The Canadian Journal of Statistics
Vol. 49, No. 4, 2021, Pages 992–1017
La revue canadienne de statistique

Estimated reproduction ratios in the SIR model

Sean ELLIOTT[1*] and Christian GOURIÉROUX[1,2,3]

[1]*Department of Economics, University of Toronto, Toronto, Ontario M5S 2E9, Canada*
[2]*Toulouse School of Economics, Toulouse 31000, France*
[3]*Center for Research in Economics and Statistics, Palaiseau 91764, France*

Key words and phrases: Approximate maximum likelihood; COVID-19; EpiEstim; final size; reproduction ratio; SIR model.

MSC 2020: Primary 62P10; secondary 62M05.

Abstract: The aim of this article is to understand the extreme variability in estimates of the reproduction ratio R_0 observed in practice. For expository purposes, we consider a discrete-time, stochastic version of the susceptible-infected-recovered model and introduce different approximate maximum likelihood estimators of R_0. We carefully discuss the properties of these estimators and illustrate, by a Monte Carlo study, the widths of confidence intervals for R_0. *The Canadian Journal of Statistics* 49: 992–1017; 2021 © 2021 Statistical Society of Canada

Résumé: Le but de cet article est de comprendre l'extrême variabilité du taux de reproduction R_0 observée en pratique. Pour la présentation, nous considérons une version stochastique en temps discret du modèle SIR (Susceptible, Infecté, Guéri) et introduisons différents estimateurs du Maximum de Vraisemblance Approché (MVA) de R_0. Nous analysons en détail les propriétés de ces estimateurs et les illustrons par une étude de Monte-Carlo des largeurs d'intervalles de confiance de R_0. *La revue canadienne de statistique* 49: 992–1017; 2021 © 2021 Société statistique du Canada

1. INTRODUCTION

In the standard epidemiological model, the reproduction ratio—introduced by McDonald (1952)—measures the expected number of persons who are infected by a newly infectious individual. The value of this ratio describes the explosive episode in the early phase of an epidemic, the peak number of infections, as well as an epidemic's final size (Ma & Earn, 2006). It may be estimated daily or weekly as a simple indicator of either an approaching or receding peak (Public Health Ontario [PHO], 2020) and is often used for containment policy. In some cases, R_0 is used to fix the conditions of a partial lockdown or to justify the closing of an international border to foreigners arriving from other countries. "Alert levels are frequently based on this new totemic figure" (Adam, 2020).

The estimated reproduction ratio is a forward-looking notion whose definition involves an expectation that is conditional on both the size of the susceptible population and recovery rates. This is a model-based notion that depends on the information and dynamic model used to evaluate the expectation. The purpose of estimating R_0 is to predict the rate of transmission of an infectious disease. This forward-looking notion must be distinguished from its model-free, retrospective analogue, which simply counts the number of persons infected by a given individual. This

Additional Supporting Information may be found in the online version of this article at the publisher's website.
* *Corresponding author: sean.elliott@mail.utoronto.ca*

difference is similar to the difference between life expectancy and lifetime or between volatility and realized volatility. However, the model-free approach cannot be carried out in the absence of an accurate tracing process and is not immediately useful from a prediction perspective (White & Pagano, 2008).

In practice, estimation of R_0 generates large uncertainty regarding its value (Sanchez & Blauer, 1997; Obedia, Haneef & Boelle, 2012; Cori et al., 2013). For instance, the first R_0 estimates for COVID-19 in Wuhan, China, were between 1.9 and 6.4 (Li et al., 2020; Riou & Althaus, 2020; Sanche et al., 2020; Wu et al., 2020). Estimating R_0 is so important that, "to calculate the official ratio of the United Kingdom, a dedicated government committee reaches consensus on a possible range from ten estimations performed independently" (Adam, 2020). The range in estimates is due to different interpretations and definitions of the ratio in the models that underlie the estimation methods, the estimation methods themselves (see Obedia, Haneef & Boelle, 2012; Cori et al., 2013 for standard estimation packages), and the way the ratio is estimated using rolling calibration windows (Wallinga & Teunis, 2004; Cori et al., 2013). Moreover, estimates are generally provided without confidence bands. These bands can be large, especially in the early phases of an epidemic. Furthermore, estimators can be inconsistent for the reproduction ratio of interest, even if applied to a large population.

The aim of this article is to precisely analyze the uncertainty and lack of robustness of the reproduction ratio estimators. For expository purposes, we focus on the standard susceptible-infected-recovered (SIR) model, initially introduced by Kermack & McKendrick (1927) and widely used in the literature. This model is used to unambiguously define the reproduction ratio. This approach can feasibly be implemented to estimate COVID-19 reproduction ratios using publicly available Canadian surveillance data.

In Section 2, we introduce a discrete-time, stochastic version of the SIR model and discuss the possibility of aggregating the individual infection histories without loss of information. We also rigorously define different notions of the reproduction ratio and how these ratios evolve during an epidemic. Statistical inference for the SIR model is the topic of Section 3. Since the binomial distributions that underlie the SIR model can be approximated by either Poisson or Gaussian distributions, depending on the structure of the population and on transition probabilities, we consider different approximate maximum likelihood estimators of the reproduction ratio. These estimators do not provide the same estimated value, nor do they have the same distribution when we perform estimation in a Gaussian asymptotic framework. They can even be inconsistent in a Poisson asymptotic framework. This leads to Section 4, which contains a Monte Carlo study to find valid confidence intervals for the different estimators and under various designs. We introduce the matrix-variate definition of the reproduction ratio for a SIR model with heterogeneity in Section 5. This leads to the introduction of within- and between-compartment reproduction ratios. Section 6 discusses an alternative definition of the reproduction ratio, called the instantaneous reproduction number, introduced by Fraser (2007), which is based on a renewal equation for the evolution of infected individuals. This notion is the basis of a Bayesian estimation approach to the reproduction ratio, implemented in the EpiEstim R package (Cori et al., 2013). The EpiEstim estimator is usually computed in a rolling way, but ideally should provide reasonable results in the standard SIR model. We discuss precisely why this approach considers a parameter of interest that does not correspond to the initial definition of the reproduction ratio and illustrate this feature by a Monte Carlo study. We also discuss an alternative approach of the same type based on autoregressions of counts of newly infected individuals. We present conclusions in Section 7. The Supplementary Material provides a review of the main properties of the continuous-time, deterministic model and its Euler time discretization. Proofs of some estimation results, additional Monte Carlo results, and a summary of methods currently implemented in popular software packages are also given in the Supplementary Material.

2. MODEL AND OBSERVATIONS

We consider a discrete-time, stochastic version of the SIR model with three states: $S = 1$ (susceptible), $I = 2$ (infected or infectious), and $R = 3$ (recovered, immunized, or removed). We also discuss the aggregation of observations and the notion of the reproduction ratio.

2.1. The Model of Individual Histories

The model specifies the joint distribution of individual medical histories. For each individual i ($i = 1, \dots, n$), and date t ($t = 0, 1, \dots, T$), the variable $Y_{i,t}$ provides the state j ($j = 1, 2, 3$) of individual i at date t.

Assumption 1. *The individual histories* $(Y_{i,t} : t = 0, 1, \dots, T)$ *for* $i = 1, \dots, n$ *are such that*

(i) the variables $Y_{i,t}$, *for* $i = 1, \dots, n$, *are independent conditional on past histories*

$$\underline{Y_{t-1}} = \left\{ \left(Y_{i,t-1}, Y_{i,t-2}, \dots, Y_{i,0} \right) : i = 1, \dots, n \right\};$$

(ii) they have the same transition probability matrix $P_t = (p_{jk}(t))$, *where* $p_{jk}(t)$ *is the probability of migrating from state* j *at date* $t - 1$ *to state* k *at date* t *conditional on the past of the entire population process,* $\underline{Y_{t-1}}$; *and*

(iii) the structure of the transition probability matrix is

$$P_t = \begin{pmatrix} 1 - aN_2(t-1)/n & aN_2(t-1)/n & 0 \\ 0 & 1 - c & c \\ 0 & 0 & 1 \end{pmatrix},$$

where $N_2(t-1)$ *is the number of individuals in state* $I = 2$ *at date* $t - 1$, *and* a *and* c *are parameters such that* $a > 0$ *and* $0 < c < 1$.

Assumption 1 requires that we consider a homogeneous segment of individuals. The case of several segments demands an extension of the SIR model to account for the contagion both between and within segments. This extension is discussed in Section 5. The structure of the transition probability matrix characterizes the SIR model.

(i) The last row of the matrix means that state $R = 3$ is an absorbing state, implying that an individual cannot be infected twice.
(ii) The zero in the second row means that, after infection, an individual recovers, is immunized, and then cannot become at risk.
(iii) The zero in the first row means that an individual cannot recover without being infected first.
(iv) The parameter c is constant and represents the intensity of recovery.
(v) The parameter a characterizes the contagion rate and the intensity of being infected for an individual at risk and is proportional to the proportion of infectious people.

Under Assumption 1, we can deduce the joint distribution of $Y_{i,t}$, with $i = 1, \dots, n$ and $t = 1, \dots, T$, given the initial conditions $Y_{i,0}$ for $i = 1, \dots, n$. Nothing is said about the initial distribution of $Y_{i,0}$, for $i = 1, \dots, n$. This conditional joint distribution is parameterized by a and c, which are assumed to be independent of both n and T. The assumption of a fixed population size is a simplifying assumption for statistical analysis. In some contexts, this size can vary over time. If this is the case, it is necessary to model variations due to births and deaths or due to travel between regions.

 DOI: 10.1002/cjs.11663

TABLE 1: Aggregate counts.

	1	2	3	Total
1	$N_{11}(t)$	$N_{12}(t)$	0	$N_1(t-1)$
2	0	$N_{22}(t)$	$N_{23}(t)$	$N_2(t-1)$
3	0	0	$N_{33}(t)$	$N_3(t-1)$
Total	$N_1(t)$	$N_2(t)$	$N_3(t)$	n

2.2. Aggregated Counts

Under Assumption 1, it is possible to aggregate individual data without losing information on the parameters a and c. We define

- $N_{jk}(t)$, for $j,k = 1,2,3$, as the number of individuals transitioning from state j to k between dates $t-1$ and t;
- $N_j(t)$, for $j = 1,2,3$, as the number of individuals in state j at date t;
- $\hat{p}_{jk}(t) = N_{jk}(t)/N_j(t-1)$ as the sample analogue of $p_{jk}(t)$; and
- $\hat{p}_j(t) = N_j(t)/n$, as the proportion of individuals in state j at date t.

