Published in Image Processing On Line on YYYY–MM–DD.
Submitted on YYYY–MM–DD, accepted on YYYY–MM–DD.
ISSN 2105–1232 © YYYY IPOL & the authors CC–BY–NC–SA
This article is available online with supplementary materials,
software, datasets and online demo at
https://ipolcore.ipol.im/demo/clientApp/demo.html?id=304

A daily measure of the SARS-CoV-2 effective reproduction number for all countries

Tahar Zamene Boulmezaoud[1], Luis Alvarez[2], Miguel Colom[3], and Jean-Michel Morel[3]

[1] Université Paris-Saclay, UVSQ, Laboratoire de Mathématiques de Versailles, 78035, Versailles Cedex, France.
(tahar.boulmezaoud@uvsq.fr)
[2] Departamento de Informática y Sistemas, Universidad de Las Palmas de Gran Canaria, Spain
(lalvarez@ulpgc.es)
[3] Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli, F-94235, Cachan, France
({miguel.colom-barco, jean-michel.morel}@ens-paris-saclay.fr)

PREPRINT November 24, 2020

Abstract

We describe a transparent method calculating an "effective reproduction number (ERN)" from the daily count (incidence) of newly detected cases in each country, in the EU and in each US state. We aim at getting a result as faithful as possible to the observed data, which are very noisy. The noise, being specific of administrations, shows a seven days period. Hence the incidence curve is first filtered by a seven days mean or median filter. Then the ERN is computed by a classic reproduction formula due to Nishiura. To do so requires knowledge of the serial interval function $\Phi(s)$ which models the time between the onset of symptoms in a primary case and the onset of symptoms in secondary cases, or equivalently the probability that a person confirmed infected today was actually infected s days earlier by another confirmed infected person. We use and compare several recently proposed evaluations of Φ, and verify that their variation has moderate practical incidence on the evaluation of the ERN. The method we present derives from Nishiura's formula but we prove that for the adequate choice of parameters it is identical to one of the methods proposed by the classic EpiEstim (Estimate Time Varying Reproduction Numbers from Epidemic Curves) software. We find that the same method can be applied to compute an effective reproduction number from the daily death count, which yields therefore another prediction of the expansion of the pandemic. Although this application has no clear theoretical justification, we find good experimental fit of the ERN curves obtained from the incidence and from the death curve, up to a time shift. In most countries, both curves appear to be similar, with a time delay that depends on each country's detection and administrative processing delays. Both ERNs can be consulted daily online in the demo tag associated with this paper. We refer the readers to the online demo[1] to experiment by themselves. In the case of France, an ERN based on hospitalizations, new entries in ICU's and deaths at hospitals is also computed daily[2].

[1] https://ipolcore.ipol.im/demo/clientApp/demo.html?id=304
[2] https://ipolcore.ipol.im/demo/clientApp/demo.html?id=77777000102

Source Code

The code is quite simple. To simplify its presentation we use a single file where all basic procedures are included, from the management of data to the parameter optimization. The reviewed source code and documentation for this algorithm are available from the web page of this article[3]. Compilation and usage instruction are included in the `README.txt` file of the archive.

Keywords: SARS-CoV-2, epidemic, daily reproduction number, ERN.

1 Introduction and previous work

Several websites give daily statistics and estimates of the number of detected positive cases and of the death toll of SARS-CoV-2 infections. For example in [1] these data are given for each country, but it is not raw data. The curves displayed are clearly the result of a semi-global estimation. In [11] one finds a presentation of the raw curve of current deaths, together with a curve giving an estimate of the daily effective reproduction number based on the death curve. Yet, only a smoothed curve is given for the daily infections and no reproduction curve is associated with it.

In the website [9] by clicking the *Résultats* tag, one can observe a prediction of the daily reproduction number by country and region, obtained by two different estimation methods, and based on three different estimates of the *serial interval*. The method for computing the ERNs is explained in [8], where the *reproduction number* $R(t)$ is explained in the following terms (we translate):

> One of the key parameters in an epidemic is the reproduction number R which characterizes the number of people infected by a contagious person during the course of his or her infection. At the beginning of an epidemic, when the whole population is susceptible (i.e. not immune), this number takes on a particular value noted R_0 and called the basic reproduction number. During the course of the epidemic, when the proportion of immunized people becomes large enough to slow down the transmission of the virus (by an effect comparable to a dilution of the individuals still susceptible), this number is called the effective, or temporal, reproduction number, $R(t)$.
>
> Intuitively, if $R(t) > 1$, then a person infects more than one person on average and the epidemic is in a growing phase. As the COVID-19 epidemic spreads, $R(t)$ decreases, as an increasing proportion of the population becomes immune. When the threshold for group immunity is exceeded $R(t)$ falls below 1, an epidemic peak is reached and the epidemic decreases. Public health control measures can also decrease $R(t)$ and thus reach an epidemic peak before the threshold of group immunity is reached. In addition, as observed in [14] the pandemic spread may become sub-exponential due to confinement measures, or grow again after they are relaxed. In short, $R(t)$ can vary in a somewhat erratic way by human intervention. At time t therefore, knowing the value of $R(t)$ is essential to determine the status of the epidemic.

The online computations are based on several references. We translate again from [9]:

> The 2013 EpiEstim software developed by Cori et al. [7] updated by Thompson et al. [19] is based on an approach different from the $R0$ software, motivated by the fact that in the situation where the epidemic under study would still be ongoing, and more particularly when it comes to evaluating the effectiveness of control measures (a very current situation

[3] https://ipolcore.ipol.im/demo/clientApp/demo.html?id=304

therefore), the total number of infections caused by the latest cases detected is not yet known. This weakness is an opportunity to highlight two approaches to the number of temporal reproduction, namely that of Wallinga & Teunis [20], Obadia et al. [18] computing a case (or cohort) replication number, which is retrospective: its calculation is based on the number of secondary cases actually caused by a cohort of infectors detected from the date on which the latter were detected.

In [5] the approach to ERN is based on the classic SEIR model. The simplest SEIR model classifies individuals as susceptible (S), exposed (E), infectious (I), recovered (R), and dead (D), to which one can add a variable for the cumulative number of new detected cases (C) This leads to a system of six differential equations linking these numbers, depending on four parameters. The transmission rate is β. Infectious individuals either recover or die at the mean rates γ and δ, respectively. Individuals in latent period (E) progress to the infectious class at the rate k (where $\frac{1}{k}$ suggests the mean latent period). So the resulting SEIRDC system reads

$$\begin{aligned}
\frac{dS(t)}{dt} &= -\beta\frac{S(t)I(t)}{N} \\
\frac{dE(t)}{dt} &= \beta\frac{S(t)I(t)}{N(t)} - kE(t) \\
\frac{dI(t)}{dt} &= kE(t) - (\gamma+\delta)I(t) \\
\frac{dR(t)}{dt} &= \gamma I(t) \\
\frac{dD(t)}{dt} &= \delta I(t) \\
\frac{dC(t)}{dt} &= kE(t).
\end{aligned} \tag{1}$$

The basic reproduction number (BRN) is obtained as

$$R_0 = \frac{\beta}{\gamma+\delta}.$$

The parameters β and δ can be obtained by a best data fit between the observed temporal statistics of S, E, I, R, D and C and the solution of (2) in the time period where the pandemic evolves freely. If lock down or other mitigation measures are taken, the parameter β becomes actually time dependent, and such a global estimate is no longer possible.

