - \tilde{z}(s_0)\} + \log(R_0) - \log N\\ &\quad + \log\{\tilde{z}(t) - \tilde{z}(s)\}\\ &= \log\tilde{x}_0 + \log p^* - \tfrac{R_0}{N}\{\tilde{z}(s) - \tilde{z}(s_0)\} + \log(R_0) - \log N\\ &\quad + \log\{\tilde{z}(t) - \tilde{z}(s)\}\\ &\approx \log p^* + \log(R_0) - \tfrac{R_0}{N}\{\tilde{z}(s) - \tilde{z}(s_0)\} + \log\{\tilde{z}(t) - \tilde{z}(s)\}. \end{aligned}$$

Therefore, for the log of the differences, the intercept $a = \log(R_0)$ is replaced with $a^* = \log(R_0) + \log(p^*)$, and we can define $R_0^* = p^*R_0 < R_0$. Thus, the observed data would display a lower intercept, but the same slope, as long as $p = 1$. If both p and p^* are less than one, the relationship expressed in (9) is affected but as long as p^* is close to one the change is negligible:

$$-\frac{\exp(a^*)}{b_pN} = -\frac{\exp\{\log(R_0) + \log(p^*)\}}{b_pN} < -\frac{\exp\{\log(R_0)\}}{b_pN}$$

Given that the slope is affected by the proportion p of followed-up cases and that p^* acts on the intercept but can be expected to be close to 1 in practice (see our examples in Section 3.1), we prefer to base the estimation on the intercept.

2.3. Stochastic approach

Our main analysis tackles the spread of an epidemic in a large population and, therefore, the proposed stochastic versions X_t, Z_t of susceptibles and removed are supposed to revolve "around" the deterministic solutions described in Section 2. The main purpose is: on one hand to have $E(X_t) = \tilde{x}(t), E(Z_t) = \tilde{z}(t)$; on the other hand, to preserve, if possible, one or both Eqs. (3) and (5). First, we note that, in practice, we observe new cases and new removals, which should be realizations of the increments of such processes:

$$(\tilde{x}_0 - X_t) - (\tilde{x}_0 - X_s) = X_s - X_t \quad \text{and} \quad Z_t - Z_s, \;\; s < t.$$

The simplest idea is to follow Leduc (2011) (a model related to the SIR system that makes sense for the SEIR one as well). In such an approach, we postulate two independent non homogeneous Poisson processes of related intensities, corresponding to: (i) the difference between the initial number of susceptibles $\tilde{x}_0$ and susceptibles at a fixed time t and (ii) removed until time t, namely

$$\begin{cases} V_t = \tilde{x}_0 - X_t & \text{such that} \quad E(V_t) = \tilde{x}_0 - \tilde{x}(t), \\ Z_t & \text{such that} \quad E(Z_t) = \tilde{z}(t). \end{cases}$$

In other words, V_t is the cumulative number of cases who have fallen ill until time t (some may still be active, i.e. still infected whether exposed or infectious, some are already removed). Obviously we have:

$$V_{s,t} = V_t - V_s = X_s - X_t \text{ such that } \mu_{V,s,t} = E(V_{s,t}) = \tilde{x}(s) - \tilde{x}(t).$$

In other words, the intensity functions of the processes V_t and $Z_t, t \geqslant 0$ are, respectively:

$$\lambda_V(t) = \frac{\beta}{N}\tilde{x}(t)\tilde{y}(t) \quad \text{and} \quad \lambda_Z(t) = \gamma\tilde{y}(t),$$

and are related by the equation

$$\lambda_V(t) = \frac{R_0}{N}\tilde{x}(t)\lambda_Z(t). \tag{11}$$

This kind of Poisson model is akin to developments in Rizoiu et al. (2018), Southall et al. (2020) or O'Driscoll et al. (2021) but these authors deal only with the process corresponding to new cases, and therefore do not exploit formulas like (11).

Further, our basic idea in estimation is to resort to stochastic equivalents of relation (5), and therefore we focus on the link function:

$$\log \mu_{V,s,t} = \log(R_0) - \frac{R_0}{N}\tilde{z}(s) + \log\{\tilde{z}(t) - \tilde{z}(s)\}, \tag{12}$$

with offset $\log\{\tilde{z}(t) - \tilde{z}(s)\}$. How we make use of this model in estimation is discussed in Section 2.4.

Moreover, in estimation, we propose to consider the case where we observe only a proportion p of the data, i.e., the stochastic equivalent of (7) with $0 < p < 1$. This type of observation corresponds to the thinned processes, $\tilde{Z}_t$ and $\tilde{V}_t$, where:

$$\tilde{Z}_t = \begin{cases} \sum_{j=1}^{Z_t} I_{Z,j}, & Z_t > 0, \\ 0, & \text{otherwise}, \end{cases} \quad \text{and } \tilde{V}_t = \tilde{x}_0 - \tilde{X}_t = \begin{cases} \sum_{j=1}^{V_t} I_{V,j}, & V_t > 0, \\ 0, & \text{otherwise}, \end{cases} \tag{13}$$

with $I_{Z,j}, I_{V,j}, j = 1, 2, \ldots$ i.i.d. Bernoulli variables of success probability p (and independent of Z_t, V_t). Thus, we have: $E\left(\tilde{Z}_t|Z_t\right) = pZ_t, E\left(\tilde{V}_t|V_t\right) = pV_t$. Moreover, by using the technique in Leduc (2011) and Froda and Leduc (2014) that adapt the proof of Theorem 2C of Chapter 4 in Parzen (1962), we conclude that $\tilde{Z}_t, \tilde{V}_t$ are also non homogeneous Poisson processes, of respective intensities $\lambda_{\underset{Z}{\sim}}(t) = p\lambda_Z(t)$, and $\lambda_{\underset{V}{\sim}}(t) = p\lambda_V(t)$. Therefore, $\lambda_{\underset{V}{\sim}}(t), \lambda_{\underset{Z}{\sim}}(t)$ are also related by Eq. (11).

2.4. Estimation

Our main point is that, given the deterministic formula (5) of the SIR (or SEIR) model, if we deal with a large population of known fixed size N we could try to apply straightforward estimation methods of the GLM-type and estimate R_0 by relating observed new cases, $V_t - V_s = X_s - X_t, s < t$ to observed cumulative removed, $Z_s, s \geqslant 0$, while taking into account newly removed, $Z_t - Z_s, s < t$.

In practice, we deal with count data observed at consecutive discretized times, like daily data. If the data follow the stochastic model considered in Section 2.3, then in such consecutive disjoint time intervals, the increments $V_t - V_s, s < t$ and $Z_t - Z_s, s < t$ are independent Poisson variables. On the other hand, the link function defined in (12) depends linearly on the unknown deterministic value $\tilde{z}(s)$. Thus, we have to replace $\tilde{z}(s)$ by its natural estimate, i.e. the observed value z_s, since $E(Z_s) = \tilde{z}(s)$. Therefore, we propose to fit a Poisson regression model (Agresti, 2015; Dobson and Barnett, 2018) to the increments $V_t - V_s, s < t$,

$$\log \mu_{V_{s,t}} = \log(R_0) - \frac{R_0}{N}z_s + \log(z_t - z_s), \tag{14}$$

where $\log(z_t - z_s)$ is an estimate of the offset. Then $b = -R_0/N$ is the slope in (14), while $a = \log(R_0)$ is the intercept.