It is known that the set of aggregates $\{N_{jk}(t) : j,k = 1,2,3; t = 1,\dots,T\}$ is a sufficient statistic for the analysis (see Appendix A.2 of the Supplementary Material). Therefore, the analysis can be based on these aggregates only. In the SIR framework, with a constant population size n, these aggregates are related as shown in Table 1. In particular, the following relationships provide the cross-sectional counts in terms of the transition counts: $N_1(t) = N_{11}(t), N_2(t) = N_{12}(t) + N_{22}(t), N_3(t) = N_{23}(t) + N_{33}(t), N_1(t-1) = N_{11}(t) + N_{12}(t), N_2(t-1) = N_{22}(t) + N_{23}(t)$, and $N_3(t-1) = N_{33}(t)$. For the SIR model, these equations can be solved to get the transition counts in terms of the marginal counts. We have that

$$
\begin{aligned}
&N_{11}(t) = N_1(t),\\
&N_{12}(t) = N_1(t-1) - N_1(t) = -\Delta N_1(t),\\
&N_{22}(t) = N_2(t) + \Delta N_1(t),\\
&N_{23}(t) = N_2(t-1) - N_2(t) - \Delta N_1(t) = -\Delta N_1(t) - \Delta N_2(t) = \Delta N_3(t),\\
&\text{and}\\
&N_{33}(t) = N_3(t-1),
\end{aligned}
$$

and are able to deduce the following result:

Proposition 1. *For the SIR model under Assumption 1, the collection of sequences* $N(t) = [N_1(t), N_2(t), N_3(t)]^\top, t = 0,\dots,T$, *is also a sufficient statistic. Moreover, the process* $(N(t))$ *is a homogeneous Markov process.*

Thus, we have the same information in the transition counts and in the cross-sectional counts. That is, the conditional distribution of $(N(t))$ given $\underline{Y_{t-1}}$ is equal to the conditional distribution of $(N(t))$ given $N(t-1)$ only. This property is specific to the SIR model. It is not satisfied in general, for instance, in models with heterogeneity or with more compartments.

2.3. Reproduction Ratio

Numerous summaries of the development of a disease have been introduced in the epidemiological literature. An important concept is the reproduction (or reproductive) ratio (or number). It is defined by computing the expected number of individuals at risk that a newly infected individual will infect during his/her infectious period. In our framework with a constant recovery intensity, the length of the infection/infectious period is stochastic and follows a geometric distribution with the elementary probability mass function $P(X = x) = c(1 - c)^{x-1}$, the survival function $P(X \geq x) = (1 - c)^{x-1}$, and expectation $EX = 1/c$ for $x = 1, 2, \ldots$. As in Farrington & Whitaker (2003), we deduce the expected number of individuals infected by an individual who was infected at date t to be

$$R^*_{0,t} = \frac{a}{n} \sum_{x=1}^{\infty} E_t[N_1(t + x - 1)](1 - c)^{x-1} = \frac{a}{n} \sum_{x=0}^{\infty} E_t[N_1(t + x)](1 - c)^x]. \tag{1}$$

This expectation depends on the transmission rate a and the survival function for the infectious period, but also on the expected proportion of people at risk. For instance, if the population at risk disappears (i.e., $N_1(t) \simeq 0$), then $R_{0,t} = 0$. To adjust for the size of the population at risk and the medical notion of transmission, it is common to also consider

$$R_{0,t} = \frac{a}{N_1(t)} \sum_{x=0}^{\infty} [E_t[N_1(t + x)](1 - c)^x]. \tag{2}$$

The quantities in Equations (1) and (2) are called basic and effective reproduction numbers, respectively. Under Assumption 1, $E_t N_1(t + x) = g(a, c, N_1(t), N_2(t), N_3(t))$ by the homogeneous Markov property, where g is a nonlinear function independent of time. Therefore, $R_{0,t}$ and $R^*_{0,t}$ also depend on time through the marginal counts at time t.

In the literature, this time dependence is often disregarded by focusing on the very early phase (outbreak) of an epidemic (Hethcote, 2000). At $t = 0$, the following assumptions are made:

(i) First, we have that $N_1(0) = n - \varepsilon, N_2(0) = \varepsilon$, and $N_3(0) = 0$, where ε is a very small, positive number. This ε corresponds to the number of initially infected individuals or the size of the first cluster. Without this initial infection, the disease cannot appear in the population. In other words, the SIR model assumes a "closed economy," except at the initial date.

(ii) In the following time, $N_1(t) = n - \varepsilon(t)$, where $\varepsilon(t)$ is also small. An approximate formula for the reproduction ratio is

$$R_{0,0} = R^*_{0,0} \simeq a \sum_{x=0}^{\infty} (1 - c)^x = a/c,$$

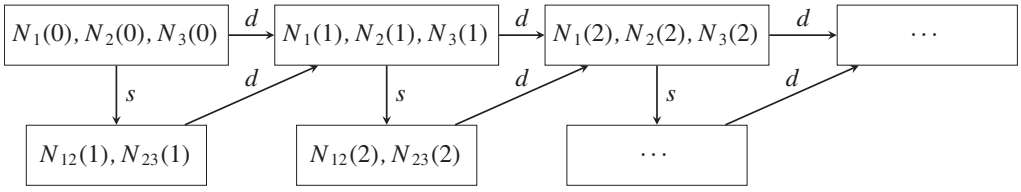
that is, the transmission rate times the expected length of the infection episode. This common value is called the initial reproduction ratio. However, during the epidemic, this measure can differ significantly.

2.4. Simulation

The conditional distributions of the count variables are easily deduced from Assumption 1.

 DOI: 10.1002/cjs.11663

TABLE 2: Simulation scheme.



Proposition 2. *Under Assumption 1,*

(i) $N_{12}(t)$ *and* $N_{23}(t)$ *are independent given the past* $N_{12}(t)$ *follows the binomial distribution* $\mathcal{B}\left(N_1(t-1), a\dfrac{N_2(t-1)}{n}\right)$, $N_{23}(t)$ *follows the binomial distribution* $\mathcal{B}\left(N_2(t-1), c\right)$, *and*

(ii) the process $(N_1(t), N_2(t))$ *is a Markov process, with a conditional distribution obtained from that of* $(N_{12}(t), N_{23}(t))$ *by the change of variables* $N_1(t) = N_1(t-1) - N_{12}(t)$ *and* $N_2(t) = N_2(t-1) + N_{12}(t) - N_{23}(t)$.

That is, under Assumption 1, we obtain a structural dynamic model for the counts in a homogeneous segment, which implies conditional heteroskedasticity and a set of binomial distributions that are specific to the count data. This differs from the model inference approach, which considers a reduced-form regression model with ad hoc errors added to the conditional means.

These results can be used to simulate aggregate counts given parameter values a and c and given starting counts $N_1(0), N_2(0)$, and $N_3(0)$, following the simulation scheme in Table 2, where $\overset{s}{\to}$ denotes a draw from one of the binomial distributions of Proposition 1(i) and $\overset{d}{\to}$ denotes the application of one of the deterministic relations in Proposition 1(ii).

For simulations, and by analogy with COVID-19, we set the parameter values $c = 0.07$, which corresponds to an expected infection period of approximately 14 days and $R_{0,0} = a/c$ between 0.5 and 1.5, which corresponds to a between 0.095 and 0.105. It is worth noting that, in the SIR model, the infected and infectious periods are assumed to be the same, which is not the case with COVID-19. The initial structure of a population corresponding to the city of Toronto, say, can be $n = 3{,}000{,}000$ with a first cluster of $N_2(0) = 50$ (with $N_3(0) = 0$). Thus for $a \simeq 0.1$ at date $t = 0, p_{12}(0) \simeq \dfrac{0.1 \times 50}{3{,}000{,}000} = \dfrac{1}{600{,}000}$. We see that $p_{23}(t)$ and $p_{12}(t)$ are small at the beginning of the epidemic. A simulated path is given in Figure 1, where we observe the standard patterns:

- A decreasing pattern is seen for the size of the population at risk.
- An increasing pattern is seen for the number of immunized people.
- The peak of the epidemic for the number of infected people occurs at around 250 days in this simulation. The figure is given for a rather large number of days to highlight asymptotic behaviour. For this SIR model, there is herd immunity (Allen, 1994). The immunity ratio—that is, the proportion of those who were infected and subsequently recovered—is around 55%.

Let us now explain how we will compute the population basic and effective reproduction numbers corresponding to given parameter values. The corresponding theoretical expressions involve conditional expectations that can be approximated by Monte Carlo simulation. More precisely, at each date t, we simulate and average several future paths $N_1(t+x)$, for $x = 1, \ldots, 30$,

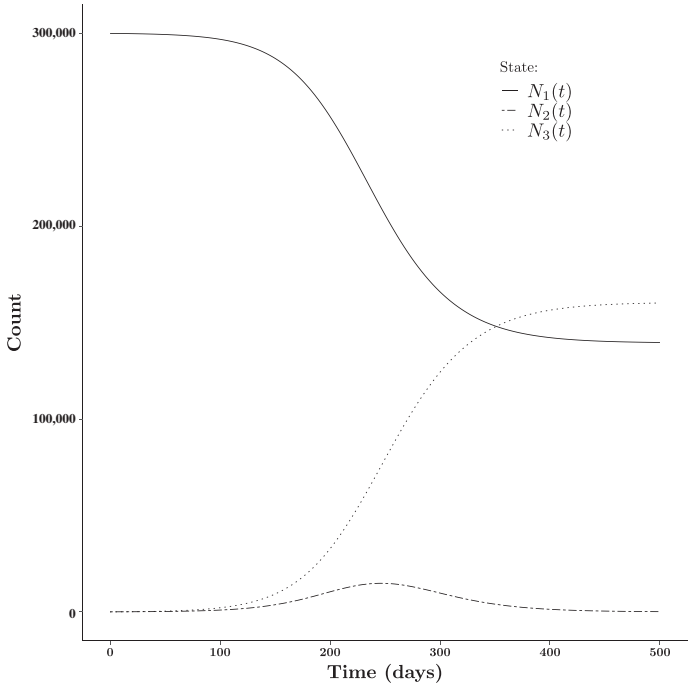


FIGURE 1: Counts for the different states.

to approximate the basic and effective reproduction numbers at t. These paths are reported in Figure 2 with $S = 100$ replications. We observe that even the basic reproduction number, that is, the number adjusted by the size of the population at risk, is not constant during the epidemic in the time-discretized version of SIR. We also observe that the final value of the effective reproduction number is equal to its starting value. Indeed, for large t, the size of the population at risk coincides with the final size of the epidemic and $R_0(\infty) = a/c$. The evolutions shown in Figure 2 are obtained with future paths $N_1(t)$ with lengths of 100 days. In practice, the sum in the definition of R_0 can be truncated by changing the maximum value taken by x: such a truncation can have an impact on the evaluation of R_0. Figure 3 provides the evolutions of reproduction ratios computed with 30-, 60-, and 100-day truncations, respectively.

3. ESTIMATION

This section introduces exact and approximate maximum likelihood approaches and discusses their asymptotic properties. Finite-sample properties are more important in practice and will be considered in Section 4.