Similarly, in the sophisticated recent study on effects of the lock-down in France [15], a numerical analysis of the daily hospital data (arrivals in regular and critical care units, releases and deaths), is provided using extended SEIR models. These models involve ratios of evolutionary timescales to branching fractions, assumed uniform throughout a country, and the basic reproduction number, R_0 before and during the national lock-down, for each region of France. The study is based on a joint-region Bayesian analysis. The extended SEIR model becomes SEIHCDRO, by inclusion of the new categories H (hospitalized), and O (other recovered, concerning cases that did not pass through hospital), then SEIFHCDRO which splits the infectious compartment into one where people effectively contaminate the Susceptibles, and another where people are too ill to go outside and contaminate Susceptibles. These new compartments are called the Asymptomatic and Feverish phases. The various models estimate, among other parameters, two R0 factors (before and after lock down). They therefore do not provide a daily reproduction rate.

1.1 The serial interval function

A short cut for computing the ERN $R(t)$ passes by the knowledge of a serial interval function, sometimes also called serial interval. As explained in [2]:

> Rather than resorting to fully parametric models and seeing $R(t)$ as the by-product of its identification, a more phenomenological, semi-parametric approach can be followed [7], [18] [19]. This approach has been reported as robust and potentially leading to relevant estimates of $R(t)$, even for epidemic spreading on realistic contact networks, where it is not possible to define a steady exponential growth phase and a basic reproduction number [19]. The underlying idea is to model incidence data[4] $i(t)$ at time t as resulting from a Poisson distribution with a time evolving parameter adjusted to account for the data evolution. This parameter can be written as
>
> $$R(t)\sum_{s\geq 1}\Phi(s)i(t-s),$$
>
> where $i(t-s)$ accounts for the past incidence data, as convolved with a function $\Phi(s)$ standing for the distribution of the serial interval.
>
> **Definition 1.** *The serial interval function $\Phi(s)$ models the time between the onset of symptoms in a primary case and the onset of symptoms in secondary cases, or equivalently the probability that a person confirmed infected today was actually infected s days earlier by another infected person.*
>
> The serial interval function is thus an important ingredient of the model, accounting for the biological mechanisms in the epidemic evolution. Assuming the distribution Φ to be known (which can be questionable), the whole challenge in the actual use of the semi-parametric Poisson-based model thus consists in devising estimates $\hat{R}(t)$ of $R(t)$ that have better statistical performance (more robust, reliable and hence usable) than the direct brute-force and naive form:
>
> $$\hat{R}_{naive}(t)=\frac{i(t)}{\sum_{s\geq 1}\Phi(s)i(t-s)}.$$

This approach derives from [12], where the authors proposed to measure explicitly the reproduction number and generation time, by recording all individual-level transmission events. They found that the classical concept of the basic reproduction number is untenable in realistic populations, and does not provide any conceptual understanding of the epidemic evolution. This departure from the classical theoretical picture is not due to behavioral changes and other exogenous epidemiological determinants. Rather, it can be simply explained by the (clustered) contact structure of the population. This led the authors to promote methodologies aimed at estimating the instantaneous (or effective) reproduction number (which we call ERN) to characterize the correct epidemic dynamics from incidence data. They assume that the number of cases cases $i(t)$ at time t can be approximated by a Poisson according to the following equation:

$$i(t)\simeq Poisson\left(R(t)\sum_{s=1}^{t}\Phi(s)i(t-s)\right),$$

[4] For coherence with the present text, we change the original notation of the paper. In it, the daily incidence $i(t):=C'(t)$ was denoted $C(t)$.

where Φ is the generation time distribution and $R(t)$ is the effective reproduction number at time t. The likelihood L of the observed time series of cases from day 1 to T is thus given by

$$L = \Pi_{t=1}^{T} \mathbb{P}\left(i(t), R(t)\sum_{s=1}^{t}\Phi(s)i(t-s)\right),$$

where $\mathbb{P}(k, \Phi)$ is the probability mass function of a Poisson distribution (i.e., the probability of observing k events if these events occur with a known rate Φ). The posterior distribution of $R(t)$ is then explored using MCMC sampling. As we shall see, though, the Poisson model for administrative cases and deaths statistics is questionable.

In [2] the problem of estimating $R(t)$ by maximum likelihood estimation of L is complemented by a piecewise regularity term for $R(t)$, instead of using a Bayesian framework. This regularity term in the variational model is complemented by a spatial regularity term to ensure that neighboring French districts have similar values for $R(t)$.

1.2 Available serial interval functions for COVID-19

We now discuss what serial interval functions Φ are available for COVID-19. The computation of an effective, or instantaneous, reproduction number is much more problematic than its global estimate on a large period where the pandemic runs free. In [6] for example, the reproduction number of the Spanish influenza was estimated from daily case notification data using several variants of a SEIR model, but the estimate was based on a long period, was therefore not time dependent as it should be in periods where lock-down strategies or other distancing measures are being applied.

As we saw, the *serial interval* in epidemiology refers to the time between successive observed cases in a chain of transmission. In the case of COVID-19 this interval seems to range between 3 and 8 days. The authors of [10] define this interval as follows:

> The serial interval of COVID-19 is defined as the time duration between a primary case (infector) developing symptoms and secondary case (infectee) developing symptoms.

Hence, by a careful inquiry on many pairs of patients, where one is the probable cause of the infection of the other, one may obtain the distribution of the serial interval in practice, as it has been done in [10] on 468 cases. The conclusion of the authors is that the serial interval mean and standard deviation are 3.96 and 4.75 respectively with 12.6% of reports indicating pre-symptomatic transmission.

The observed serial distribution in [10] had cases on negative days, meaning that in many instances the infectee had developed symptoms up to 10 days before the infector.

The obtained serial interval is shown in Figure 1 where we also show the truncated time series obtained by removing the bins on negative days.

In [9], the serial interval is defined as the length of time a person is contagious. It can be estimated by tracking contacts (i.e., infector-infected pairs) and by counting the number of days between the dates of onset of symptoms in the infecting and infected individuals respectively. Another serial interval function was estimated on a few (28) patients by [17], which is obviously problematic. This distribution and two parametric regressions are shown in Figure 2.

Figure 3 shows the ERN of the EU calculated between March 15 and June 15. From left to right, the computation is based on the serial intervals of [10, 13, 17]. There is very good agreement between all results. This is reassuring, given the differences between the three evaluations of the serial interval.