Let $\hat{a}$ and $\hat{b}$ be the maximum likelihood estimators of a and b. Then, we define our estimator of R_0 as $\hat{R}_0 = \exp(\hat{a})$. A fact to be exploited in estimation is that by factoring out in (4) the exponential in $\tilde{z}(t)$ the approximations (5) and (7) still hold if we replace the regressor $\tilde{z}(s)$ with $\tilde{z}(t)$. Then, it is possible to obtain two estimators of R_0 : $\hat{R}_{0,s}$ by regressing on z_s and $\hat{R}_{0,t}$ by regressing on z_t.

Since the neglected term in (5) is positive in one case and negative in the other it makes sens to use the average $\widehat{R}_0^{(a)} = \left(\hat{R}_{0,s} + \hat{R}_{0,t}\right)/2$; therefore, $\widehat{R}_0^{(a)}$ is our proposed estimator of R_0. Further, if we observe only an unknown proportion p of the data, we preserve the link function but the slope of the linear relationship is altered, as seen in (7); still, the intercept stays unchanged and can be used to estimate R_0 as above. We can estimate p by referring to the relationship (9) and the processes defined in (13). We regress on the observed z_s, estimate R_0 by exponentiating the intercept and this gives $\hat{R}_{0,s}$; the estimator of the slope is $\hat{b}_p$. Then the estimator of p is $\hat{p} = -\hat{R}_{0,s}/\left(\hat{b}_p N\right)$.

3. Numerical results

In this section, we first explore the performance of our proposed method on simulated data sets and then we apply the method to a real data set.

3.1. Simulation study

In order to study the performance of our method in a broad range of setups, we consider various scenarios of transmissibility; all seem plausible in the context of the Covid-19 epidemic. The parameters correspond to the case where time is measured in days.

3.1.1. Parameters

For all scenarios, let $N = 100\ 000$ be the size of the population (the size plays no role in this type of model), and let $y_0 = 0.02\%$ which comes to $\tilde{y}_0 = 20$ initial cases. Moreover we consider a completely susceptible population and let $\tilde{w}_0 = \tilde{z}_0 = 0$. Although the maximum percentages of infected, newly infected and newly removed depend on R_0 and $\tilde{x}_0$ only, the timings of the maxima change according to R_0 and the infectious and latency periods, with a longer epidemic if $1/\gamma$ and $1/\sigma$ are larger. Trying to mimic what was reported in the early literature on Covid-19, in simulation we settle for a mean latency time of $1/\sigma = 4.5$ days and $1/\gamma = 7$ days of average infectivity. Note that, as seen in Section 2.4, the estimation method uses neither the time distributions nor these specific parameter values.

We consider three different values for R_0, namely $R_0 = 1.8, 2.4$ and 3. Note that once R_0 is chosen, one can deduce the value of the parameter β from $\beta = R_0 \cdot \gamma$. With γ and σ given above, Table 1 reports the timing, t_x and t_z, of the maximum number of new cases and new removals respectively and the proportion of removed at time t_z, i.e. $z(t_z)$, for each value of R_0. These quantities are important since they can have an impact on the quality of the estimation of R_0.

Further, we propose to explore to what extent observing only some proportion p of epidemic cases (as mentioned in Section 2.2) influences the results, and take $p = 0.2, 0.5, 0.8$ and 1 in Eq. (7). Moreover, we propose to perform the estimation by taking into account the time window d where the epidemic is observed: we estimate over 45, 60, and 90 days (corresponding roughly to 6, 8 and 12 weeks) that start at day $s_0 = 1, 15$ and 30 of the "true" epidemic, as mentioned in Section 2. So the maximum observation time after the true, typically unknown, beginning of the epidemic would be 120 days. As pointed out in the next section, these times can correspond to different moments in each outbreak, i.e. before the peak, close to the peak, after the peak, where the peak refers to the maximum number of either new cases or new removals. Moreover, note that when $s_0 = 30$ and $R_0 = 1.8, R_0 = 2.4$, or $R_0 = 3$, then (p^*, R_0^*) is equal to, respectively: $(0.9976, 1.796), (0.9937, 2.385)$, and $(0.9849, 2.955)$; we see that p^* is close to 1. With our parameters it makes sense to limit the analysis to the case where the observation time starts at $s_0 \leqslant 30$ but not later since, in a population of size one hundred thousand, we would have, depending on R_0, 240, 630, 1510 cases by day 30, and these are large tallies to pass completely "unnoticed".

In Section 3.1.2 we concentrate on $p = 0.2$ and 1 while leaving to Appendix B various tables that summarize the statistical properties of our estimators.

3.1.2. Results

For each scenario, i.e. each combination of R_0, p, d, and s_0, we performed $n_{\text{sim}} = 500$ simulations yielding n_{sim} estimates, $\widehat{R}_{0,1}^{(a)}, \ldots, \widehat{R}_{0,n_{\text{sim}}}^{(a)}$, of R_0. Boxplots of these estimates for scenarios in which $p = 0.2$ or 1 are given in Fig. 1.

Table 1
Days t_x and t_z where $-x'(t)$, respectively $z'(t)$, are maximum; proportion $z(t_z)$.

R_0	t_x	t_z	$z(t_z)$
1.8	122	133	0.351
2.4	80	90	0.407
3.0	61	71	0.433

Moreover, we calculated two statistics, namely, the mean value $\bar{R}_0$ and the mean relative absolute bias $\overline{B}_{\text{abs}}$ expressed as a percentage, i.e.:

$$\bar{R}_0 = \frac{1}{n_{\text{sim}}}\sum_{i=1}^{n_{\text{sim}}} \widehat{R}_{0,i}^{(a)}, \ \overline{B}_{\text{abs}} = \left(\frac{1}{n_{\text{sim}}}\sum_{i=1}^{n_{\text{sim}}} \frac{|\widehat{R}_{0,i}^{(a)} - R_0|}{R_0}\right) \times 100.$$

These values are reported in Tables B.5 and B.6 of Appendix B.

In the standard case where data is fully observed over a long period of time, i.e. $p = 1, s_0 = 1, d$ large, the performance is very good. Otherwise, under the different constraints considered here, it appears that the performance is best if data are close enough to the time of the respective peak and, in such a case, even $d = 45$ can give very good results. A worthy feature is the good performance in the case $p = 0.2$ where the behaviour is very similar to the case of fully observed data, i.e. $p = 1$. This is of great relevance in practice where we can expect to deal with partially observed data.