3.1. Challenges

The estimation of a SIR model and, more generally, of any epidemiological model, is challenging for three main reasons:

(i) The SIR model is a continuous-time, nonlinear, dynamic model with chaotic properties (see Appendix A.1 of the Supplementary Material). This implies that small changes in the parameter values a and c can have a strong impact on the evolution of the process in the medium and long run. It is known (Allen, 1994) that the deterministic, discrete-time version

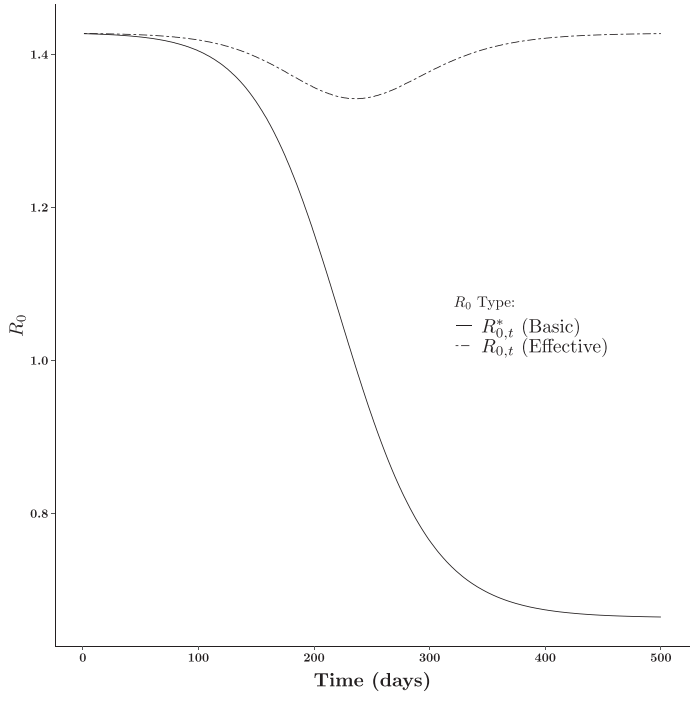


FIGURE 2: Evolution of basic and effective reproduction ratios.

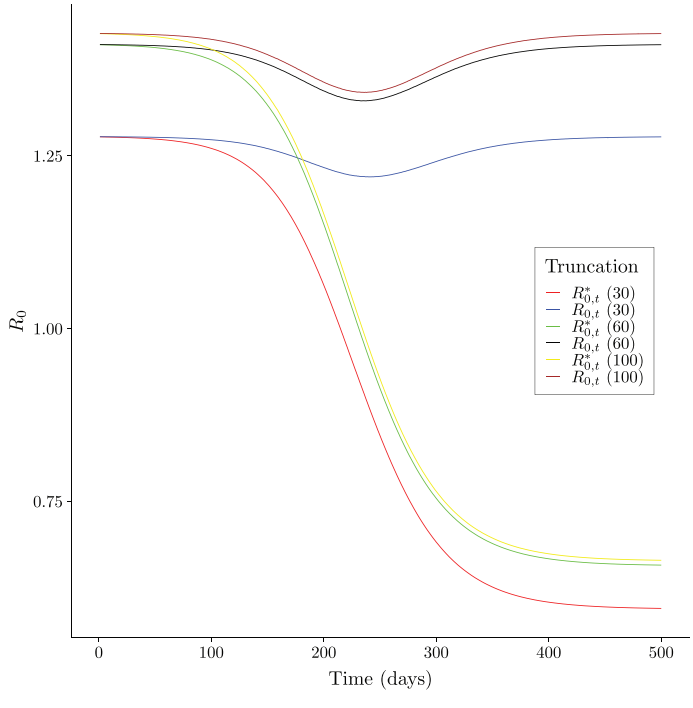


FIGURE 3: Evolution of the reproduction ratio under truncation.

of the SIR model guarantees herd immunity. However, in our stochastic framework, the level of herd immunity and the time at which it is reached are very sensitive to the values of a and c and to the initial conditions.

(ii) The counts of susceptible, infected, and recovered individuals are nonstationary processes, as seen in Figure 1. If $R_{0,0} > 1$, the proportion of infected individuals increases up to a peak and then decreases towards an asymptotic stationary state. This nonstationarity makes it difficult to analyze the properties of the estimators as functions of the number of observation dates T. Moreover, T is usually small, between 20 and 60 days, at the beginning of an epidemic.

(iii) In contrast to the previous point, the population size n is very large and we expect some asymptotic results when n tends to infinity and T is fixed. However, Proposition 2 shows the key role of the binomial distributions $\mathcal{B}\left(N_1(t-1), p_{12}(t)\right)$, and $\mathcal{B}(N_2(t-1), c)$ for $t = 1, \ldots, T$. For an asymptotic analysis, what matters is not just n, but also the marginal counts $N_1(t-1)$ and $N_2(t-1)$. Whereas the susceptible population is often very large, at least at the beginning of an epidemic, the number of infected individuals is much smaller.

However, for large $N_1(t-1)$ and $N_2(t-1)$, we may apply the standard asymptotic results for a binomial distribution, that is, the possibility to approximate it by either a Poisson or Gaussian distribution. For example, if the relevant asymptotic results hold, the Poisson approximation may be preferred because of the fact that it produces closed-form expressions for quantities of interest (see Section 3.3). This approximation of $\mathcal{B}(N, p)$ is either $\mathcal{P}(Np)$ if $N \to \infty$ and $p \to 0$ such that $Np \to \lambda > 0$, where $\mathcal{P}(\lambda)$ denotes the Poisson distribution with parameter λ, or $N(Np, Np(1-p))$ if $N \to \infty$ with p fixed. In our framework, both $p_{23}(t) = c$ and $p_{12}(t)$ are small. The choice between the approximations depends on the magnitudes of $N_1(t-1)p_{12}(t)$ and $N_2(t-1)p_{23}(t)$ for $t = 1, \ldots, T$: that is, the numbers of newly infected and newly recovered individuals, respectively.

For example, if these counts are smaller than 45–50, the Poisson approximation can be used. Otherwise, one may use the Gaussian approximation. But at the beginning and end of an epidemic, $N_{12}(t)$ and $N_{23}(t)$ are rather small. These counts are larger around the peak of the epidemic. Therefore, the approximation will depend on the observation date and also on the size n of the population of interest. For instance, this size is smaller if we want to consider a subpopulation of Toronto, say, males older than 75 (see Zhang et al., 2020 for an analysis restricted to the outbreak on the *Princess Diamond* cruise ship).

3.2. Mechanistic Model

A major part of the literature is based on a deterministic, dynamic model that implicitly assumes the possibility of closely approximating the theoretical transition probabilities using their empirical counterparts, that is, to use the Gaussian approximation. More precisely, under Assumption 1, we have that

$$E_{t-1}\hat{p}(t) = P[\hat{p}_2(t-1)]^\top \hat{p}(t-1).$$

Therefore if $\hat{p}(t)$ is equivalent to $p(t)$, where $p(t)$ is the vector of state occupancy probabilities, we get the following deterministic dynamic model for the $p(t)$s:

$$p(t) = P[p_2(t-1)]^\top p(t-1).$$

This is often called the mechanistic model (see Breto et al., 2009 and Appendix A.1 of the Supplementary Material for its link with the continuous-time SIR model).

3.3. (Approximate) Maximum Likelihood Estimator

In our framework, the log-likelihood function $L(a, c)$ can be decomposed as the sum $L(a, c) = L_1(a) + L_2(c)$. This allows us to separately estimate a and c by focusing on the first and second rows of the (observed) transition matrix, respectively (see Appendix A.2 of the Supplementary Material). Different log-likelihood functions, such as the true one based on the binomial distributions or approximate ones based on either the Poisson or Gaussian approximations, can be considered.

3.3.1. Binomial log-likelihood

Following Proposition 2, we can deduce that

$$L_1(a) = \sum_{t=1}^{T} \left\{ N_{11}(t) \log[1 - a\hat{p}_2(t-1)] + N_{12}(t) \log[a\hat{p}_2(t-1)] \right\}$$

and

$$L_2(c) = \sum_{t=1}^{T} \left\{ N_{22}(t) \log(1-c) + N_{23}(t) \log c \right\}.$$

The maximum likelihood (ML) estimator of a is the solution to the first-order condition

$$-\sum_{t=1}^{T} \left[\frac{N_{11}(t)\hat{p}_2(t-1)}{1 - \hat{a}\hat{p}_2(t-1)} \right] + \frac{1}{\hat{a}} \sum_{t=1}^{T} N_{12}(t) = 0,$$

and has no closed-form expression. The ML estimator of c is

$$\hat{c} = \sum_{t=1}^{T} N_{23}(t) / \sum_{t=1}^{T} N_2(t-1) = \sum_{t=1}^{T} \left\{ \frac{N_2(t-1)}{\sum_{t=1}^{T} N_2(t-1)} \hat{p}_{23}(t) \right\}$$

and is a weighted combination of dated transition frequencies.

3.3.2. Poisson approximate log-likelihood

Here we have that

$$L_1^P(a) \propto \sum_{t=1}^{T} \left\{ N_{12}(t) \log[aN_1(t-1)\hat{p}_2(t-1)] - aN_1(t-1)\hat{p}_2(t-1) \right\}$$

and

$$L_2^P(c) \propto \sum_{t=1}^{T} \left\{ N_{23}(t) \log[N_2(t-1)c] - N_2(t-1)c \right\}.$$

We can obtain Poisson approximate maximum likelihood (AML) estimators with closed-form expressions as

$$\hat{a}_P = n \sum_{t=1}^{T} N_{12}(t) / \sum_{t=1}^{T} [N_1(t-1)N_2(t-1)] \tag{3}$$

and

$$\hat{c}_P = \sum_{t=1}^{T} N_{23}(t) / \sum_{t=1}^{T} N_2(t-1) = \hat{c}.$$

Equation (3) shows that $\hat{a}_P$ is a weighted average of the dated estimated transition coefficients $\hat{a}_t = N_{12}(t)/[N_1(t-1)\hat{p}_2(t-1)]$, with weights proportional to $N_1(t-1)\hat{p}_2(t-1)$. We deduce an analytic formula for the corresponding estimator of the initial reproduction ratio:

$$\hat{R}_{0,P} = \frac{n \sum_{t=1}^{T} N_{12}(t) \sum_{t=1}^{T} N_2(t-1)}{\sum_{t=1}^{T} [N_1(t-1)N_2(t-1)] \sum_{t=1}^{T} N_{12}(t)}. \tag{4}$$

Equation (4) can be used if $\sum_{t=1}^{T} N_{23}(t)$ is nonzero, that is, if recovery has been observed.