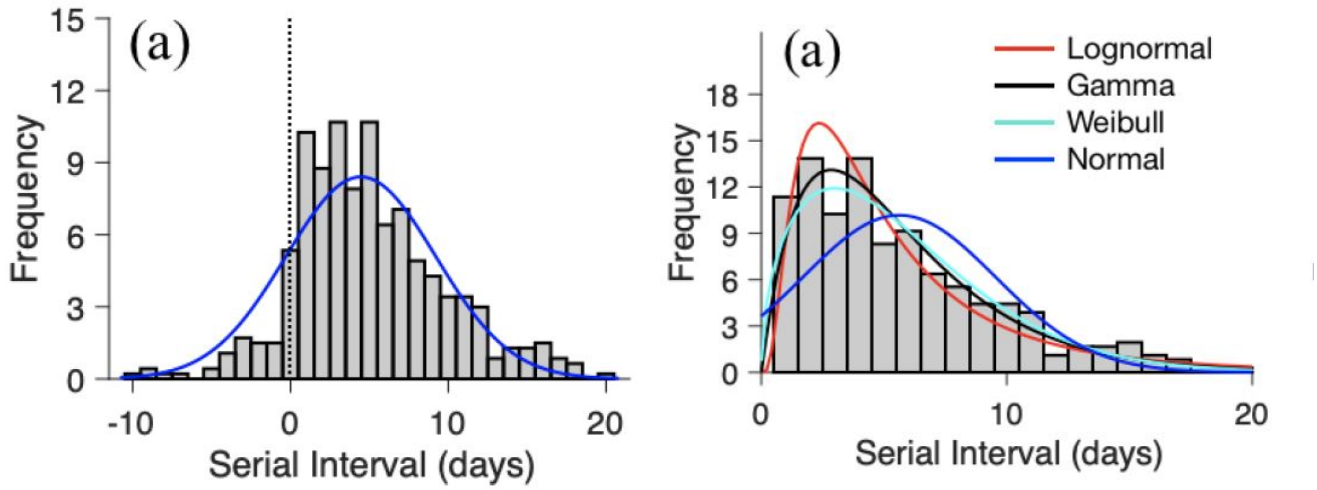

Figure 1: Left: Estimated serial interval distribution for COVID-19 based on 468 reported transmission events in China between January 21, 2020 and February 8, 2020. Bars indicate the number of infection events with specified serial interval and blue lines indicate fitted normal distributions for all infection events (N = 468) reported across 93 cities of mainland China by February 8, 2020. Negative serial intervals (left of the vertical dotted lines) suggest the possibility of COVID-19 transmission from asymptomatic or mildly symptomatic cases. Right: the truncated data removing all non-positive values for all 468 infection events. In the demo we accumulated at day 0 all days before 0 of histogram (a) [10], to obtain to obtain Φ = [1237, 1026, 873, 1068, 783, 1068, 636, 702, 486, 429, 342, 342, 302, 84, 130, 130, 150, 81, 63, 0, 24] / 9956.

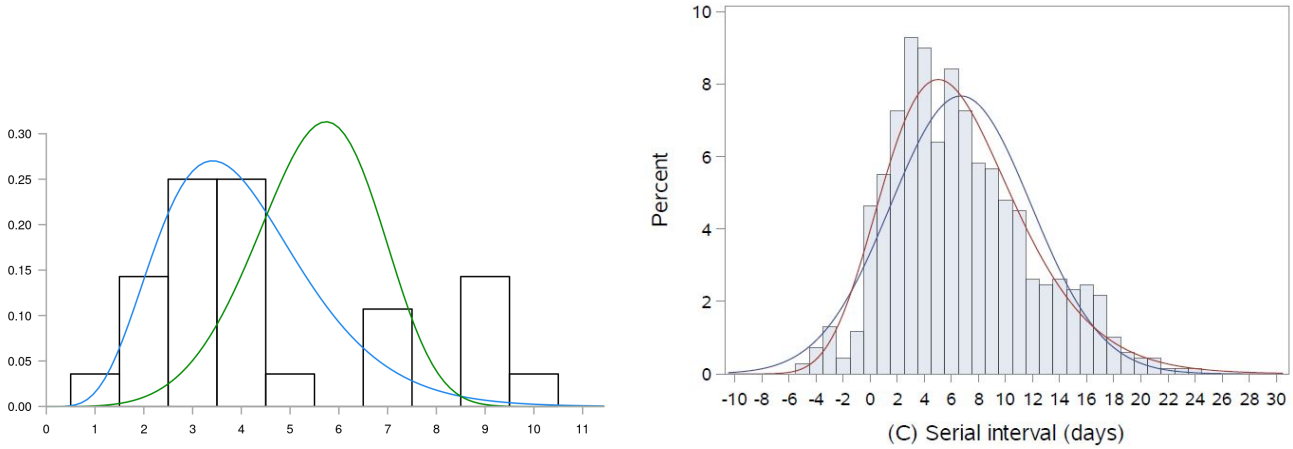

Figure 2: Left: three serial intervals considered in [9]: the Nishiura et al. histogram obtained on 28 patients, a Gamma(6.5,0.62) distribution (in blue) and a Weibull(5,6) distribution (in green). We convolved this histogram with $[1, 2, 1]$ to obtain Φ = [1.5, 4, 6.25, 5.5, 2.25, 1, 1.5, 1.75, 2.25, 1.5] / 27.5. Right: serial interval proposed in [13]. It has values on negative days. Hence, like for the serial interval proposed in [10], we shall shift and add these values on time 0. We used Φ = [848, 550, 724, 927, 900, 637, 840, 725, 580, 565, 477, 450, 260, 246, 260, 230, 245, 215, 100, 55, 45, 45, 15, 15, 15]/9969.

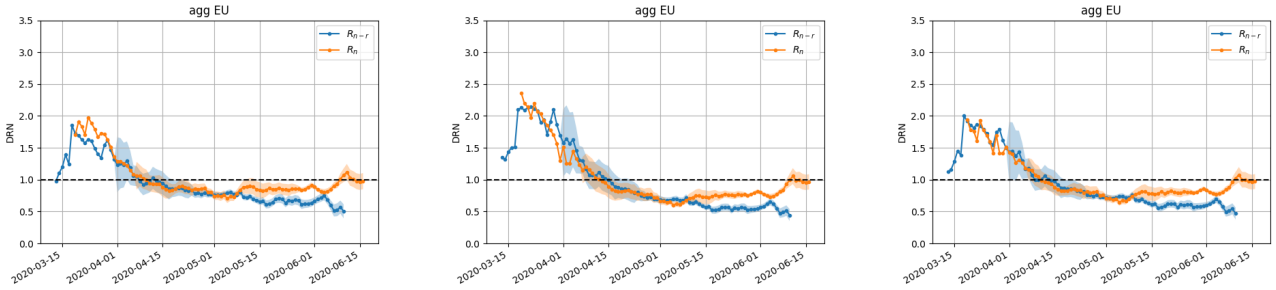

Figure 3: Computation of the ERN of the EU. f From left to right, the ERN time series obtained using respectively the serial intervals of [10, 13, 17].

2 Computing the ERN

2.1 The Nishiura formula

The simple formula underlying the calculation of the reproduction number at time t, $R(t)$, from the so-called *incidence*, namely the number of new detected cases $i(t)$ at time t, and the *serial interval function*, $\Phi(t)$ is given in Nishiura 2007 [16] as

$$i(t) = \int_0^t i(t-\tau)R(t-\tau)\Phi(\tau)d\tau. \tag{2}$$

Here $\Phi(\tau)$ can be interpreted as the probability that an incident case was contaminated τ days ago. According to the author, Equation (2) is a reformulation of a 2004 method proposed for SARS in [20]. The advantages of formula (2) is that it only requires the time of onset of cases (i.e. the model does not require the total number of susceptible individuals or detailed contact information) and the time-dependent reproduction number can therefore be reasonably estimated using a far simpler equation than other population dynamics models.