Another set of results concerns the coverage probabilities of the estimates, that are smaller than the nominal value; the coverage deteriorates if the observation time interval goes well beyond the peak of $z'(t)$. In part this phenomenon can be explained by the fact that the asymptotic variance in the GLM gets quite small once we use data collected well after the time of the peak of the epidemic, where the percentage $z(t)$ of those already infected and removed becomes quite large. A similar problem concerning the widths of the confidence intervals as one progresses over time has been noticed by O'Driscoll et al. (2021) who applied a different set of estimation methods. Most likely, as soon as one relies on properties of a deterministic system, be it SIR or SEIR, but performs the inference based on noisy data, some form of measurement error is introduced and needs to be corrected. In order to attain the nominal coverage probability, we propose to use confidence intervals based on the jackknife [eq. 12.5] (Tibshirani and Efron, 1993), meaning that the bounds of a confidence interval of level $1 - \alpha$ are $\widehat{R}_0^{(a)} \pm z_{1-\alpha/2}\widehat{se}\left(\widehat{R}_0^{(a)}\right)$, where z_α is the α-quantile of a standard normal variable and $\widehat{se}\left(\widehat{R}_0^{(a)}\right)$ is the jackknife estimate of the standard error of $\widehat{R}_0^{(a)}$ [eq. 11.5] (Tibshirani and Efron, 1993). In Table 2 we consider each scenario and report the coverage probability (in percentage) computed from the $n_{sim} = 500$ confidence intervals obtained by jackknifing where $\alpha = 0.05$. These results are very good in general, with a slight tendency to overcoverage if we resort to data until the end of the epidemic. If desired, in practice one could use only earlier data once the peak has passed.

Finally, we present the performance of our method in estimating the proportion p in the scenarios where $p < 1$; in this section we give the results for $p = 0.2$. For each scenario, we obtain the estimates $\hat{p}_1, \ldots, \hat{p}_{n_{sim}}$ as described at the end of Section 2.4, and assess the performance of our estimator by computing

$$\bar{p} = \frac{1}{n_p}\sum_{i=1}^{n_{sim}} \hat{p}_i \cdot \mathbf{1}(0 \leqslant \hat{p}_i \leqslant 1), \ \text{where} \ n_p = \sum_{i=1}^{n_{sim}} \mathbf{1}(0 \leqslant \hat{p}_i \leqslant 1),$$

and $\mathbf{1}(\cdot)$ is the indicator function. In other words, we compute the performance statistics by considering only the observed values $0 \leqslant \hat{p}_i \leqslant 1, i = 1, \ldots, n_{sim}$ (and discard the other ones). Our experiments suggest that unless $p = 1$ or very close to 1 or 0, the estimates of p usually fall in the right interval $[0, 1]$ and the estimators perform quite well as long as one observes the data long enough. Of course, in practice, one could set at 1 any estimate $\hat{p} > 1$ and reject an estimate $\hat{p} < 0$. Estimating p is rather a by-product than a main objective of this research and given that R_0 is well estimated (by

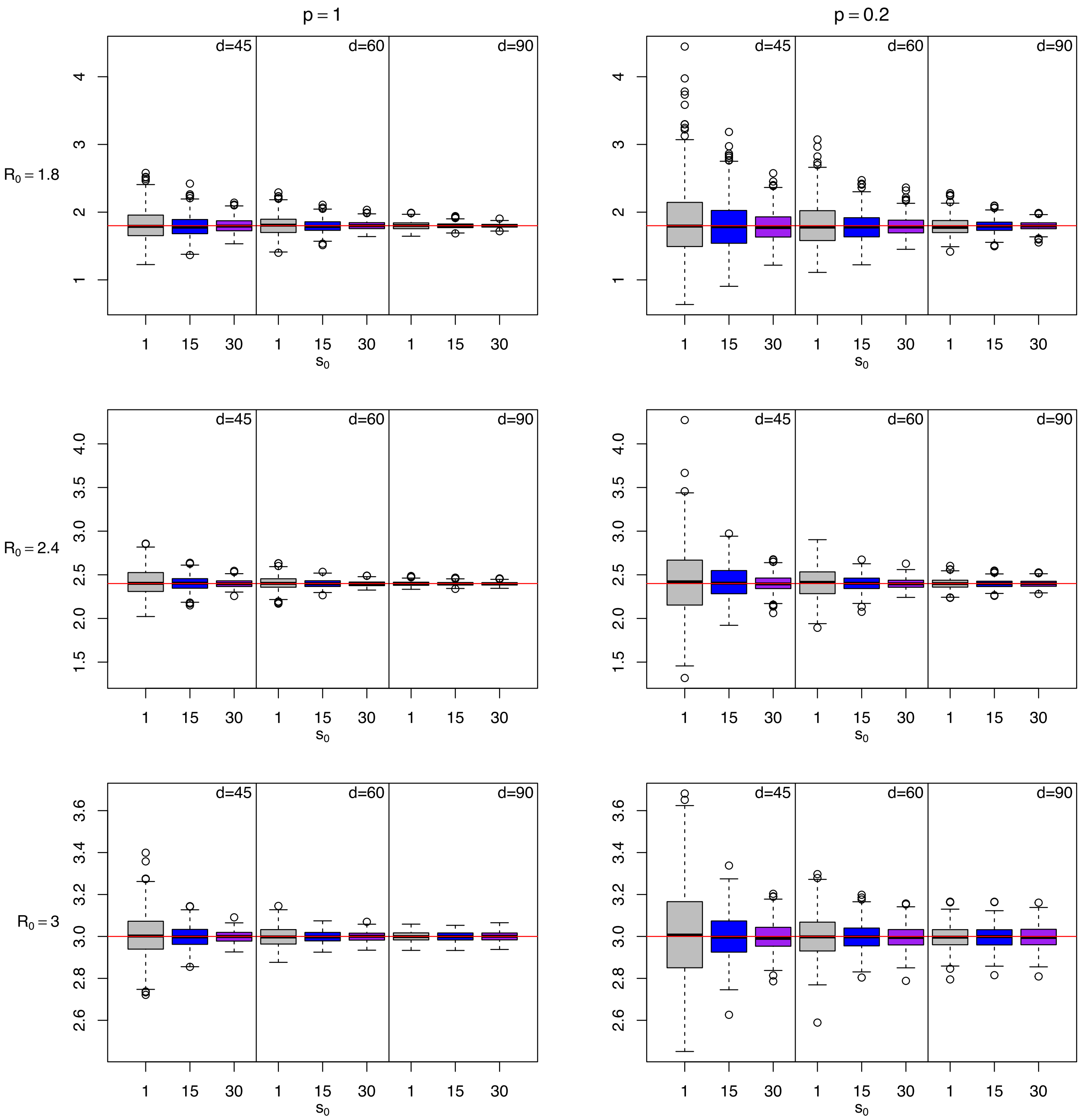

Fig. 1. Boxplots of $\widehat{R}^{(a)}_{0,1}, \ldots, \widehat{R}^{(a)}_{0,500}$: true R_0 is 1.8 (upper panels), 2.4 (middle panels) and 3 (lower panels) and the observed proportion $p = 1$ (left panels) and $p = 0.2$ (right panels). In each panel, we have 9 different combinations of $s_0 = 1, 15$ or 30 and time window $d = 45, 60$ or 90.

exponentiating the intercept) it seems that the issue with estimating p is the estimator of the relatively small slope $-R_0/(pN)$. The values of $\bar{p}$ and n_p/n_{sim} (in percentage) are reported in Table 3.