3.3.3. Gaussian and unfeasible Gaussian approximate log-likelihood

Using the Gaussian approximation to the binomial distribution, we have that

$$L_1^G(a) \propto -\frac{1}{2} \sum_{t=1}^{T} \log\left(a\hat{p}_2(t-1)[1 - a\hat{p}_2(t-1)]\right) - \frac{1}{2} \sum_{t=1}^{T} N_1(t-1) \frac{[\hat{p}_{12}(t) - a\hat{p}_2(t-1)]^2}{a\hat{p}_2(t-1)[1 - a\hat{p}_2(t-1)]}$$

and

$$L_2^G(c) \propto -\frac{T}{2} \log[c(1-c)] - \frac{1}{2} \sum_{t=1}^{T} N_2(t-1) \frac{[\hat{p}_{23}(t) - c]^2}{c(1-c)}.$$

The unfeasible log-likelihood is obtained by replacing the variance $a\hat{p}_2(t-1)[1 - a\hat{p}_2(t-1)]$ by the estimate $\hat{p}_{12}(t)[1 - \hat{p}_{12}(t)]$, which may be inconsistent when n tends to infinity. We have that

$$L_1^{UG}(a) = -\frac{1}{2} \sum_{t=1}^{T} \left\{ N_1(t-1) \frac{[\hat{p}_{12}(t) - a\hat{p}_2(t-1)]^2}{\hat{p}_{12}(t)[1 - \hat{p}_{12}(t)]} \right\}.$$

From this expression we obtain a closed-form expression for $\hat{a}_{UG}$, which corresponds to an unfeasible, generalized least squares (GLS) estimator of a

$$\hat{a}_{UG} = \sum_{t=1}^{T} (N_1(t-1)\hat{p}_2(t-1)/[1 - \hat{p}_{12}(t)]) / \sum_{t=1}^{T} \left[\frac{N_1(t-1)\hat{p}_2(t-1)^2}{\hat{p}_{12}(t)[1 - \hat{p}_{12}(t)]} \right].$$

3.3.4. Poisson/Gaussian approximate log-likelihood

When n is large, p is small, and np is large, the Poisson distribution $\mathcal{P}(np)$ can be approximated by the Gaussian distribution $N(np, np)$. Thus, compared to the approximations in Section 3.3.3, the p^2 term in the variance is disregarded. In this approach

$$L_1^{PG}(a) \propto -\frac{1}{2} \sum_{t=1}^{T} \log[a\hat{p}_2(t-1)] - \frac{1}{2} \sum_{t=1}^{T} \left\{ N_1(t-1) \frac{[\hat{p}_{12}(t) - a\hat{p}_2(t-1)]^2}{a\hat{p}_2(t-1)} \right\}$$

and

$$L_2^{PG}(c) \propto -\frac{1}{2}T\log c - \frac{1}{2}\sum_{t=1}^{T}\left\{N_2(t-1)\frac{[\hat{p}_{23}(t)-c]^2}{c}\right\}.$$

The AML estimates are positive solutions of polynomial equations of degree two, given by

$$\frac{1}{T}\sum_{t=1}^{T}\left\{N_1(t-1)\hat{p}_2(t-1)\right\}a^2 + a - \frac{1}{T}\sum_{t=1}^{T}\left\{N_1(t-1)\hat{p}_{12}(t)\right\} = 0$$

and

$$\frac{1}{T}\sum_{t=1}^{T}N_2(t-1)nc^2 + c - \frac{1}{T}\sum_{t=1}^{T}\left\{N_2(t-1)\hat{p}_{23}(t)\right\} = 0.$$

To summarize, there are as many AML estimators of a, c, and the initial reproduction number $R_{0,0} = a/c$ as there are (approximate) log-likelihoods. This can explain the different approximations of $R_{0,0}$ published for the same series of aggregate counts.

3.4. Properties of the AML Estimators

Properties of the AML estimators can be derived by Monte Carlo simulations, as shown in Section 4. Their asymptotic properties depend on either Poisson or Gaussian asymptotics, depending on which is the most appropriate, and on the selected estimators. For instance, we may have chosen a Poisson AML estimator when the Gaussian asymptotic conditions were satisfied. In this case, $\mathcal{B}(N,p)$, which is well approximated by $N(Np, Np(1-p))$, has been replaced by $\mathcal{P}(Np)$, which is close to $N(Np, Np)$. Therefore, we have not used the right approximation and have neglected the p^2 term. Recall that, as outlined in Section 3.1, the appropriate choice of approximation relies on whether $p \to 0$ or p is fixed. For illustration, we consider

(i) the behaviour of the Poisson AML estimator $\hat{a}_P$ in the case where Poisson asymptotics are applicable and
(ii) the behaviour of the binomial ML estimator $\hat{a}$ in the case where Gaussian asymptotics are applicable.

3.4.1. Poisson AML and Poisson asymptotics

Let us consider the case where $T = 1$, that is, with two observations of the aggregates. The main results below are valid for any finite T. We have that $\hat{a}_P = nN_{12}(1)/N_1(0)N_2(0)$, $\hat{c}_P = N_{23}(1)/N_2(0)$, and $\hat{R}_{0,P} = \hat{a}_P/\hat{c}_P = \dfrac{N_{12}(1)}{N_{23}(1)}\dfrac{n}{N_1(0)}$. Conditional on $(N_1(0), N_2(0))$, the estimators $\hat{a}_P$ and $\hat{c}_P$ are independent such that $\dfrac{N_1(0)N_2(0)}{n}\hat{a}_P \sim \mathcal{P}\left(a\dfrac{N_1(0)N_2(0)}{n}\right)$, and $N_2(0)\hat{c}_P \sim \mathcal{P}(cN_2(0))$. We deduce that $E_0\hat{a}_P = a$ and $E_0\hat{c}_P = c$, which shows that the Poisson AML estimators are unbiased for $T = 1$. Their variances are $V_0\hat{a}_P = \dfrac{an}{N_1(0)N_2(0)}$ and $V_0\hat{c}_P = \dfrac{c}{N_2(0)}$. In practice, $N_1(0)$ (which is approximately n) and $N_2(0)$ are too small (<30 or 40, say) for Poisson asymptotics to be valid. Therefore, both $V_0(\hat{a}_P)$ and $V_0(\hat{c}_P)$ are not small, even for large n, and we cannot expect $\hat{a}_P$ and $\hat{c}_P$ to be consistent for large n under Poisson asymptotics. Moreover, at the very beginning of an epidemic, infected individuals have not yet recovered, meaning that $N_{23}(1) = 0$. We deduce that $\hat{R}_{0P} = \hat{a}_P/\hat{c}_P = \hat{a}_P/0 = \infty$. This illustrates the lack of accuracy of the basic reproduction ratio during the initial phase of an outbreak.

Remark 1. The unbiasedness property is specific to the case where $T = 1$. If $T = 2$, we have, by iterated expectation, that

$$\hat{a}_P = \frac{n[N_{12}(1) + N_{12}(2)]}{N_1(0)N_2(0) + N_1(1)N_2(1)}, \text{ and } E_0(\hat{a}_P) = nE_0\left[\frac{N_{12}(1) + aN_1(1)N_2(1)}{N_1(0)N_2(0) + N_1(1)N_2(1)}\right],$$

where the expectation is a complicated nonlinear function of the counts $N_1(1)$, $N_2(1)$, and $N_{12}(1)$.

3.4.2. Binomial ML and Gaussian asymptotics

When the law of large numbers and the central limit theorem are applicable, the sample proportions tend to their theoretical counterparts: $\hat{p}_{jk}(t) \to p_{jk}(t)$ and $\hat{p}_j(t) \to p_j(t)$, for $j, k = 1, 2, 3$ as n tends to infinity. The ML estimators tend to the true parameter values: $\hat{a} \to a$, $\hat{c} \to c$, and $\hat{R}_0 = \hat{a}/\hat{c} \to a/c$ at the speed $1/\sqrt{n}$. Both $\hat{a}$ and $\hat{c}$ are asymptotically independent and asymptotically normal with variances consistently estimated by

$$\hat{V}(\hat{a}) = \left\{\sum_{t=1}^{T}\left(\frac{N_{11}(t)\hat{p}_2(t-1)}{[1-\hat{a}\hat{p}_2(t-1)]^2}\right) + \frac{1}{\hat{a}^2}\sum_{t=1}^{T}N_{12}(t)\right\}^{-1} \text{ and } \hat{V}(\hat{c}) = \frac{\hat{c}(1-\hat{c})}{\sum_{t=1}^{T}N_2(t-1)},$$

respectively.

4. MONTE CARLO STUDY

Even when Gaussian asymptotics can be used, real datasets are finite: a key issue is to know whether the asymptotic results are accurate in determining confidence intervals for the parameters a, c, and R_0. In this section, we perform a Monte Carlo analysis for some of the estimators introduced in Section 3. We fix the design as follows: $N_1(0) = 3{,}000{,}000$; $N_2(0) = 100$, 1000; $T = 20$; $c = 0.07$; and $R_0 = 2$. This design corresponds to estimators computed on the period $[0, T]$. Note that the process of marginal counts is Markov. Therefore, this simulation exercise also applies to a rolling estimator computed on $(t, t + T)$, where the marginal counts at t are the counts fixed for $N_1(0)$ and $N_2(0)$. This explains why we allow a large value of $N_2(0)$ in the design. Figures 4 and 5 correspond to the parameters estimated by the approximated Poisson likelihood with $N_2(0) = 100, 1000$, respectively. The figures provide the finite-sample distributions of the parameters a, c, and $R_0 = a/c$.

Whereas some skewness can be observed in the estimated contagion parameter distribution in Figure 5, this feature largely disappears for the estimated reproduction number. This is due to the nonlinear transformation used to compute R_0 and the dependence between $\hat{a}$ and $\hat{c}$. The initial number of infected individuals also has an impact on the width of the estimated distribution of R_0 which is known at $\pm 20\%$ for $N_2(0) = 100$, at $\pm 10\%$ for $N_2(0) = 1000$.

To have more insight into the finite-sample properties of these estimators, we provide summary statistics for different designs (a, c), $N_2(0)$, and T in Tables 3–5. Finite-sample distributions for the estimators computed using the unfeasible Gaussian approximate likelihood are given in Appendix A.3 of the Supplementary Material.