Our method requires the observation of:

- the number of new daily cases of SARS-CoV-2 infections, denoted as $i_n = i(n)$ on day n.
- or the number of new deaths per day attributable to SARS-CoV-2 contamination, denoted $d_n = d(n)$ on day n.
- an empirical probability distribution $\Phi = (\Phi_1, \cdots, \Phi_f)$.

Nevertheless, our computation will be justified for the analysis of i_n. Applying the same process to d_n is exploratory, as we use the same serial interval as for i_n.

Let us consider an infected person detected today. This person was infected one day ago with a probability Φ_1, two days ago with a probability Φ_2, ..., f days ago with a probability Φ_f. We assume that the sum of these coefficients $\Phi_1, \cdots, \Phi_f$ coefficients is 1. As proposed in [7, 4], the ERN is estimated as the ratio

$$R_n^\star := \frac{i_n}{\Phi_1 i_{n-1} + \Phi_2 i_{n-2} + \cdots + \Phi_f i_{n-f}}, \tag{3}$$

where f can range from 10 to 20 depending on the estimation of the serial interval function Φ. This

equation can be derived from the discrete version of (2),

$$i_n = \sum_{k=1}^{f} i_{n-k} R_{n-k} \Phi_k, \tag{4}$$

by assuming that the reproduction number was nearly constant over the past f days. (A similar assumption is made in [7].) Then we get

$$R_n^\star = \frac{i_n}{\sum_{k=1}^{f} i_{n-k}\Phi_k},$$

which is precisely (3). This being said, we also see that the estimated $R_n^\star$ is estimated at time n, but appears to be an average value over the interval $[n-f, n-1]$. Thus, although obtained on day n, it informs us on the ERN occurring roughly nine days ago, hence the "$\star$" notation to recall this hidden observation delay.

In one sentence, Equation (3) measures the ratio between the number i_n of people that are detected today as infected, to a weighted number of reported infected people who tentatively contaminated them in the past $f = 20$ days. Notice that the reported infected are only a proportion of the infected, but if this proportion is constant the ratio remains valid. This assumption of a constant ratio of reported infections should be highlighted because for COVI-19 and for other diseases there is a strong variation in reporting along days of the week. This will be another argument if favor of working with weekly averages only.

Exactly the same calculation can be made with the daily death count, by replacing the daily number i_n of reported infected persons at time n with the number of reported deaths d_n at time n, that is

$$R_{n-s}^\star := \frac{d_n}{\Phi_1 d_{n-1} + \Phi_2 d_{n-2} + \cdots + \Phi_f d_{n-f}}$$

where s is a time delay. Nevertheless, we can offer no justification for the fact that we are using the same time interval $(\Phi_k)_{k=1,\cdots f}$ for the deaths than for the incident case. This choice would be adequate if there was a fixed time between infection and death, but instead the time distribution is spread out.

2.2 The Cori et al. formula

We shall compare our method with the method proposed by Cori et al. [7]. In this article, the serial interval is called "infectivity profile". These authors assume that the instantaneous reproduction number R_t, can be estimated by the ratio of the number of new infections generated at time step t, i_t, to the total infectiousness of infected individuals at time t, given by $\sum_{k=1}^{f} i_{t-k}\Phi_k$, the sum of infection incidence up to time step $t-1$, weighted by the infectivity profile $(\Phi_k)_{k=1,\cdots f}$. Thus, R_t is the average number of secondary cases that each infected individual would infect if the conditions remained as they were at time t. But the authors also remark that it is convenient to average the estimate over a time interval with length τ. Thus at each time step t, they propose to calculate the reproduction number over a time window of size τ ending at time t and assuming that R_t is constant on the interval $[t-\tau+1, t]$. These estimates, denoted $R_{t,\tau}$, quantify the average transmissibility over a time window of length τ ending at time t. Cori et al. show that assuming a Gamma distributed prior, $\Gamma(a, b)$ with parameters a, b, for $R_{t,\tau}$, the posterior joint distribution for $R_{t,\tau}$ is the

Gamma distribution $\Gamma(a+\sum_{s=t-\tau+1}^{t} i_s, (b^{-1}+\sum_{s=t-\tau+1}^{t}\sum_{k=1}^{f} i_{s-k}\Phi_k)^{-1})$. In particular, the mean and standard deviation of this distribution are given by

$$E(R_{t,\tau}) = \frac{a+\sum_{s=t-\tau+1}^{t} i_s}{b^{-1}+\sum_{s=t-\tau+1}^{t}\sum_{k=1}^{f} i_{s-k}\Phi_k}, \tag{5}$$

$$\sigma(R_{t,\tau}) = \frac{\sqrt{a+\sum_{s=t-\tau+1}^{t} i_s}}{b^{-1}+\sum_{s=t-\tau+1}^{t}\sum_{k=1}^{f} i_{s-k}\Phi_k}, \tag{6}$$

Reorganizing the terms of the expression (5) we obtain that

$$E(R_{t,\tau}) = \frac{\frac{a}{\tau}+\frac{\sum_{s=t-\tau+1}^{t} i_s}{\tau}}{\frac{b^{-1}}{\tau}+\sum_{k=1}^{f}\left(\frac{\sum_{s=t-\tau+1}^{t} i_{s-k}}{\tau}\right)\Phi_k}, \tag{7}$$

which means that the Cori et al. estimation can be interpreted as pre-processing the data i_t using a moving average of τ days, that is

$$\hat{i}_{t,\tau} = \frac{\sum_{s=t-\tau+1}^{t} i_s}{\tau},$$

and then we compute $E(R_{t,\tau})$ using the formula

$$E(R_{t,\tau}) = \frac{\frac{a}{\tau}+\hat{i}_{t,\tau}}{\frac{b^{-1}}{\tau}+\sum_{k=1}^{f}\hat{i}_{t-k,\tau}\Phi_k}.$$

The default parameter for τ is obviously $\tau = 7$. For a and b, Cori et al. propose to use the constant values $a = 1$ and $b = 5$. We point out that the assumption of a Gamma distributed prior, $\Gamma(a, b)$, with constant parameters a, b for $R_{t,\tau}$ is valid at the beginning of the epidemic spread but when R_t starts to change in a significant way the parameters a, b should adapt to the expected value of $R_{t,\tau}$. In fact, in the interval $[t-\tau+1, t]$, a natural condition we can impose to $\Gamma(a, b)$ is that its mean satisfies:

$$E(\Gamma(a,b)) = ab = \frac{\sum_{s=t-\tau+1}^{t} i_s}{\sum_{s=t-\tau+1}^{t}\sum_{k=1}^{f} i_{s-k}\Phi_k}, \tag{8}$$

As shown in the following lemma, if the prior Gamma distribution, $\Gamma(a, b)$ of $R_{t,\tau}$, satisfies condition (8), then, the posterior joint distribution has the same mean as the prior distribution.