3.2. Data illustration

To illustrate our methodology we applied it to Canadian Covid-19 data by considering the four provinces with the highest population (Alberta, British Columbia, Ontario, Quebec). In this context, whether in Canada or elsewhere, mitigation measures have been enforced, and one could argue that the "true" R_0 cannot be fully appraised. On the other hand, as explained in Appendix A, in the same spirit as in Britton (2010) we can estimate the reduced value R_0^{**} given by (A.2) that can incorporate mitigation measures as well and thus allows us to compare the success of such mitigation measures. Thus, our analysis is geared towards this comparison.

The analyzed data has the advantage of coming from the same database, namely the Canada Public Health Infobase during the so-called "second wave"; in Canada, the first wave was considered elapsed by the end of June; the data on the second wave was counted since September 2020 (Detsky and Bogoch, 2021). During such a wave there is a clear epidemic behaviour, like in a flu season. In analyzing European data, some authors, e.g. Proverbio et al. (2022), resorted to the recent literature on early warning signals of disease re-emergence, EWS (Southall et al., 2021) in order to better assess the beginning of a new wave, an important practical issue. For our analysis, we proceeded empirically, and considered the time interval from September 1st 2020 till January 31st 2021, a period when the second wave was surely under way. Indeed, in

Table 2
Coverage probability in percentage computed with $n_{sim} = 500$ confidence intervals based on the jackknife; nominal coverage $1 - \alpha = 0.95$.

$R_0 = 1.8$						
d	$s_0 = 1$		$s_0 = 15$		$s_0 = 30$	
	$p = 0.2$	$p = 1$	$p = 0.2$	$p = 1$	$p = 0.2$	$p = 1$
45	94.4	94.2	94.2	94.8	94.8	94.0
60	94.4	95.0	93.4	93.8	94.8	93.0
90	95.4	93.6	94.2	95.4	94.4	96.6
$R_0 = 2.4$						
d	$s_0 = 1$		$s_0 = 15$		$s_0 = 30$	
	$p = 0.2$	$p = 1$	$p = 0.2$	$p = 1$	$p = 0.2$	$p = 1$
45	92.8	95.8	94.2	97.0	97.0	96.4
60	94.8	96.0	97.6	95.6	96.0	97.2
90	96.4	97.0	97.0	96.4	97.0	96.6
$R_0 = 3$						
d	$s_0 = 1$		$s_0 = 15$		$s_0 = 30$	
	$p = 0.2$	$p = 1$	$p = 0.2$	$p = 1$	$p = 0.2$	$p = 1$
45	95.4	96.2	97.8	96.0	96.0	98.0
60	97.4	97.0	97.4	98.2	97.2	98.2
90	97.8	98.6	97.8	98.8	97.4	98.2

Table 3
Values of $\bar{p}$ and n_p/n_{sim} (in percentage) rounded off to three decimals for $n_{\text{sim}} = 500$ and the true value $p = 0.2$.

$R_0 = 1.8$						
d	$s_0 = 1$		$s_0 = 15$		$s_0 = 30$	
	$\bar{p}$	n_p/n_{sim}	$\bar{p}$	n_p/n_{sim}	$\bar{p}$	n_p/n_{sim}
45	0.014	46	0.037	47	0.090	50
60	0.032	48	0.118	53	0.208	68
90	0.214	68	0.239	96	0.200	100
$R_0 = 2.4$						
d	$s_0 = 1$		$s_0 = 15$		$s_0 = 30$	
	$\bar{p}$	n_p/n_{sim}	$\bar{p}$	n_p/n_{sim}	$\bar{p}$	n_p/n_{sim}
45	0.070	53	0.201	72	0.197	100
60	0.206	75	0.197	100	0.195	100
90	0.195	100	0.196	100	0.197	100
$R_0 = 3$						
d	$s_0 = 1$		$s_0 = 15$		$s_0 = 30$	
	$\bar{p}$	n_p/n_{sim}	$\bar{p}$	n_p/n_{sim}	$\bar{p}$	n_p/n_{sim}
45	0.205	75	0.192	100	0.193	100
60	0.192	100	0.193	100	0.194	100
90	0.195	100	0.195	100	0.195	100

principle, our method can be applied to data in any time window. On the other hand, the simulations indicated that the estimation is improved by including data close to the peak, which in this case occurred around the Christmas/holidays season. We stop at the end of January in order to separate the effects of vaccination from those of provincial sanitary measures prior to the vaccination campaign. Although one cannot suppose homogeneous mixing at the provincial level, in each province the dynamics of contacts and infections is driven by what happens in its most populous city, with a large share of cases (Montreal, Toronto, Calgary, and Vancouver); such cities can be considered as having a high level of good mixing.

As for the recorded data, we noticed issues with the reported daily cases and recoveries in this data set, given the many corrections and updates operated over time that introduced artificial variability. For example, there were instances where data accumulated over periods of 2 to 5 days between two dates was reported in the last day of the period while leaving sequences of zeros in-between dates. Besides, recoveries of mild cases were indirectly assessed but in a similar way across provinces. In the present illustrative treatment, in order to deal with such sequences of zero values, we considered the first positive value after such a sequence and redistributed it equally among the preceding days with missing data. The final estimate of R_0 was not greatly affected by this imputation but we report the results with the imputed data in Table 4. Other refinements involving more sophisticated preprocessing of the available data, like some form of smoothing, could

Table 4
Comparison of four Canadian provinces: estimate $\widehat{R}_0^{(a)}$, 95% confidence interval (CI) for R_0 obtained by jackknifing, and coefficient of determination, R_V^2; we use 153 days of data starting September 1st 2020.

Province	$\widehat{R}_0^{(a)}$	CI	R_V^2
Alberta	1.697	$(1.553, 1.840)$	0.51
British Columbia	1.548	$(1.352, 1.744)$	0.23
Ontario	1.349	$(1.281, 1.418)$	0.33
Quebec	1.345	$(1.260, 1.430)$	0.30