5. REPRODUCTION NUMBER UNDER HETEROGENEITY

5.1. Model with Heterogeneity

Another source of variability in estimating R_0 is due to latent heterogeneity and concerns the definition of R_0 itself. For illustration, we consider a situation with two homogeneous populations, population 1 and population 2. The SIR model is replaced by a $(\text{SIR})^2$ model with

 DOI: 10.1002/cjs.11663

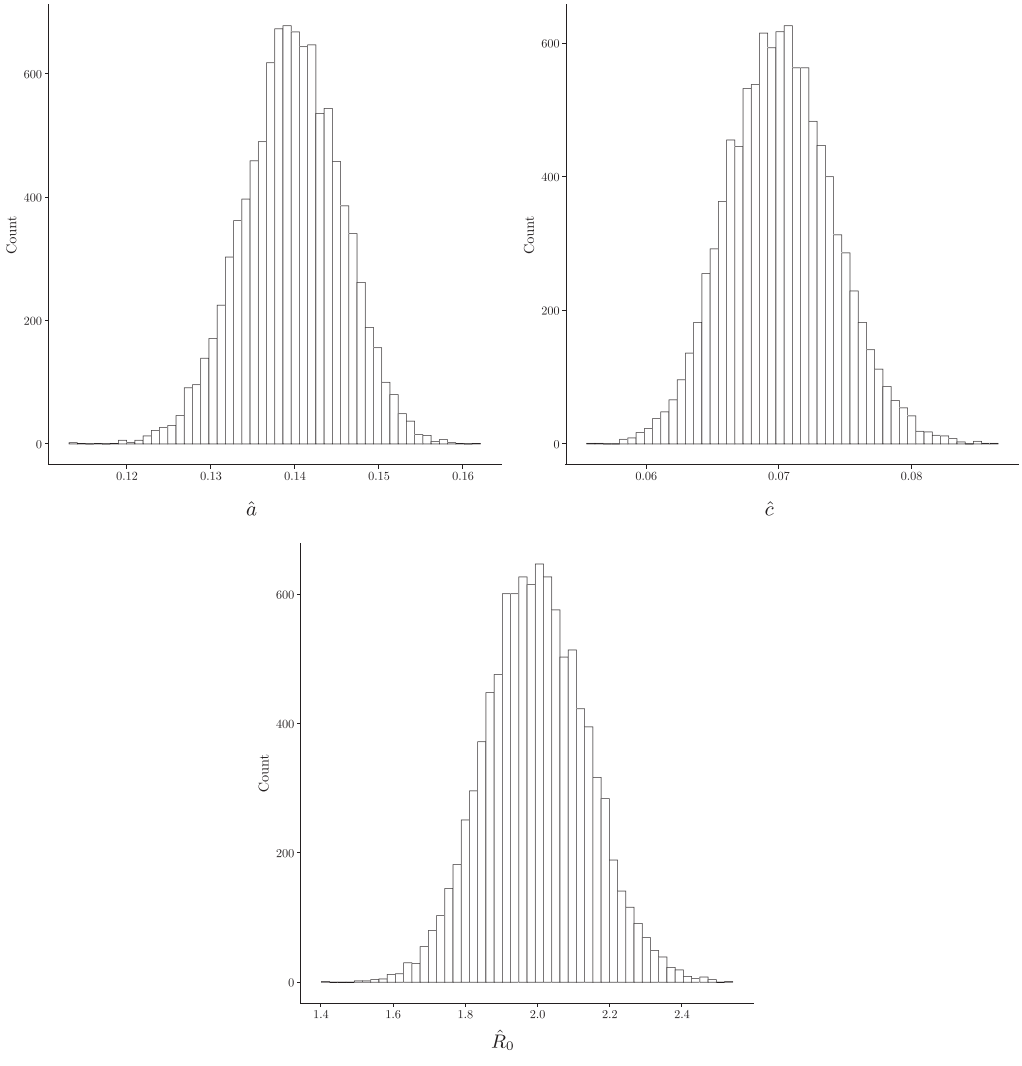


FIGURE 4: Distributions of approximate Poisson ML estimators under $N_2(0) = 100$.

six states: $S_1I_1R_1S_2I_2R_2$, in the terminology of Appendix 1 of Gourieroux & Jasiak (2020a). The (6×6) transition matrix is block diagonal with the jth diagonal block

$$P_{j,t} = \begin{pmatrix} 1 - a_{j1}\dfrac{N_2^1(t-1)}{N^1} - a_{j2}\dfrac{N_2^2(t-1)}{N^2} & a_{j1}\dfrac{N_2^1(t-1)}{N^1} + a_{j2}\dfrac{N_2^2(t-1)}{N^2} & 0 \\ 0 & 1 - c_j & c_j \\ 0 & 0 & 1 \end{pmatrix},$$

for $j = 1, 2$, where $N_2^j(t)$ (respectively, n^j) is the number of infected people in population j (respectively, the size of population j). Typically, the two populations can correspond to two age categories, say, young and old. Now, the contagion parameter has a matrix form $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$. Indeed, there are contagions within each population, described by a_{11} and a_{22}, and between the populations described by a_{12} and a_{21}.

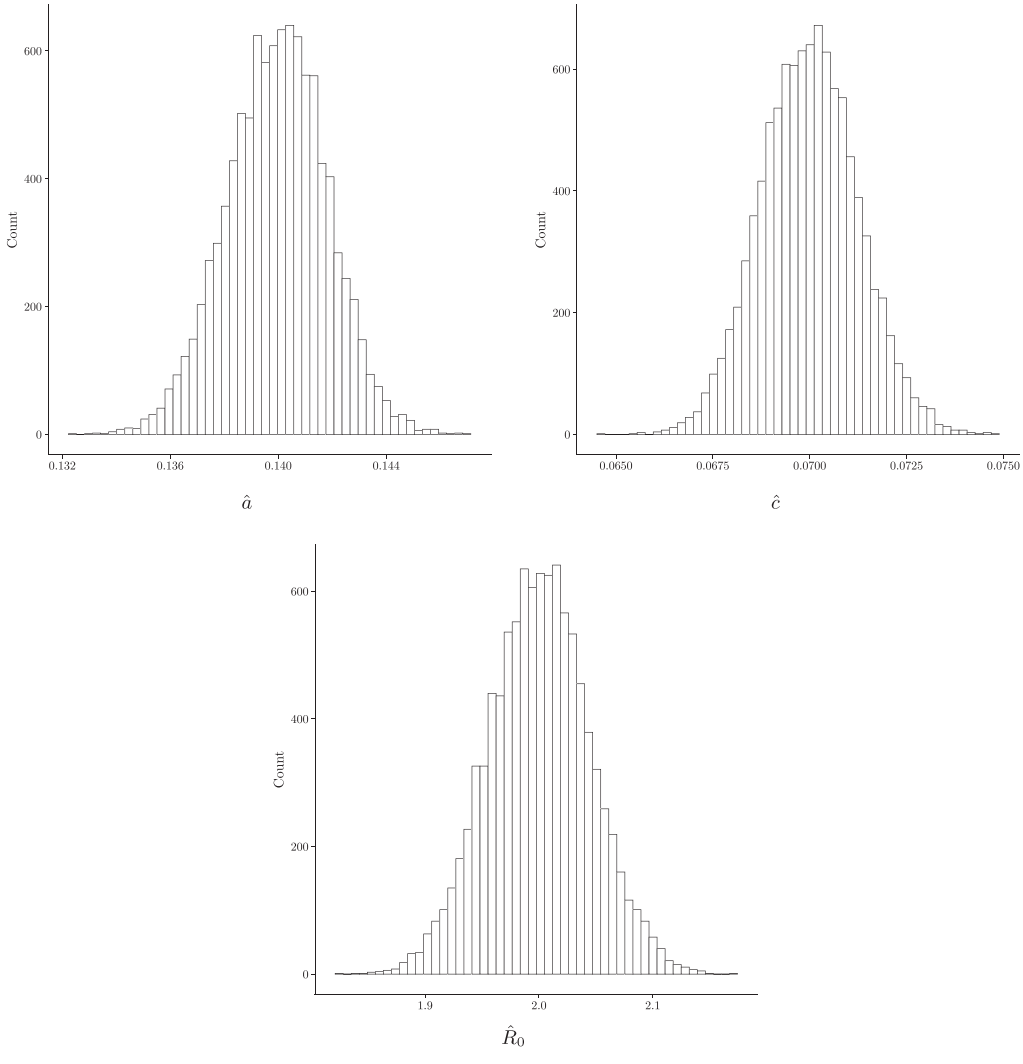


FIGURE 5: Distributions of approximate Poisson ML estimators under $N_2(0) = 1000$.

The $(\text{SIR})^2$ model can be constrained by introducing degrees of infectiveness and of infection vulnerability, denoted by α_j and β_j, respectively. The contagion matrix is $A = \beta\alpha'$. This matrix has reduced a rank equal to 1. The existence of between- and within-population contagions modifies the notion of the reproduction number, which now must account for the different types of contagions. The initial reproduction number now has a matrix form $R_{0,0} = \beta\tilde{\alpha}^\top$, with $\tilde{\alpha}_j = \alpha_j/c_j$, for $j = 1, 2$. The diagonal elements of $R_{0,0}$ can be very different. For instance, if one segment includes super-spreaders, the reproduction number can vary from a value of around 2 (WHO, 2020) to a value between 4.5 and 11.5 (Kochanczik, Grabowski & Lipniacki, 2020).

5.2. Omitted Heterogeneity

Let us now assume an underlying $(\text{SIR})^2$ model and aggregate the two subpopulations in $S = S_1 \cup S_2$, $I = I_1 \cup I_2, R = R_1 \cup R_2$. There is an aggregation bias, which implies that the

TABLE 3: Select summary statistics for the estimated distribution of $\hat{a}$ and correlation (ρ) between $\hat{a}$ and $\hat{c}$.

$N_2(0)$	T	a	c	R_0	Mean($\hat{a}$)	Var($\hat{a}$)	Median($\hat{a}$)	$\rho(\hat{a},\hat{c})$
5	20	0.035	0.07	0.5	0.031	0.00046	0.030	−0.112
5	20	0.140	0.07	2.0	0.131	0.00010	0.135	−0.246
5	40	0.105	0.07	1.5	0.097	0.00052	0.101	−0.380
5	40	0.140	0.07	2.0	0.133	0.00045	0.137	−0.489
100	20	0.140	0.07	2.0	0.139	0.00003	0.140	−0.005
100	40	0.070	0.07	1.0	0.069	0.00002	0.070	0.006
200	20	0.070	0.07	1.0	0.070	0.00002	0.070	−0.027
200	40	0.070	0.07	1.0	0.070	0.00001	0.070	−0.009
300	20	0.070	0.07	1.0	0.070	0.00001	0.070	−0.008
300	40	0.035	0.07	0.5	0.035	0.00001	0.035	0.000

TABLE 4: Select summary statistics for the estimated distribution of $\hat{c}$ and correlation (ρ) between $\hat{a}$ and $\hat{c}$.

$N_2(0)$	T	a	c	R_0	Mean($\hat{c}$)	Var($\hat{c}$)	Median($\hat{c}$)	$\rho(\hat{a},\hat{c})$
50	40	0.035	0.07	0.5	0.0709	0.00007	0.0703	−0.004
100	40	0.070	0.07	1.0	0.0703	0.00002	0.0702	0.006
100	40	0.105	0.07	1.5	0.0702	0.00001	0.0701	−0.007
200	20	0.105	0.07	1.5	0.0701	0.00001	0.0700	−0.007
200	20	0.140	0.07	2.0	0.0701	0.00001	0.0700	−0.007
300	20	0.035	0.07	0.5	0.0701	0.00001	0.0701	−0.004
500	20	0.035	0.07	0.5	0.0701	0.00001	0.0702	−0.003
500	20	0.105	0.07	1.5	0.0701	0.00000	0.0700	0.008
500	40	0.035	0.07	0.5	0.0701	0.00001	0.0702	0.012
1000	20	0.035	0.07	0.5	0.0701	0.00000	0.0704	0.006

TABLE 5: Select summary statistics for the estimated distribution of $\hat{R}_0$.