Lemma 1. *If the mean, ab, of the Gamma prior distribution, $\Gamma(a, b)$, for $R_{t,\tau}$ satisfies:*

$$ab = \frac{\sum_{s=t-\tau+1}^{t} i_s}{\sum_{s=t-\tau+1}^{t}\sum_{k=1}^{f} i_{s-k}\Phi_k} \tag{9}$$

Then, the posterior distribution $\Gamma(a+\sum_{s=t-\tau+1}^{t} i_s, (b^{-1}+\sum_{s=t-\tau+1}^{t}\sum_{k=1}^{f} i_{s-k}\Phi_k)^{-1})$ has the same mean that the prior distribution and its standard deviation is given by

$$\sigma(R_{t,\tau}) = \frac{\sqrt{\sum_{s=t-\tau+1}^{t} i_s}}{\sqrt{\rho+1}\sum_{s=t-\tau+1}^{t}\sum_{k=1}^{f} i_{s-k}\Phi_k}, \tag{10}$$

where

$$\rho = \frac{a}{\sum_{s=t-\tau+1}^{t} i_s} = \frac{b^{-1}}{\sum_{s=t-\tau+1}^{t}\sum_{k=1}^{f} i_{s-k}\Phi_k}. \tag{11}$$

Proof: Following (9) and (11) we obtain that

$$a = \rho \sum_{s=t-\tau+1}^{t} i_s \quad \text{and} \quad b^{-1} = \rho \sum_{s=t-\tau+1}^{t} \sum_{k=1}^{f} i_{s-k}\Phi_k$$

then the mean of the posterior distribution $\Gamma(a + \sum_{s=t-\tau+1}^{t} i_s, (b^{-1} + \sum_{s=t-\tau+1}^{t} \sum_{k=1}^{f} i_{s-k}\Phi_k)^{-1})$ is given by

$$E(R_{t,\tau}) = \frac{a + \sum_{s=t-\tau+1}^{t} i_s}{b^{-1} + \sum_{s=t-\tau+1}^{t} \sum_{k=1}^{f} i_{s-k}\Phi_k} = \frac{(\rho+1)\sum_{s=t-\tau+1}^{t} i_s}{(\rho+1)\sum_{s=t-\tau+1}^{t} \sum_{k=1}^{f} i_{s-k}\Phi_k} = ab$$

and the standard deviation is given by

$$\sigma(R_{t,\tau}) = \frac{\sqrt{(\rho+1)\sum_{s=t-\tau+1}^{t} i_s}}{(\rho+1)\sum_{s=t-\tau+1}^{t} \sum_{k=1}^{f} i_{s-k}\Phi_k} = \frac{\sqrt{\sum_{s=t-\tau+1}^{t} i_s}}{\sqrt{\rho+1}\sum_{s=t-\tau+1}^{t} \sum_{k=1}^{f} i_{s-k}\Phi_k}.$$

2.3 Filtering "administrative noise"

The raw data curves of i_n and d_n are extraordinarily noisy, and the *administrative noise* has unfortunately little to do with the Poisson noise used in all aforementioned publications. Government statistics are affected by changes of testing and polling policies, political decisions, and week-end reporting delays. Here is for example a list of explanations for the undue peaks (and even negative counts) in official death and cases statistics in France[5]:

- Start of adding death cases from Établissement d'hébergement pour personnes âgées dépendantes (EHPADs - Retirement homes) since 1 April, previously not taken into account.
- A new laboratory transmits data since May 4, retrospectively from March 16. The new number of cases in the last 24 hours takes this into account.
- The increase in cases compared to data of the previous day is an aggregation of additional data from 13th May, previously not taken into account.
- Some positive patients were counted twice, this is no longer the case, therefore the decrease in cases compared to data of the previous day.

This administrative impulse noise together with the "week-end" 7-periodic noise clearly dominate the alleged Poisson noise inherent in any counting procedure, as illustrated in Figure 4. These curves illustrate the extremely noisy character of observations, a peculiar feature of these data that we called *administrative noise*. This noisy character is observed regardless of the amount of cases. In the case of Greece it is enhanced by the very low number of observed cases, but in France which had up to 1000 times more cases daily, the SNR is similar. This illustrates the fact that administrative noise is not a Poisson noise. If it were, the SNR of France would be about 30 times larger than the SNR of Greece. Indeed, the SNR of a Poisson noise scales as the square root of the count.

So we opted for filtering the raw time series of incidence and death with a sliding median or a sliding mean over a 7 day neighborhood. This filter is applied to the daily statistics before applying Formula (3). Hence, the last estimated R_n value is obtained for the day before the penultimate day. The median is sometimes preferable to the mean, as it also filters out huge peaks caused by

[5] https://en.wikipedia.org/wiki/COVID-19_pandemic_in_France

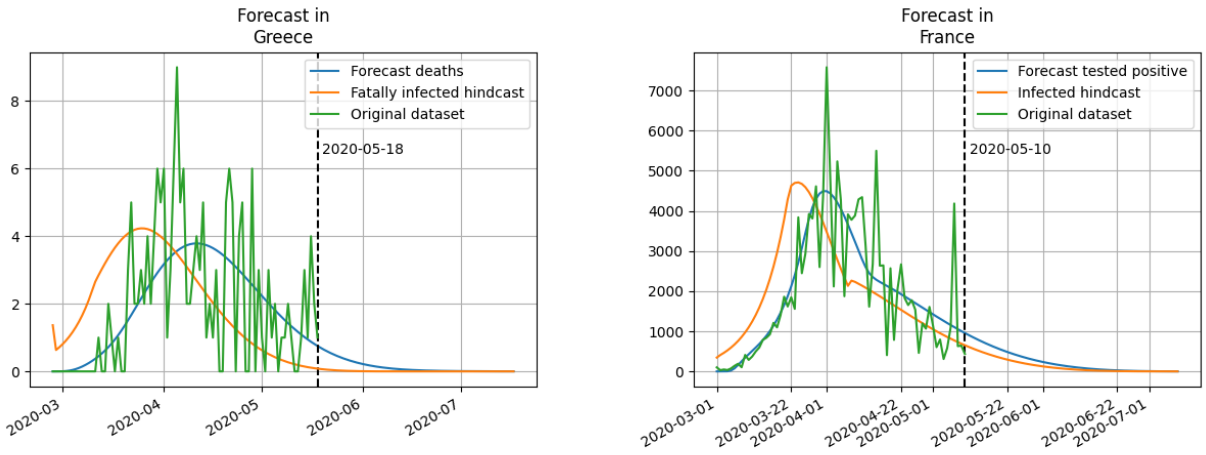

Figure 4: In green: daily count of new positive cases in Greece and France between March 1st and mid-May. These curves illustrate the *administrative noise* and the fact that it is not Poisson. The forecast and hindcast (curves in blue and orange) presented in this figure have been computed using the method proposed in [3].

impulse administrative noise. It can be objected that the median is not mean conservative. Hence, by applying a median to the time series the overall case count will be altered. This is why we leave the option of applying the mean, which has nevertheless the defect of diffusing an often considerable impulse noise.

3 Experiments

All of the experiments made here can be found in the archive of the online demo and can also be directly run online again at the present date.