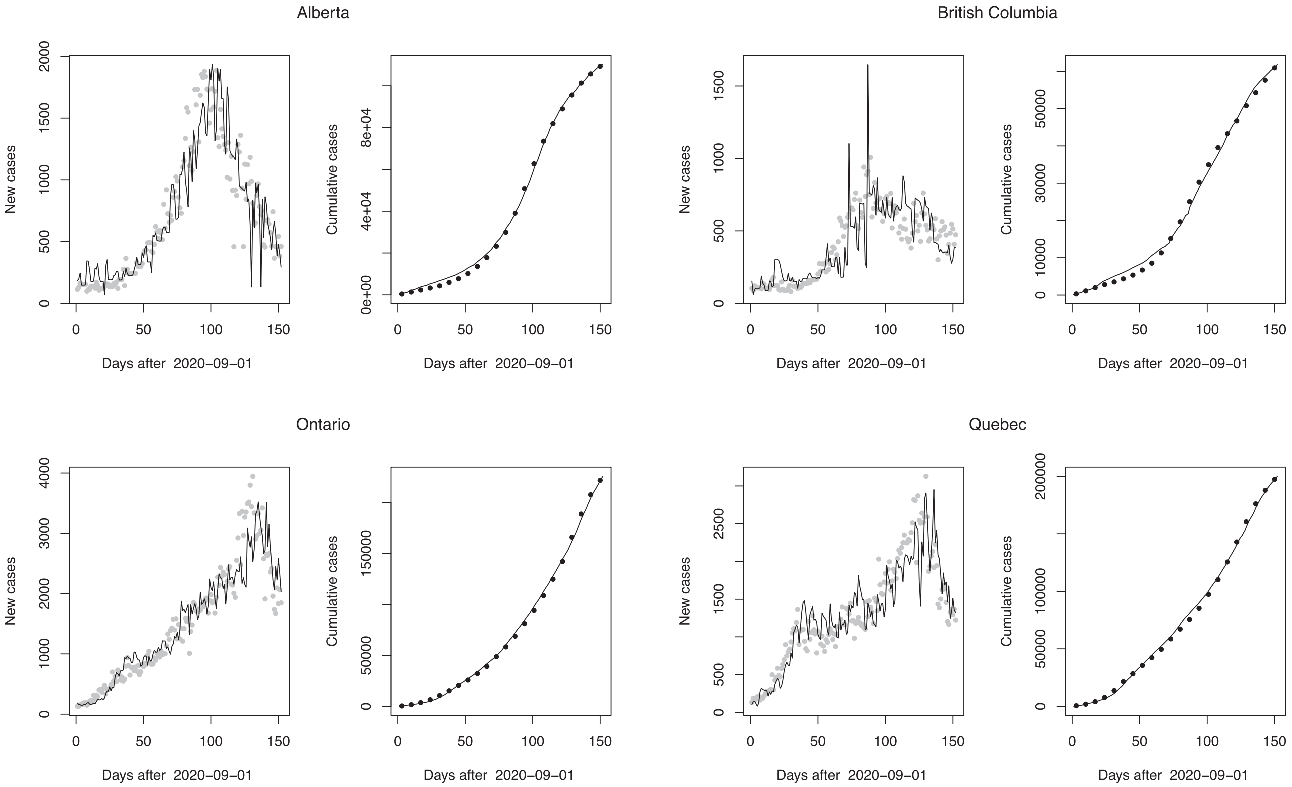

Fig. 2. COVID-19 cases in four Canadian provinces. New daily cases (left panels): fitted (full line) and observed (grey dots). Cumulative cases (right panels): fitted (full line) and observed (black dots), where, for ease of presentation, the dots correspond to weekly data.

prove helpful. In Fig. 2 we present, on one hand, the imputed daily data (new cases, grey dots) and the estimated values obtained from the GLM fitting (curve). On the other hand, we display the cumulative cases: fitted (curve) and observed (dots), that follow closely the estimated curve. Moreover, in Table 4 we report the coefficient of determination R_V^2 as defined in Zhang (2017). Its values range from 0.23 to 0.51 which indicate a satisfactory fit for such noisy data, but also that the fit could be improved by taking into account additional explanatory variables.

We can expect the estimates to express the differences in mitigation measures among provinces. Thus Ontario and Quebec have similar $\widehat{R}_0^{(a)}$ estimates while British Columbia and Alberta have higher values, with the highest $\widehat{R}_0^{(a)}$ for Alberta, which may reflect that some mitigation measures were delayed in these provinces (during the second wave).

4. Discussion and conclusion

In this paper, we propose to estimate the basic reproduction number R_0 by resorting to stochastic equivalents of a known equation relating new cases and cumulative removed (i.e. dead or cured) that characterize some SIR and SEIR deterministic systems. We are using a straightforward GLM approach that is geared towards large populations counts. Our analyses indicate that this relationship remains of interest in observed count data as well; on the other hand, the fit could be improved by taking into account explanatory variables. Moreover, we assess the impact of various forms of missing information, namely: (i) unknown starting time of the epidemic, (ii) partially observed data, and (iii) a large proportion of initially removed. Theoretical appraisal and simulation studies indicate that our estimation methodology works well as far as (i) and (ii) are concerned. In practice, this would allow to base the estimation on a specific part of the population, like hospitalizations. In case (iii), for instance if massive vaccination takes place, given the available data our method recuperates a reduced R_0^{**} value that comes essentially to the effective reproductive rate at time 0 as defined in Anderson and May (1992). In the case where mitigation is enforced one could produce an estimate of R_0 by resorting to the proportionality $R_0^{**} = R_0 N^*/N$ deduced from (A.3), where N^* is the part of the population that participates in the spread of the disease. The unknown $N^* < N$ can be estimated by relying on results like those in Mossong et al. (2008), for instance. Thus, our method seems to be more appropriate for providing comparisons.

This being said, some of the theoretical properties described in Section 2 could impact any estimation method based on models like (1) or (A.1), and some issues we raise in this paper have a larger scope. In particular, given that the timing of the main events and the initial values are unknown to the practitioner one should exercise some caution on basing evaluations or predictions by assuming that one has observed the n-th case, the n-th death, etc, as these may be inaccurate. Indeed, even methods based on initial growth are impacted by the initial proportion of susceptibles in the population, a proportion that is related to vaccination or early mitigation. One merit of our approach is that an equation like (4) is valid during the whole progress of the epidemic.

Another aspect is the assumption of homogeneous mixing. Models like (1) and (A.1) are less realistic in the case of a larger population, but the proposed method for estimating R_0 could be applied to smaller and more homogeneous communities. Since in a given time frame like a specific flu season, for example, an epidemic is generally characterized by a single R_0, one can consider to estimate it from data collected in such smaller communities. How such estimates of R_0 could be refined and combined is a development that goes beyond the scope of the present paper.

Data Availability statement

The data that support the findings of this study are openly available from the Canadian Public Health Infobase at https://health-infobase.canada.ca/ and can be downloaded as a.csv file.

CRediT authorship contribution statement

Marie-Hélène Descary: Methodology, Formal analysis, Writing – review & editing. **Sorana Froda:** Conceptualization, Methodology, Writing – original draft.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors are thankful to Hugues Leduc, René Ferland, and Patrick Fournier for helpful discussions and suggestions. Patrick Fournier contributed a real time formatting of the data set. Two anonymous referees made very insightful comments and suggestions that greatly improved the paper. This research is supported by the National Sciences and Engineering Research Council of Canada (NSERC).