$N_2(0)$	T	a	c	R_0	Mean($\hat{R}_0$)	Var($\hat{R}_0$)	Median($\hat{R}_0$)
5	40	0.035	0.07	0.5	0.433	0.08843	0.430
50	20	0.035	0.07	0.5	0.499	0.01498	0.492
50	20	0.140	0.07	2.0	1.99	0.04052	1.989
100	20	0.070	0.07	1.0	0.999	0.01404	0.994
100	40	0.070	0.07	1.0	0.993	0.00690	0.994
200	20	0.105	0.07	1.5	1.499	0.00917	1.499
300	40	0.035	0.07	0.5	0.499	0.00160	0.499
500	40	0.035	0.07	0.5	0.500	0.00096	0.500
500	40	0.070	0.07	1.0	0.999	0.00137	0.999
1000	20	0.070	0.07	1.0	1.000	0.00138	0.999

cross-sectional counts $N_1(t) = N_1^1(t) + N_1^2(t)$, $N_2(t) = N_2^1(t) + N_2^2(t)$, and $N_3(t) = N_3^1(t) + N_3^2(t)$ no longer define a Markov process. However, it is still possible to compute the transition matrix at a horizon of one. Let us, for instance, consider the probability that an individual who is at risk at date $t-1$ (i.e., in state S at $t-1$ denoted as $S_{t-1,p}$ where p is subpopulation) is infected at date t, denoted as I_t, by a newly infectious individual. By Bayes' formula

$$\begin{aligned}
P(I_t|S_{t-1}) &= P(I_t|S_{t-1,1})P(S_{t-1,1}|S_{t-1}) + P(I_t|S_{t-1,1})P(S_{t-1,2}|S_{t-1}) \\
&= \frac{N_1^1(t-1)}{N_1(t-1)}\left[a_{11}\frac{N_2^1(t-1)}{N^1} + a_{12}\frac{N_2^2(t-1)}{N^2}\right] \\
&\quad + \frac{N_1^2(t-1)}{N_1(t-1)}\left[a_{21}\frac{N_2^1(t-1)}{N^1} + a_{22}\frac{N_2^2(t-1)}{N^2}\right] \\
&= \left[\beta_1\frac{N_1^1(t-1)}{N_1(t-1)} + \beta_2\frac{N_1^2(t-1)}{N_1(t-1)}\right]\left[\alpha_1\frac{N_1^1(t-1)}{N^1} + \alpha_2\frac{N_2^2(t-1)}{N^2}\right] \\
&= a_t\frac{N_2(t-1)}{N},
\end{aligned}$$

where a_t is the dated transmission parameter in the SIR model with omitted heterogeneity. Therefore, the use of the standard SIR model when there is heterogeneity implies a time-varying contagion parameter. A similar effect, known as the mover-stayer phenomenon, exists for infection state recovery intensity, and leads to a time-varying c_t and, therefore, a time-varying reproduction number $R_{0,0,t} = a_t/c_t$. This type of decomposition can easily be extended to more than two homogeneous subpopulations (Alipoor & Boldea, 2020).

6. INSTANTANEOUS REPRODUCTION NUMBER

There exist different packages on the market for estimating reproduction numbers. These typically use a rolling calibration window. That is, instead of using the entire history of past infections, only a subset of the most recent data (e.g., the past week) is used to estimate a reproduction number. We discuss below two types of estimators. The first, called a generic estimator, approximates the instantaneous reproduction number, a notion that differs from the basic reproduction number. An alternative, called the autoregressive estimator, defines R as an exponential rate of the diffusion of a disease and is usually estimated by either log-regression or Poisson regression (Wallinga & Lipsitch, 2007). Computation and associated software for the instantaneous reproduction number can be found in Cori et al. (2013) and the EpiEstim package (see Appendix A.4 of the Supplementary Material). For the time-dependent reproduction number, see the RO package (Obedia et al., 2012). These are used, for instance, for the official reproduction numbers provided by (PHO, 2020). Both the generic and autoregression estimators use a rolling calibration window and are presented as estimating a time-varying reproduction number. But, methodologies are expected to work in a framework with a weakly time-dependent reproduction number. This is why the discussion here is done under the SIR model with constant parameters.

6.1. The Linearized Mechanistic Model

Both estimation approaches are based on a linearization of the mechanistic model, which assumes a population of infinite size.

6.1.1. The mechanistic model

Let us consider a mechanistic model of infection derived from the SIR model. As in Section 3.2, we denote by $p_1(t)$ and $p_{12}(t)$ the theoretical probabilities corresponding to the frequencies $\hat{p}_1(t) = N_1(t)/n$ and $\hat{p}_{12}(t) = N_{12}(t)/n$. We assume that $\hat{p}$ tends to p when n tends to infinity. In this case, $p(t)$ is also equal to the (unconditional) expectation of $\hat{p}(t)$. Let us focus on the mechanistic component of the model for infection, that is, without considering recovery.

When n varies, we need to appropriately adjust the contagion parameter to derive the mechanistic model by replacing a with $a_n = a/n$. Then we have that

$$E_{t-1}\left(\frac{N_{12}(t)}{n}\right) = a\frac{N_1(t-1)}{n}\frac{N_2(t-1)}{n}.$$

Let us now decompose the count $N_2(t-1)$ as

$$N_2(t-1) = \sum_{s=1}^{t} N_2(t-1;s),$$

where $N_2(t-1;s)$ is the number of individuals infected at $t-s$ for the first time who are still infectious at $t-1$. In the SIR model with geometric infection durations, we have that

$$E_{t-1}\left[\frac{N_2(t-1,s)}{n}\right] = \frac{N_{12}(t-s)}{n}(1-c)^{s-1}$$

and so

$$E\left[\frac{N_2(t-1,s)}{n}\right] = (1-c)^{s-1}E\left[\frac{N_{12}(t-s)}{n}\right].$$

Making n tend to infinity in these relations and using the fact that the limit of the p is deterministic, we get the deterministic recursive equation

$$p^*_{12}(t) = ap_1(t-1)\sum_{s=1}^{t}\left[(1-c)^{s-1}p^*_{12}(t-s)\right],$$

or equivalently,

$$p^*_{12}(t) = a\left[1-\sum_{s=1}^{t}p^*_{12}(t-s)\right]\sum_{s=1}^{t}\left[(1-c)^{s-1}p^*_{12}(t-s)\right], \tag{5}$$

where $p^*_{12}(t) = \lim_{n\to\infty}[N_{12}(t)/n]$. Also, $p^*_{12}(t)$ differs from $p_{12}(t)$ in its denominator: n instead of $N_1(t-1)$, except at the beginning of the disease. From Equation (5), we see that the series $p^*_{12}(t) = E(N_{12}(t)/n)$ satisfies a quadratic recursive equation with an order that tends to infinity with t.

6.1.2. Linearization

A first-order approximation assumes that $p_1(t-1)$ is close to 1. This approximation is reasonable and standard at the beginning of the disease, but will induce biases in the medium run

(when looking for the peak) and in the long run (when looking for final size of the epidemic and herd immunity). Under this approximation

$$p^*_{12}(t) \simeq a \sum_{s=1}^{t} \left[(1-c)^{s-1} p^*_{12}(t-s)\right] = \frac{a}{c} \sum_{s=1}^{t} \left[w(s) p^*_{12}(t-s)\right]$$

$$= R_{0,0} \sum_{s=1}^{t} \left[w(s) p^*_{12}(t-s)\right] \tag{6}$$

with $w(s) = c(1-c)^{s-1}$. The relation in Equation (6) on the expected new infection rates is the basis of the methodology introduced in Fraser (2007).

6.2. The Generic Estimator

6.2.1. Definitions

A generic approach has been introduced in Fraser (2007) and Cori et al. (2013), following a similar idea in Wallinga & Teunis (2004). The method requires knowledge of only the sequence $N_{12}(t)$ of new infections with t varying. The count at time t is written via the lagged counts as

$$N_{12}(t) \simeq \sum_{s=1}^{S} \gamma_s N_{12}(t-s)$$

and the regression coefficients can be normalized as $\gamma_s = w_s \gamma$, where $\sum_{s=1}^{S} w_s = 1$.

The estimated "instantaneous reproduction number" is defined in EpiEstim (Cori et al., 2013) as

$$\hat{R}^i_t = \frac{N_{12}(t)}{\sum_{s=1}^{t} N_{12}(t-s)\hat{w}_s}, \tag{7}$$

where the sum in the denominator starts at the first occurrence of an infection, and $\hat{w}_s$ is a Bayesian estimate of the infectiousness profile. The infectiousness profile of w_s is not necessarily estimated, but, rather, chosen by the practitioner, possibly through a prior (see, e.g., Cori et al. 2013 and the discussion below). The estimator in Equation (7) is not necessarily robust: it depends on the length of the estimation period, the number of lags in the sum appearing in the denominator, and the choice of the infectiousness profile w_s. But more importantly, any generic approach will work well under some implicit assumptions if the notion of interest is correctly defined under these assumptions.

6.2.2. Properties of the EpiEstim approach

Let us illustrate the properties of the EpiEstim approach (see Appendix A.4 of the Supplementary Material for additional details). This estimator is usually computed using a rolling calibration window. It is based on a Bayesian approach with a prior on the distribution of the serial interval, that is, the time from symptom onset in a primary case (infector) to symptom onset in a secondary case (infectee). The log-normal prior depends on two parameters, a mean and a standard deviation. In our EpiEstim1 setting, we have retained the same log-normal prior with a mean of 4.5 days and a standard deviation of 2.5 days, as chosen in PHO (2020). This is close to the prior in Nishiura et al. (2020) with a mean of 4.7 days and a standard deviation of 2.5 days, based on 18 infector–infectee pairs, but different from the prior in Du et al. (2020) with a mean of 3.96 days and a standard deviation of 4.15 days, based on 468 pairs.