We first summarize the algorithm used:

Summary of the algorithm computing ERNs

1. Input: time series i_k, $k = 1, \cdots, n$ and a serial interval function $\Phi = (\Phi_1, \cdots, \Phi_f)$.
2. Apply a median (or a mean) of seven days to the time series i_k, $k = 1, \cdots n$. The resulting series $\hat{i}_n$ therefore ends at $n - 3$ (Algorithm 5). The first three values of the series remain unfiltered.
3. Compute the ERN at time k as a ratio between $\hat{i}_k$ and a weighted average of the $\hat{i}_{k-p}$ weighted by the Φ_p (Algorithm 3 for cases and Algorithm 2 for deaths, where the only difference is that the ERN computed at time n is attributed to time $n - r$ and r is fixed by the user. Its default value is 6.) The ratio is tempered for small values of its denominator s to avoid aberrant values. More precisely, the division by s ,$\frac{1}{s}$, is replaced by the multiplication by $\frac{1}{s+\frac{5}{1+s}}$. For large s, this creates little bias and is close to $\frac{1}{s}$. For small values of s, (e.g. ≤ 10), the denominator gets a minimal value equal to 5. The obtained value of R is also capped at $R_{max} = 3.5$.
4. Compute for each k a linear regression of the ERNs in the past 20 days. Compute its RMSE on this interval and multiply it by 2. Display the confidence interval $[\text{ERN}_k - 2 \times \text{RMSE}, \text{ERN}_k + 2 \times \text{RMSE}]$ at time k (Algorithm 4).

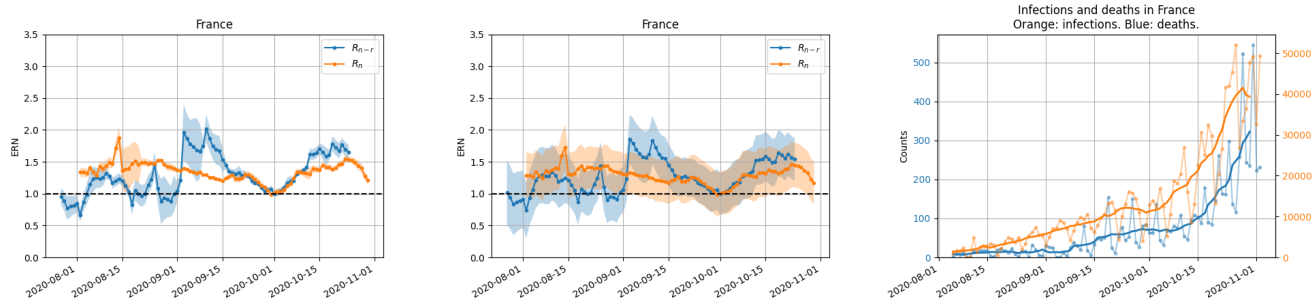

Figure 5: Comparison of our method with our interpretation of Cori et al. [7] when applying a seven days mean for both methods. Left: the ERN according to Nishiura's formula with the seven days mean option. Middle: the Cori et al. [7] method with our proposed parameters for the gamma distribution. Both ERN curves are nearly identical, as expected. Only the evaluation of their confidence interval differs. The computation was made on November 3, 2020 on France's incidence and death data (right).

5. The same procedure is applied for Cori et al.'s method, but R_n and R_{n-r} are computed with Algorithms 8 and 9, the confidence interval is computed with Algorithm 11, and the mean along τ days is the default filter. Figure 5 shows that the results are nearly identical with our proposed choice of parameters for both.

There is generally a good "fit" between the shape of the blue ERN curve obtained by the deaths and the orange ERN curve obtained by the new cases. This is illustrated in Fig. 6 showing plots of the ERNs of Italy between March 1st and June 15. On the left of this figure we display in orange $R^\star_n$ the daily reproduction number computed from the daily count of reported infected persons (orange curve, right). In blue, left, $R^\star_{n-s}$, the ERN computed from the daily count of reported deaths (right, also in blue). The time delay between both curves was fixed at $s = 6$ days to get the best fit between both curves, up to a time delay. This is equivalent to assuming that the average time delay between (reported) infections and (reported) deaths is about 6 days, which may depend on the country. The value 6 seems to be the most adequate in most western countries. However, in the case of Germany, where a policy of early screening has been practiced, this time lag sees to be larger, about $s = 13$ as shown in Figure 7.

A *caveat* about our graphic presentation is that, although $R^\star_n$ is plotted at time n because it relies on values available at time n, it **does not** indicate the actual reproduction number at time n, but rather at a time $n - p$ where p would be the average time between infection and actual detection of the infection. This time, again, may depend on the country's health system and administrative processing time, both of which can vary a lot at different times or with different countries. An empirical observation seems to indicate that a p between 6 and 8 is a most likely value for most countries, which would mean that the ERN is being observed with a 6 to 8 days delay. So the code and online demo leave the users fix p based on their own observation of the incidence and death curves.

The results on USA (Fig. 8) show that the ERN remained close to 1 for a long time (a similar pattern can be observed in Sweden and Poland). This might imply that a slightly better effort would have put the ERN neatly below 1, thus enabling a solid containment of the pandemic.

France's ERN (Fig 9) was very good after April 14, whereas it was catastrophically high on March 15. The fit of both ERNs is decent before and during lock down. After May 5, both curves diverge significantly because France changed its testing policy, thus detecting progressively many more benign

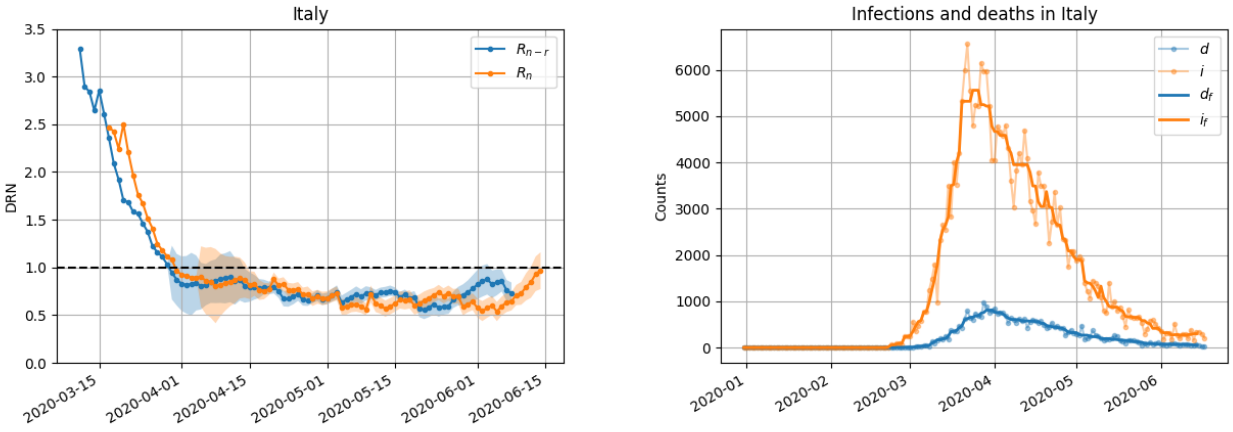

Figure 6: Plot of the ERNs of Italy from March 1st to June 15, 2020. Left: in orange $R_n^\star$ the daily reproduction number computed from the daily count of reported infected persons (orange curve, right). In blue, left, $R_{n-s}^\star$, the ERN computed from the daily count of reported deaths (right, in blue). The time delay between both curves was fixed at $s = 6$ days to get a best fit between both curves, up to a time delay. This is equivalent to assuming that the average time delay between (reported) infections and (reported) deaths is about 6 days, which may depend on countries.