Appendix A. General formulas, case $\tilde{z}_0 > 0$

The SEIR system expressed in percentages can be written as follows:

$$\begin{cases} x'(t) &= -\beta x(t)y(t), \\ w'(t) &= \beta x(t)y(t) - \sigma w(t), \\ y'(t) &= \sigma w(t) - \gamma y(t), \\ z'(t) &= \gamma y(t), \end{cases} \tag{A.1}$$

with $\beta > 0, \gamma > 0, \sigma > 0$. The initial values (x_0, w_0, y_0, z_0) satisfy either $x_0 > 0, y_0 > 0$ and $w_0 = 0, z_0 = 0$ or $x_0 > 0, y_0 > 0, z_0 > 0$ and $w_0 = 0$. The second case expresses the fact that, at the beginning of the epidemic, part of the population is "removed" without experiencing the disease, in other words it is immune (either naturally or by vaccination). For modeling purposes we can include in this framework the mitigation/suppression case where part of the population is not circulating and thus it is "removed" and cannot get infected.

Indeed, we can follow an argument given in Britton (2010) concerning vaccination. Let the true parameters be β, γ; then, in a vaccination program we replace $\tilde{z}_0 = 0$ with a positive $\tilde{z}_0$ and the susceptible population of size $\tilde{x}_0 = N - \tilde{y}_0 - \tilde{w}_0 = N - \tilde{y}_0$ is replaced with a susceptible population of size $\tilde{x}_0 = N - \tilde{z}_0 - \tilde{y}_0 - \tilde{w}_0 = N - \tilde{z}_0 - \tilde{y}_0 < N - \tilde{y}_0$. Further, let $\tilde{y}_0 \approx 0$ and let $N^* = N - \tilde{z}_0$. Then, as long as we can assume that the contact (infection) rate per individual does not change, we have the relationship

$$\frac{\beta}{N} = \frac{\beta^*}{N^*} \Rightarrow \beta^* < \beta.$$

So, the "apparent" basic reproduction number is the value R_0^{**} that satisfies

$$R_0^{**} = \frac{\beta^*}{\gamma} < \frac{\beta}{\gamma} = R_0, \tag{A.2}$$

since the denominator γ remains unchanged. In other words: if $\tilde{z}_0$ individuals are immune at the beginning of the epidemic, an outbreak in a susceptible population of size N and basic reproduction number R_0 behaves like an outbreak in a susceptible population of size $N^* = N - \tilde{z}_0$ and reduced basic reproduction number R_0^{**}. This property is at the basis of any vaccination campaign and, based on field data, any inference method would estimate R_0^{**} rather than R_0. Indeed, the value $R_0^{**} = R_0 N^*/N \approx R_0 x_0$ comes to the effective reproductive rate at time 0, as defined in Anderson and May (1992).

It is commonly accepted (Bjørnstad et al., 2020) that the argument for reducing R_0 used in the case of vaccination can be, approximately, translated to the case of mitigation/suppression while the measures are enforced, as part of the susceptible population is not circulating, is "removed", although not immune, as in the case of vaccination. The two basic reproduction values can be compared by computing their ratio:

$$\frac{R_0}{R_0^{**}} = \frac{\beta}{\beta^*} = \frac{N}{N^*} \approx \frac{N}{N - \tilde{z}_0} = \frac{\tilde{x}_0 + \tilde{z}_0}{\tilde{x}_0} = \frac{1}{p^{**}}, \tag{A.3}$$

Further, we assess the behaviour of the formulas developed in Section 2 in the case where we deal with $(\tilde{x}(t), \tilde{w}(t), \tilde{y}(t), \tilde{z}(t))$ and the sequence of observed times starts at time 0, but $\tilde{z}_0 > 0$. In this case, we refer to (10) with $s_0 = 0$ and $\tilde{z}_0 > 0$ and write

$$\log\{\tilde{x}(s) - \tilde{x}(t)\} \approx \log(R_0) + \log(\tilde{x}_0) - \log N + \log\{\tilde{z}(t) - \tilde{z}(s)\} - \frac{R_0}{N}\{\tilde{z}(s) - \tilde{z}_0\}.$$

Given that in the case where $\tilde{z}_0 > 0$ the available information is not on $\tilde{z}(s)$ but on $\tilde{z}(s) - \tilde{z}_0$, in practice one "regresses" on the difference $\tilde{z}(s) - \tilde{z}_0$. Therefore, the estimation method described in Section 2.4 can give good results as in the case $\tilde{z}_0 = 0$, in the following sense: by neglecting $\tilde{y}_0$ and letting $N \approx N - \tilde{y}_0 = \tilde{x}_0 + \tilde{z}_0$, the exponential of

$$\log R_0 + \log\frac{\tilde{x}_0}{N} \approx \log R_0 + \log\frac{\tilde{x}_0}{\tilde{x}_0 + \tilde{z}_0} = \log R_0 + \log p^{**}$$

is $R_0 p^{**} = R_0^{**}$ and one estimates R_0^{**} as defined in (A.2).

Finally, we have seen in Section 2.2 that the same reduction in R_0 occurs when data is observed not from time 0 but from time $s_0 > 0$, in which case $\tilde{x}(s_0) = p^*\tilde{x}_0$. Further, suppose that $\tilde{z}_0 > 0$ as well, and therefore $\tilde{x}_0 = p^{**}N$. By combining the assumptions $s_0 > 0, z_0 > 0$ we can write:

$$\log\{\tilde{x}(s) - \tilde{x}(t)\} \approx \log(R_0) + \log\tilde{x}(s_0) - \log N - \frac{R_0}{N}\{\tilde{z}(s) - \tilde{z}(s_0)\} + \log\{\tilde{z}(t) - \tilde{z}(s)\},$$

of intercept

$$\begin{aligned} \log(R_0) + \log(p^*\tilde{x}_0) - \log N \approx &\ \log(R_0) + \log(p^*p^{**}N) - \log N \\ = &\ \log(R_0) + \log p^* + \log p^{**}. \end{aligned}$$

Thus the intercept is reduced from $\log(R_0)$ to $\{\log(R_0) + \log p^* + \log p^{**}\}$. As noted above, we can expect the values of p^* to be close to one, i.e. to observe the epidemic early enough. Thus, in practice, the reduction in the estimate of R_0 reflects its reduction due to vaccination or mitigation measures.

Appendix B. Additional numerical results

We start this section by presenting two statistics defined in Section 3.1.2, namely the mean value $\bar{R}_0$ and the mean relative absolute bias $\overline{B}_{\text{abs}}$ expressed as a percentage. For ease of presentation the results on $\bar{R}_0$ are rounded off to three decimals and on $\overline{B}_{\text{abs}}$ to two decimals (or one decimal if the bias is above 10%). In

Table B.5 we consider $p = 0.2$ and 1, while in Table B.6 we take $p = 0.5$ and 0.8.

The rest of the section presents additional results for the cases $p = 0.5$ and $p = 0.8$, in tabular format. The coverage probabilities (in percentage) for confidence intervals based on the jackknife where $\alpha = 0.05$ are reported in Table B.7. Finally, Tables B.8 and B.9 contain the values of $\bar{p}$ and n_p/n_{sim} (in percentage) for $p = 0.5$ and $p = 0.8$ respectively.

Table B.5
Values of $\bar{R}_0$ and $\overline{B}_{\text{abs}}$ (in percentage) for $n_{\text{sim}} = 500$.