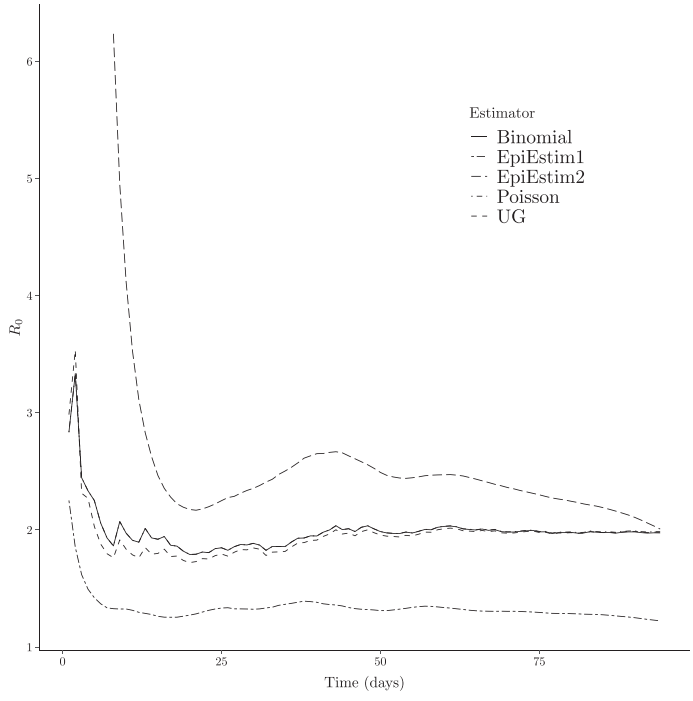


FIGURE 6: Comparison using EpiEstim on simulated SIR model data.

In Figure 6, we display different estimates computed from a simulated series satisfying the SIR model. The EpiEstim1 estimate is calculated using a window of 7 days. The approximate ML estimates (binomial, Poisson, and unfeasible Gaussian) are computed at each date t using all the data from the outbreak. The Poisson and binomial estimates cannot be distinguished. All estimates have poor properties at the beginning, when the number of new infections is rather small and there are almost no recoveries. The ML estimators exhibit decreasing variability, which becomes negligible after 30 days. The estimators then converge to the true value of the basic reproduction number.

Let us now discuss the evolution of the EpiEstim1 estimator. This evolution is strongly dependent on the Bayesian prior used. Indeed, if the estimate is computed using a rolling calibration window, only seven observations are taken into account at each date t, which gives significant weight to the prior. This explains the lack of variability in this estimate over time. Moreover, the level of the estimate is strongly dependent on the selected prior and clearly does not vary around the true value of R_0, even though it accounts for information in the counts of newly infected individuals. In EpiEstim1, we have followed the current practice in which the prior relies on pre-existing estimates of the serial interval distribution. In practice, these estimates may correspond to a different disease or to the same disease in a different country. In the case of COVID-19, this distribution has been estimated using a small number of observations: 18 endogenously selected pairs in Nishiura et al. (2020) (12 of these pairs correspond to transmission within a family and the remaining 6 to short transmissions). Furthermore, these means and standard deviations are estimated using the definition of the serial interval as the time between symptomatic cases (Thompson et al., 2019), which underestimates the mean time between primary and secondary infections and contains uncertainty due to the presence of asymptomatic infection periods and/or individuals.

The choice of a log-normal prior instead of a gamma prior, that is, of a thin-tailed prior instead of a fat-tailed prior, can also lead to an underestimation of the level. A further implication of the Bayesian approach and choice of prior distribution is that the software will generate a nonzero reproduction ratio estimate whether or not new infections are observed in the data. This implies that, based on the estimated R_0, the disease may appear to be contagious when in reality it may be that herd immunity has been achieved in the population.

In order to check the role of the prior, we also display in Figure 6 the plot corresponding to the EpiEstim estimator with a log-normal prior with the same mean and standard deviation as the geometric distribution with a mean of 14 days. This is an unfeasible estimator, assuming that the infectivity profile is fixed at its true value (see the discussion in Section 6.1.2, and Eq. (6)). Convergence to the true value of R_0 is now observed. These drawbacks of the EpiEstim approach have been recently mentioned by some of the authors of the R software package (Thompson et al., 2019), who propose an improved version. This will be discussed below, although the most recent version of the package has not yet been implemented.

The objective of the following sections is to discuss the origin of the EpiEstim approach so as to explain the differences between the estimates observed in Figure 6.

6.2.3. SIR model with stochastic infectious period durations

To understand Equation (7), we have to extend the basic SIR model. We retain a constant contagion parameter a but introduce a stochastic duration of infectiousness D that is not necessarily geometrically distributed. The distribution of D is characterized by the survival function $\gamma(s) = P(D \geq s)$ for $s = 1, 2, \ldots$. The expression of the basic reproduction number is then easily derived (see Section 2.3) as

$$R_{0,t} = \frac{a}{N_1(t)} \sum_{s=0}^{\infty} \left\{ E_t(N_1(t+s)\gamma(s) \right\}. \tag{8}$$

Let us now write this expression in terms of new infections. We have that

$$N_1(t) - N_1(t-1) = -N_{12}(t)$$

and then

$$N_1(t+s) = N_1(t) - \sum_{k=1}^{s} N_{12}(t+k).$$

By replacing $N_1(t+s)$ by this expression in Equation (8), we get, with the convention that $\sum_{k=1}^{0} = 0$,

$$\begin{aligned} R_{0,t} &= \frac{a}{N_1(t)} \sum_{s=0}^{\infty} \left\{ \gamma(s) \left(N_1(t) - E_t \left[\sum_{k=1}^{s} N_{12}(t+k) \right] \right) \right\} \\ &= a \sum_{s=0}^{\infty} \gamma(s) - \frac{a}{N_t(t)} \sum_{s=1}^{\infty} \sum_{k=1}^{s} \left[\gamma(s) E_t(N_{12}(t+k)) \right] \\ &= a \sum_{s=0}^{\infty} \gamma(s) - \frac{a}{N_1(t)} \sum_{k=1}^{\infty} \left[E_t N_{12}(t+k) \sum_{s=k}^{\infty} \gamma(s) \right]. \end{aligned}$$

The partial sums of the survival function $\gamma(s)$ can be rewritten in terms of the moments of the stochastic duration of infectiousness as

$$R_{0,t} = aE(D) - \frac{a}{N_1(t)} \sum_{k=1}^{\infty} \left\{ E\left[(D-k)^+\right] E_t\left(N_{12}(t+k)\right] \right\}, \tag{9}$$

where $x^+ = \max\{x, 0\}$.

Remark 2. In the standard SIR model, Equation (9) becomes

$$R_{0,t} = (a/c) \left\{ 1 - \sum_{k=1}^{\infty} \left[(1-c)^k E_t(N_{12}(t+k)) \right] \right\}.$$

Let us now discuss the conditional expectation E_t. In the SIR framework, the conditioning set includes the current and lagged values of $N_{jk}(t)$ for $j, k = 1, 2, 3$, or equivalently, of the cross-sectional counts $N_k(t)$ for $k = 1, 2, 3$. Therefore, a sufficient summary of the past information requires two sequences of counts. By considering a single sequence of counts, e.g., the counts of newly infected people only, the generic approach changes the information set and modifies the definition of the dated reproduction number (see the discussion in Section 6.3). With this restricted information set, the new reproduction number is

$$R^N_{0,t} = aED - \frac{a}{N_1(t)} \sum_{k=1}^{\infty} \left\{ E[(D-k)^+] E\left[N_{12}(t+k) | \underline{N_{12}(t)} \right] \right\}, \tag{10}$$

where the superscript N indicates the restriction to new infections. Can we expect a linear prediction formula for the counts of newly infected people, such as

$$E\left[N_{12}(t+k) | \underline{N_{12}(t)} \right] = \sum_{h=0}^{\infty} \beta_{kh} N_{12}(t-h),$$

with time-independent coefficients β_{kh}? Likely not, considering the nonlinear dynamics of the contagion model during a nonstationary episode.

6.2.4. Which definition of the reproduction number?

To understand the significant difference between Equation (7) for $\hat{R}^i_t$ and Equation (10) for R^N_{0t}, it is useful to come back to the paper in which the notion of the instantaneous reproduction number was introduced (Fraser, 2007). This notion is based on the renewal equation

$$I(t) = \sum_{s=1}^{\infty} \beta(t, s) I(t-s), \tag{11}$$

where $I(t)$ is the incidence proportion—see CDC (2012) for different definitions of incidence depending on the selected denominator—at t, also called the attack rate (approximated by $N_{12}(t)/N_1(t-1)$) and where $\beta(t, s)$ is the effective contact rate between infectious and susceptible individuals, taking into account the generation of newly infected people. Both the SIR model and the renewal equation appear in Kermack & McKendrick (1927) and are compatible. Under the SIR model, the contact rate $\beta(t, s)$ is a complicated nonlinear function of the sufficient summary counts, that is, the newly infected and newly recovered counts between dates $t - s$

and t. Therefore, in the SIR framework, the renewal equation (Eq. (11)) involves a "lagged endogenous" contact rate which is, in fact, an equilibrium contact rate.

Let us now give the definitions of the reproduction ratios in Fraser (2007). Two notions, called the "case reproduction ratio" and the "instantaneous reproduction ratio," are introduced with the main objective of getting a ready-to-use measure based on simple analytic formulas. These notions have new names since they significantly differ from the standard basic and effective reproduction numbers. Moreover, they do not have the same interpretation. For instance, the instantaneous reproduction number is defined in Equation (11) by considering what reproduction can be expected if "conditions remain unchanged," i.e., $I(t-s) = I$ for $s = 1, 2$. The ratio is then defined as (see Eq. (3) in Fraser, 2007)

$$R_t^i = I_t / I = \sum_{s=1}^{\infty} \beta(t, s).$$

This practice disregards the endogeneity of the contact rates. Indeed, the contact rates also depend on the evolution of the number of newly infected individuals, which is assumed to be unchanged in the "linear" component of the renewal equation but not in the (nonlinear) contact rate. Moreover, the assumption of unchanged conditions is not necessarily compatible with the evolution of infected counts observed in the SIR model and the observations of $I(t)$ or $N_{12}(t)$.

Finally, to derive the Equation (7), a decomposition of the contact rate as $\beta(t, s) = R_t^i w(s)$ is assumed, where the $w(s)$ for $s = 1, \ldots, S$, sum to 1. By taking into account this reduced rank condition, the renewal equation (Eq. (11)) is equivalent to

$$R_t^i = I_t \Big/ \sum_{s=1}^{S} \left[I_{t-s} w(s) \right],$$

which explains the generic estimate in Equation (7)—only if $N_1(t)$ is not changing greatly (see the discussion in Section 6.1)—and its interpretation as the ratio of new infections by the total infectiousness of infected individuals up to time $t-1$. A precise discussion of the assumptions underlying the generic estimator shows at least three sources of bias whose impacts can be observed in Section 6.2.2. They are (i) changes in the definitions of the reproduction number; (ii) endogeneity bias when assuming exogenous contact rates; and (iii) linearization bias when considering the linearized mechanistic model.

6.3. The Autoregression Estimate

An alternative estimator of the reproduction number can be introduced based on the approximate asymptotic relation in Equation (6). This estimator depends only on the counts of newly infected individuals and is easy to compute.