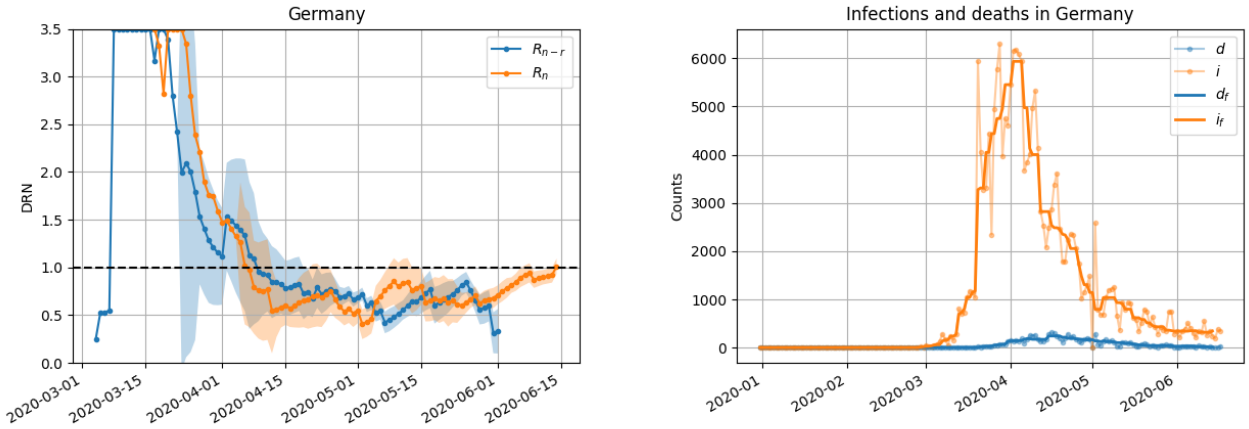

Figure 7: Plots of the ERNs for Germany with a time delay of 13 days estimated between detected cases and deaths.

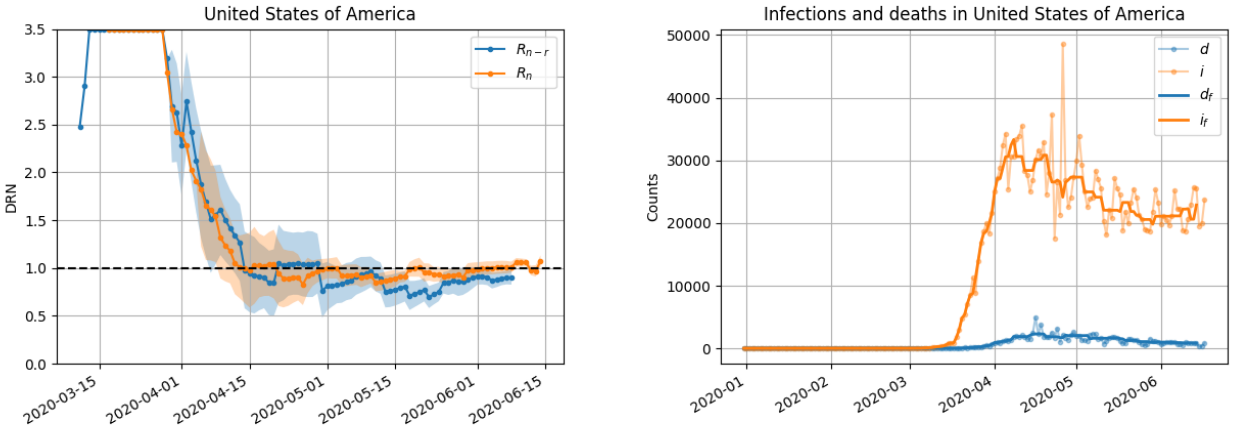

Figure 8: Plot of ERNs for the USA.

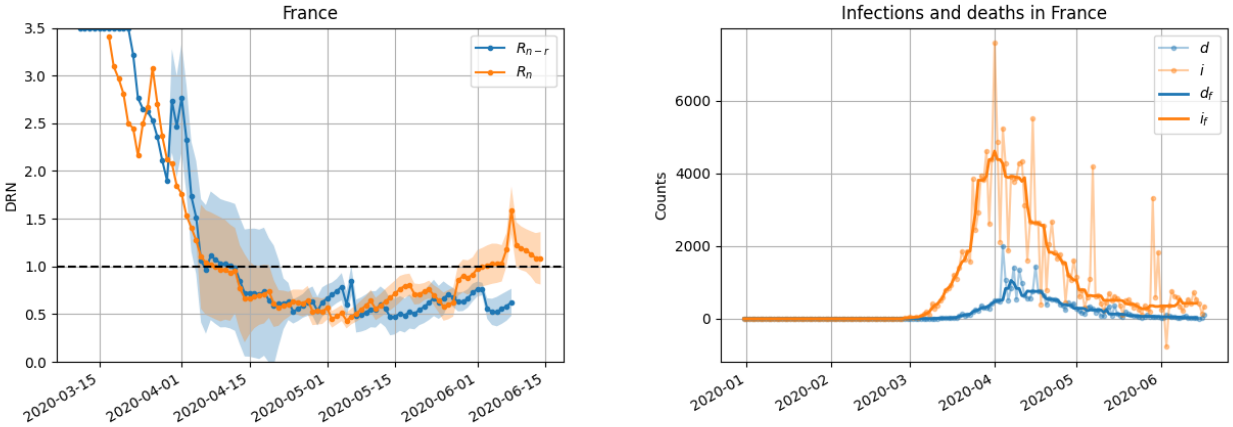

Figure 9: Plot of ERNs for France between March 1st and June 15. The discrepancy after May 15 between the ERN based on cases and the ERN based on deaths is explainable by two factors: first, the number of tests performed grew very fast in this period, thus changing the definition of a positive case. Second, the hospital treatment has arguably improved with the experience acquired.

cases. Hence, the blue curve gives a more reassuring state of affairs, given that the definition of death cannot be changed so easily by the administration.

Of course, the quality of the prediction given by these curves depends on the observational data: the earlier the tests are carried out, the shorter the laboratory delay, the shorter the administrative delay for registering infected cases or deaths, the timelier the prediction. Indeed, the delay effect of the estimation is always present: the number of positive cases measured today informs us about the number of infected people that was there several days ago, probably about 6-8 days earlier. This is why it is important to denoise these curves as much as possible, as it would allow a decent extrapolation of these curves forward. The good fit of both curves hints that merging the information from deaths and positive cases might lead to a better prediction.