$R_0 = 1.8$

	$s_0 = 1$				$s_0 = 15$				$s_0 = 30$			
	$p = 0.2$		$p = 1$		$p = 0.2$		$p = 1$		$p = 0.2$		$p = 1$	
d	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$
45	1.864	23.3	1.813	10.4	1.807	16.4	1.793	7.20	1.792	10.3	1.801	4.83
60	1.817	14.2	1.807	6.67	1.784	9.45	1.795	4.38	1.795	6.37	1.802	2.93
90	1.793	5.89	1.802	2.74	1.796	4.24	1.799	1.86	1.799	2.97	1.800	1.30

$R_0 = 2.4$

	$s_0 = 1$				$s_0 = 15$				$s_0 = 30$			
	$p = 0.2$		$p = 1$		$p = 0.2$		$p = 1$		$p = 0.2$		$p = 1$	
d	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$
45	2.423	12.6	2.416	5.20	2.412	6.48	2.402	2.64	2.404	3.15	2.399	1.53
60	2.412	6.03	2.403	2.49	2.404	3.06	2.399	1.48	2.399	2.05	2.398	0.97
90	2.399	1.99	2.398	0.90	2.398	1.55	2.398	0.73	2.398	1.44	2.398	0.67

$R_0 = 3$

	$s_0 = 1$				$s_0 = 15$				$s_0 = 30$			
	$p = 0.2$		$p = 1$		$p = 0.2$		$p = 1$		$p = 0.2$		$p = 1$	
d	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$
45	3.017	6.23	3.005	2.77	3.001	2.81	2.999	1.36	2.999	1.76	2.999	0.78
60	3.001	2.66	2.998	1.28	2.999	1.69	2.998	0.76	2.998	1.49	2.999	0.64
90	2.999	1.41	3.000	0.63	2.999	1.38	3.000	0.60	2.999	1.44	3.000	0.61

Table B.6
Values of $\bar{R}_0$ and $\overline{B}_{\text{abs}}$ (in percentage) for $n_{\text{sim}} = 500$.

$R_0 = 1.8$

	$s_0 = 1$				$s_0 = 15$				$s_0 = 30$			
	$p = 0.5$		$p = 0.8$		$p = 0.5$		$p = 0.8$		$p = 0.5$		$p = 0.8$	
d	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$
45	1.824	14.0	1.822	11.3	1.798	10.4	1.800	7.85	1.798	6.40	1.801	5.38
60	1.807	9.14	1.807	7.07	1.794	6.08	1.799	4.90	1.797	4.24	1.798	3.35
90	1.798	3.91	1.800	3.06	1.799	2.72	1.800	2.05	1.800	1.94	1.800	1.48

$R_0 = 2.4$

	$s_0 = 1$		$s_0 = 15$		$s_0 = 30$							
	$p = 0.5$	$p = 0.8$	$p = 0.5$	$p = 0.8$	$p = 0.5$	$p = 0.8$						
d	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$
45	2.413	7.57	2.405	5.94	2.402	4.12	2.396	3.24	2.401	2.12	2.398	1.73
60	2.403	3.91	2.398	2.95	2.402	2.08	2.398	1.70	2.399	1.39	2.397	1.07
90	2.399	1.34	2.397	1.02	2.398	1.04	2.397	0.78	2.398	0.96	2.397	0.74

$R_0 = 3$

	$s_0 = 1$				$s_0 = 15$				$s_0 = 30$			
	$p = 0.5$		$p = 0.8$		$p = 0.5$		$p = 0.8$		$p = 0.5$		$p = 0.8$	
d	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$	$\bar{R}_0$	$\overline{B}_{\text{abs}}$
45	3.009	4.24	3.002	3.60	3.004	1.88	3.001	1.57	2.999	1.10	2.998	0.87
60	3.005	1.79	3.003	1.53	2.999	1.06	2.998	0.88	2.999	0.93	2.999	0.71
90	2.999	0.88	2.999	0.71	2.998	0.86	2.998	0.68	2.999	0.88	2.999	0.68

Table B.7
Coverage probability in percentage computed with $n_{sim} = 500$ confidence intervals based on the jackknife; nominal coverage $1 - \alpha = 0.95$.

$R_0 = 1.8$

d	$s_0 = 1$		$s_0 = 15$		$s_0 = 30$	
	$p = 0.5$	$p = 0.8$	$p = 0.5$	$p = 0.8$	$p = 0.5$	$p = 0.8$
45	95.8	96.2	94.4	95.8	95.0	94.2
60	95.6	96.2	93.2	94.8	95.2	94.2
90	93.4	95.6	93.4	94.8	94.0	95.8

$R_0 = 2.4$

d	$s_0 = 1$		$s_0 = 15$		$s_0 = 30$	
	$p = 0.5$	$p = 0.8$	$p = 0.5$	$p = 0.8$	$p = 0.5$	$p = 0.8$
45	95.0	95.2	95.0	93.6	95.4	95.2
60	95.2	94.6	95.8	95.4	96.2	94.6
90	96.6	94.6	96.8	97.0	97.0	97.0

$R_0 = 3$

d	$s_0 = 1$		$s_0 = 15$		$s_0 = 30$	
	$p = 0.5$	$p = 0.8$	$p = 0.5$	$p = 0.8$	$p = 0.5$	$p = 0.8$
45	95.6	92.8	96.4	96.0	97.4	96.4
60	95.4	95.6	97.2	96.2	97.4	97.6
90	98.2	97.4	98.2	97.4	98.2	97.8

Table B.8
Values of $\bar{p}$ and n_p/n_{sim} (in percentage) rounded off to three decimals for $n_{\text{sim}} = 500$ and the true value $p = 0.5$.

$R_0 = 1.8$

d	$s_0 = 1$		$s_0 = 15$		$s_0 = 30$	
	$\bar{p}$	n_p/n_{sim}	$\bar{p}$	n_p/n_{sim}	$\bar{p}$	n_p/n_{sim}
45	0.049	48	0.107	49	0.247	55
60	0.105	53	0.242	54	0.429	67
90	0.448	71	0.515	96	0.493	100

$R_0 = 2.4$

d	$s_0 = 1$		$s_0 = 15$		$s_0 = 30$	
	$\bar{p}$	n_p/n_{sim}	$\bar{p}$	n_p/n_{sim}	$\bar{p}$	n_p/n_{sim}
45	0.174	53	0.447	74	0.486	100
60	0.453	77	0.485	100	0.486	100
90	0.487	100	0.490	100	0.491	100

$R_0 = 3$

d	$s_0 = 1$		$s_0 = 15$		$s_0 = 30$	
	$\bar{p}$	n_p/n_{sim}	$\bar{p}$	n_p/n_{sim}	$\bar{p}$	n_p/n_{sim}
45	0.426	73	0.475	100	0.483	100
60	0.475	100	0.483	100	0.486	100
90	0.486	100	0.487	100	0.487	100

Table B.9
Values of $\bar{p}$ and n_p/n_{sim} (in percentage) rounded off to three decimals for $n_{\text{sim}} = 500$ and the true value $p = 0.8$.