First, select an autoregressive order H, and then regress $N_{12}(t)$ on $(N_{12}(t-1), \ldots, N_{12}(t-H))$ without an intercept by OLS for $t = H+1, \ldots, T$. Defining $\hat{\gamma}(s)$ for $s = 1, \ldots, H$, as the estimated regression coefficients, the estimator of the reproduction number is

$$\hat{R}_{00}^{AR} = \sum_{s=1}^{H} \hat{\gamma}(s).$$

This estimator has a variance that will increase with H since more underlying parameters have to be estimated. This estimator also has the drawback of being computable only after at least $2H+1$ days because of the lag and the minimum number of observations necessary to identify

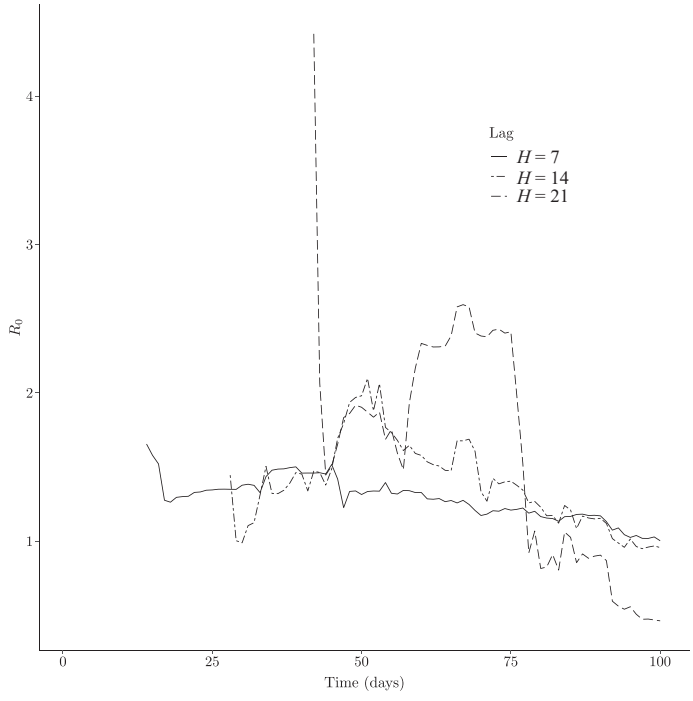


FIGURE 7: The autoregression estimate.

the autoregressive parameter. The estimates in Figure 7 have been computed for $H = 7, 14, 21$ days in a nonrolling way on the same set of simulated data used to generate Figure 6. The change in variability with H and the infeasibility of using this estimator at the beginning of an epidemic are clear. Moreover, this estimator does not converge to the true value. This approach is also subject to both causality and linearization biases, as is observed in Figure 7 with the underestimation of R_0.

7. CONCLUDING REMARKS

The reproduction number is used as a basic tool to follow the progression of an epidemic such as COVID-19 and monitor the effects of health policy. For instance, specific partial lockdown policies may be introduced if the estimated R_0 is larger than 1. Such policies neglect the variability in both definitions of R_0 and its estimators. We have considered this question in the framework of a discrete-time SIR model and have shown that this variability can be due to the definition of R_0 itself, which is time-dependent and sometimes author-dependent, or to an omitted underlying heterogeneity. This variability is also a consequence of the different estimation methods used, with bias and uncertainty that depend on the available information.

As a by-product, we have shown that the estimate of R_0 based on the Poisson approximate likelihood of the SIR model, used in a rolling fashion and possibly with a prior on model parameters (see Appendix A.5 of the Supplementary Material for Bayesian estimation), is as simple as the approach suggested in the standard EpiEstim package, but with two advantages: the former uses information from both newly infected and currently infected people, and it does not arbitrarily fix the infectiousness profile. Furthermore, our framework could be easily applied in a Canadian context using publicly available infection and recovery count data. Given the structure of the basic SIR model used here, the data currently published by the Government

of Canada at this stage of the epidemic are sufficient to compute estimates of R_0 using our framework.

As mentioned in the text, Thompson et al. (2019) highlights issues related to the use of the standard EpiEstim package and proposes an improved version to correct some drawbacks. They propose to use, in real time, two series of data: the counts of newly infected people, and "up-to-date observations of serial intervals". In this approach, there is (as in the rolling approach based on the SIR model) a larger information set. This update also introduces a path-varying mean and standard deviation for the serial interval distribution. The two approaches do not differ greatly in the underlying model on which they are based (see the discussion in Sections 6.1.1 and 6.1.2), but instead by the observations they are using to calibrate the parameters, that is, the counts of newly infected and currently infected people in the SIR-model-based estimator and the counts of newly infected people and data on infector–infectee pairs obtained by tracing.

In general, the choice of approach should largely depend on the availability (and cost) of data, especially at the beginning of the epidemic, and on the reliability of the data. In particular, available data can be incomplete if it does not account for undetected, asymptomatic people (Gourieroux & Jasiak, 2020b) and can be left-or right-censored due to infector–infectee pair tracing.

The SIR model has been used in this article since different estimation approaches are implicitly based on this model. This choice has facilitated our discussions and comparisons. Clearly, to obtain a more complete picture of R_0 in the context of the SIR model, similar experiments should be conducted using a model with more features. The aim of this extended analysis could be to account for differences between the infection and infectious periods or to incorporate a stochastically time-varying contagion parameter (Gourieroux & Lu, 2020).

ACKNOWLEDGEMENTS

We thank the anonymous referee and the editor for their helpful comments. We also gratefully acknowledge financial support from the Autorité de contrôle prudentiel et de résolution - ACPR chair "Regulation and Systemic Risk," and the Agence Nationale de la Recherche (ANR-COVID) grant (ANR-17-EUR-0010).

REFERENCES

Adam, D. (2020). A guide to R—The pandemic's misunderstood metric. *Nature*, 583, 346–348.

Alipoor, A. & Boldea, O. (2020). *The role of elementary schools in SARS-CoV-2 transmission*, Tilburg University Discussion Paper.

Allen, L. (1994). Some discrete time SI, SIR and SIS epidemic models. *Mathematical Biosciences*, 124, 83–105.

Breto, C., He, D., Ionides, E., & King, A. (2009). Time series analysis via mechanistic models. *Annals of Applied Statistics*, 3, 319–348.

Cauchemez, S., Boelle, P., & Donnelly, C. (2006). Real time estimates in early detection of SARS. *Emerging Infectious Diseases*, 12, 110–113.

Center for Disease Control and Prevention (CDC). (2012). *Principles of Epidemiology in Public Health Practice: An Introduction to Applied Methodology and Biostatistics*, 3rd ed. Section 3.

Cori, A., Ferguson, N., Fraser, C., & Cauchemez, S. (2013). A new framework and software to estimate time varying reproduction numbers during epidemics. *American Journal of Epidemiology*, 178, 1505–1512.

Du, Z., Xu, X., Wu, Y., Wang, L., Cowling, B., & Ancel Meyers, L. (2020). Serial interval of COVID-19 among publicly reported confirmed cases. *Emerging Infectious Diseases*, 26, 1341–1343.

Farrington, P. & Whitaker, H. (2003). Estimation of effective reproduction numbers for infectious diseases using serological survey data. *Biostatistics*, 4, 621–632.

Fraser, C. (2007). Estimating individual and household reproduction numbers in an emerging epidemic. *PLoS ONE*, 2. e758

Gourieroux, C. & Jasiak, J. (2020a). Time varying Markov process with partially observed aggregate data: An application to coronavirus. *Journal of Econometrics*. Forthcoming

Gourieroux, C. & Jasiak, J. (2020b). Analysis of virus transmission: A stochastic transition model representation of epidemiological models. *Annals of Economics and Statistics*, 140, 1–26.

Gourieroux, C. & Lu, Y. (2020). *SIR model with stochastic transmission*, CREST Discussion Paper.

Kermack, W. & McKendrick, A. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115, 700–721.

Kochanczik, M., Grabowski, F., & Lipniacki, T. (2020). *Accounting for super-spreading gives the basic reproduction number R_0 of COVID-19 that is higher than initially estimated*. medRxiv preprint.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., & Tong, Y. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *The New England Journal of Medicine*, 382, 1199–1207.

Ma, J. & Earn, D. (2006). Generality of the final size formula for an epidemic of a newly invading infectious disease. *Bulletin of Mathematical Biology*, 68, 679–702.

McDonald, G. (1952). The analysis of equilibrium in malaria. *Tropical Diseases Bulletin*, 49, 813–829.

Nishiura, H., Linton, N., & Akhmetzhanov, A. (2020). Serial interval of novel coronavirus (COVID-19) infections. *International Journal of Infectious Diseases*, 93, 284–286.

Obedia, T., Haneef, R., & Boelle, P. Y. (2012). The RO Package: A toolbox to estimate reproduction numbers for epidemic outbreaks. *BMC Medical Informatics and Decision Making*, 12, 147.

Public Health Ontario (PHO). (2020). *Epidemiological summary: Evolution of COVID-19 case growth in Ontario, June 2020*. https://www.publichealthontario.ca/-/media/documents/ncov/epi/covid-19-epi-evolution-case-growth-ontario.pdf?la=en.

Riou, J. & Althaus, C. (2020). Pattern of early human to human transmission of Wuhan (2019-nCoV) novel coronavirus, December 2019 to January 2020. *Eurosurveillance*, 25, 4.

Sanche, S., Liu, Y., Xu, C., Romero-Severson, E., Hengartner, N., & Rulan, K. (2020). High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2. *Emerging Infectious Diseases*, 26.

Sanchez, M. & Blauer, S. (1997). Uncertainty and sensitivity analysis of the basic reproduction rate: Tuberculosis as an example. *American Journal of Epidemiology*, 145, 1127–1137.

Smieszek, T. (2009). A mechanistic model of infection: why duration and intensity of contacts should be included in models of disease spread. *Theoretical Biology and Medical Modelling*, 6, 25.

Thompson, R., Stockwin, J., Van Gaalen, R., Polonsky, J., Kamvar, Z., Demarsh, P., & Dahlqwist, J. (2019). Improved inference of time-varying reproduction numbers during infectious disease outbreaks. *Epidemics*, 29, 10036.

Wallinga, J. & Lipsitch, M. (2007). How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences*, 274, 599.

Wallinga, J. & Teunis, P. (2004). Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology*, 160, 509–516.

White, L. & Pagano, M. (2008). A likelihood-based method for real time estimation of the serial interval and reproduction number of an epidemic. *Statistics in Medicine*, 27, 2999–3016.

World Health Organization (WHO). (2020). *Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19)*, https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf.

Wu, J., Leung, K., Bushman, M., Kishore, N., & Niehus, R. (2020). Estimating clinical security of COVID-19 from the transmission dynamics in Wuhan, China. *Nature Medicine*, 26, 506–510.

Zhang, S., Diao, M., Yu, W., Pei, L., Lin, Z., & Chen, D. (2020). Estimation of the reproductive number of novel coronavivus (COVID-19) and the probable outbreak size on the Diamond Princess cruise ship: A data-driven analysis. *International Journal of Infectious Disease*, 93, 201–204.

Received 23 December 2020
Accepted 27 July 2021