$R_0 = 1.8$						
d	$s_0 = 1$		$s_0 = 15$		$s_0 = 30$	
	$\bar{p}$	n_p/n_{sim}	$\bar{p}$	n_p/n_{sim}	$\bar{p}$	n_p/n_{sim}
45	0.075	47	0.175	48	0.382	54
60	0.186	50	0.382	54	0.613	61
90	0.626	64	0.760	87	0.789	100
$R_0 = 2.4$						
d	$s_0 = 1$		$s_0 = 15$		$s_0 = 30$	
	$\bar{p}$	n_p/n_{sim}	$\bar{p}$	n_p/n_{sim}	$\bar{p}$	n_p/n_{sim}
45	0.264	53	0.628	63	0.773	99
60	0.663	67	0.774	99	0.779	100
90	0.779	100	0.784	100	0.786	100
$R_0 = 3$						
d	$s_0 = 1$		$s_0 = 15$		$s_0 = 30$	
	$\bar{p}$	n_p/n_{sim}	$\bar{p}$	n_p/n_{sim}	$\bar{p}$	n_p/n_{sim}
45	0.608	61	0.761	100	0.773	100
60	0.760	100	0.773	100	0.778	100
90	0.778	100	0.780	100	0.780	100

References

Agresti, A., 2015. Foundations of linear and generalized linear models. John Wiley & Sons.

Anderson, R.M., May, R.M., 1992. Infectious diseases of humans: dynamics and control. Oxford University Press.

Bailey, N.T.J., 1955. Some problems in the statistical analysis of epidemic data. J. R. Stat. Soc. Ser. B (Methodological) 17, 35–68.

Bjørnstad, O.N., Shea, K., Krzywinski, M., Altman, N., 2020. Modeling infectious epidemics. Nat. Methods 17, 455–456.

Bracher, J., Held, L., 2021. A marginal moment matching approach for fitting endemic-epidemic models to underreported disease surveillance counts. Biometrics 77, 1202–1214.

Britton, T., 2010. Stochastic epidemic models: a survey. Math. Biosci. 225, 24–35.

Capaldi, A., Behrend, S., Berman, B., Smith, J., Wright, J., Lloyd, A.L., 2012. Parameter estimation and uncertainty quantification for an epidemic model. Math. Biosci. Eng. 553.

Chowell, G., Brauer, F., 2009. The basic reproduction number of infectious diseases: computation and estimation using compartmental epidemic models. In: Mathematical and statistical estimation approaches in epidemiology. Springer, pp. 1–30.

Chowell, G., Nishiura, H., Bettencourt, L.M., 2007. Comparative estimation of the reproduction number for pandemic influenza from daily case notification data. J. R. Soc. Interface 4, 155–166.

Daley, D.J., Gani, J., 2001. Epidemic modelling: an introduction, 15. Cambridge University Press.

Detsky, A.S., Bogoch, I.I., 2021. Covid-19 in canada: Experience and response to waves 2 and 3. JAMA 326, 1145–1146.

Dobson, A.J., Barnett, A.G., 2018. An introduction to generalized linear models. Chapman and Hall/CRC.

Farrington, C.P., Kanaan, M.N., Gay, N.J., 2001. Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data. J. Roy. Stat. Soc.: Ser. C (Appl. Stat.) 50, 251–292.

Froda, S., Leduc, H., 2014. Estimating the basic reproduction number from surveillance data on past epidemics. Math. Biosci. 256, 89–101.

Heesterbeek, J., Dietz, K., 1996. The concept of ro in epidemic theory. Statistica neerlandica 50, 89–110.

Hethcote, H.W., 2000. The mathematics of infectious diseases. SIAM Rev. 42, 599–653.

Isham, V., Davison, A.C., Dodge, Y., Wermuth, N., 2005. Stochastic models for epidemics. Celebrating Statistics: Papers in honour of Sir David Cox on his 80th birthday, vol. 33. OUP Oxford.

Kemp, F., Proverbio, D., Aalto, A., Mombaerts, L., d'Hérouël, A.F., Husch, A., Ley, C., Goncalves, J., Skupin, A., Magni, S., 2021. Modelling covid-19 dynamics and potential for herd immunity by vaccination in austria, luxembourg and sweden. J. Theor. Biol. 530, 110874.

Kendall, D.G., 1956. Deterministic and stochastic epidemics in closed populations, in: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 4: Contributions to Biology and Problems of Health, University of California Press. pp. 149–165.

Kermack, W.O., McKendrick, A.G., 1927. A contribution to the mathematical theory of epidemics. Proc. R. Soc. London Ser. A 115, 700–721.

Kuniya, T., 2020. Prediction of the epidemic peak of coronavirus disease in japan, 2020. J. Clin. Med. 9, 789.

Kurtz, T.G., 1970. Solutions of ordinary differential equations as limits of pure jump markov processes. J. Appl. Prob. 7, 49–58.

Kurtz, T.G., 1971. Limit theorems for sequences of jump markov processes approximating ordinary differential processes. J. Appl. Prob. 8, 344–356.

Leduc, H., 2011. Estimation de paramètres dans des modèles d'épidémies (Master's thesis). Université du Québec à Montréal.

Ma, J., Earn, D.J., 2006. Generality of the final size formula for an epidemic of a newly invading infectious disease. Bull. Math. Biol. 68, 679–702.

M'Kendrick, A., 1925. Applications of mathematics to medical problems. Proc. Edinburgh Math. Soc. 44, 98–130.

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G.S., Wallinga, J., et al., 2008. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med. 5, e74.

O'Driscoll, M., Harry, C., Donnelly, C.A., Cori, A., Dorigatti, I., 2021. A comparative analysis of statistical methods to estimate the reproduction number in emerging epidemics, with implications for the current coronavirus disease 2019 (covid-19) pandemic. Clin. Infect. Dis. 73, e215–e223.

Parzen, E., 1962. Stochastic processes holden-day. San Francisco 19622.

Proverbio, D., Kemp, F., Magni, S., Gonçalves, J., 2022. Performance of early warning signals for disease re-emergence: A case study on covid-19 data. PLoS Comput. Biol. 18, e1009958.

Rizoiu, M.A., Mishra, S., Kong, Q., Carman, M., Xie, L., 2018. Sir-hawkes: linking epidemic models and hawkes processes to model diffusions in finite populations. In: Proceedings of the 2018 world wide web conference, pp. 419–428.

Southall, E., Brett, T.S., Tildesley, M.J., Dyson, L., 2021. Early warning signals of infectious disease transitions: a review. J. R. Soc. Interface 18, 20210555.

Southall, E., Tildesley, M.J., Dyson, L., 2020. Prospects for detecting early warning signals in discrete event sequence data: Application to epidemiological incidence data. PLoS Comput. Biol. 16, e1007836.

Tibshirani, R.J., Efron, B., 1993. An introduction to the bootstrap. Monographs Stat. Appl. Prob. 57, 1–436.

Yan, P., 2008. Distribution theory, stochastic processes and infectious disease modelling. Mathematical epidemiology. Springer, 229–293.

Zhang, D., 2017. A coefficient of determination for generalized linear models. Am. Stat. 71, 310–316.