The ERN can be calculated in a granular way (by country or by organized administrative unit). However, as soon as the number of cases is too small the ERN is no longer very informative because of erratic fluctuations of the counts. To illustrate this, let us look at a country with very few cases like

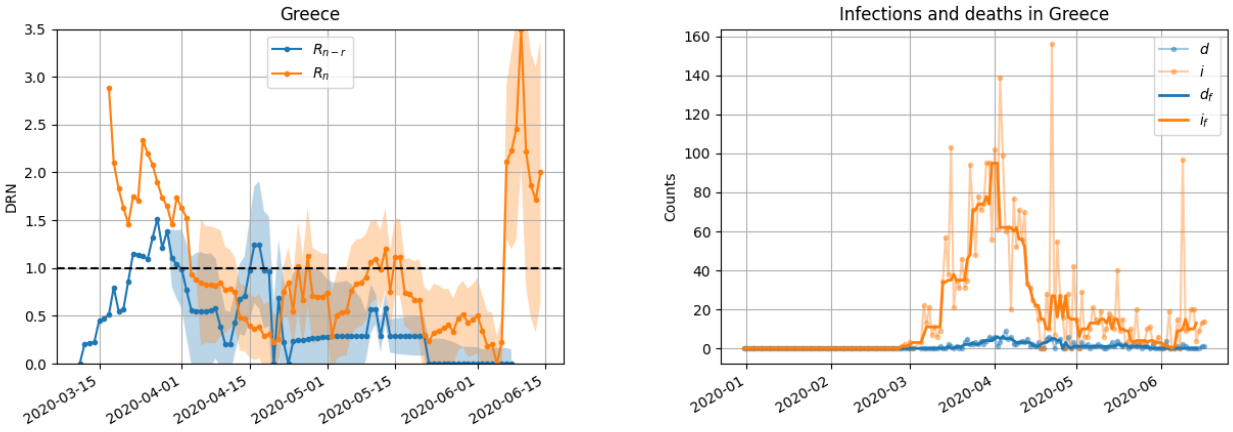

Figure 10: Plot of ERN Greece between March, 1st and June, 15. Notice the very high noise in the case data curve. It somewhat explains the oscillations of the ERN. With a very low number of cases: the ERN is sensitive to the discovery of even a small cluster.

Greece (Fig. 11). The evolution of the ERN in this graphic may seem alarming because it sometimes passes 1 for both deaths and cases. But the green curve of daily new cases in Fig. 4 shows that this daily number is tiny and oscillates between 0 and 8. Hence the oscillation of the ERN is not significant.

4 Algorithms

Algorithm 1: get_R(n, c, Φ) - Evaluate R at day n.

input : n, day index
input : c, daily cases (infections or deaths)
input : Φ, serial interval function of SARS-CoV-2 transmission
output: R evaluated at day n

$s \leftarrow \sum_{k=0}^{|\Phi|-1} \Phi[k] \times c[n-k-1]$
if $s < 1e-9$ **then**
  return 0 *// Denominator almost zero*
$R \leftarrow c[n]/\left(s + \frac{5}{1+s}\right)$ *// Tampers values of s close to zero*
$R \leftarrow \min(R, 3.5)$ *// Avoid excessive R*
return R

Algorithm 2: get_Rn_r(n, d, Φ) - Evaluate R_{n-r} at day n.

input : n, day index
input : d, daily deaths
input : Φ, serial interval function of SARS-CoV-2 transmission
output: R_{n-r} (deaths) evaluated at day n

return get_R(n, d, Φ)

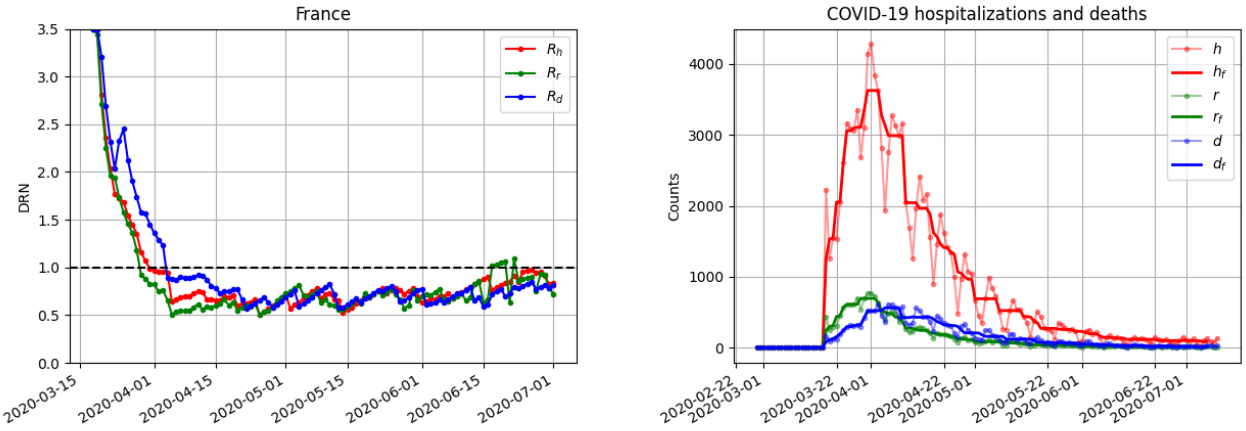

Figure 11: Plot of France's ERN between March, 1st and July, 12 computed respectively on hospital admissions (red), intensive unit admission (green), and deaths (blue). These data appear to be less subject to administrative noise and changes of policy than the incident cases (which depend on the testing policy). Notice the good agreement between all curves.

Algorithm 3: get_Rn(n, i, Φ) - Evaluate R_n at day n.

input : n, day index
input : i, daily infections
input : Φ, serial interval function of SARS-CoV-2 transmission
output: R_n (infections), evaluated at day n

return get_R(n, i, Φ)

Algorithm 4: get_conf_interval$(c,\ w = 20)$ - Computation of the confidence interval, where a linear regression is applied on the past w days including the current day.

input : c, input time series
input : $w = 20$, interval to compute the regression
output: E, confidence interval

$E \leftarrow []$
$L \leftarrow []$
$E[0, \ldots, w-1) \leftarrow 0$
for $k = w-1, \ldots, |c|-1$ **do**
  $A \leftarrow c[k-w+1, \ldots, k+1)$
  $\alpha, \beta = \text{linregress}(0, \ldots, w-1, A)$ *// Linear regression. α: slope, β: intercept.*
  $\gamma \leftarrow [\alpha + q \times \beta, q \in [0, \ldots, w)]$
  $E[k] \leftarrow 2 \times \text{std}(A - \gamma)$ *// Twice the standard deviation as confidence interval*
return E

Algorithm 5: median_filter(A, w) - Apply a w-point median filter to the input curve.

input : C, input curve
output: Filtered curve

assert $w\%2 = 0$ *// Only odd-sized kernels*
assert $|C| \geq w$ *// Input long enough*
$l_1 \leftarrow (w-1)/2$
$l_2 \leftarrow w - l_1$
$\tilde{C}_w \leftarrow \text{median}(C[k-l_1, \ldots, k+l_2-1], k \in l_1, \ldots, |C|-l_1-1)$ *// Compute median with w points*
return $C[0], \ldots C[l_1-1], \tilde{C}_w[0], \ldots, \tilde{C}_w[|C|-l_1-1]$

Algorithm 6: cori_numerator(t, c) - Numerator in Cori's R formula.

input : t, day index
input : c, daily cases (infections or deaths)
input : Φ, serial interval function of SARS-CoV-2 transmission
output: Numerator in Cori's R formula at day t

$a, \tau = 1, 7$ *// Constants*
return $a + \sum_{s=t-\tau+1}^{t} c[s]$

Algorithm 7: cori_denominator(t, c) - Denominator in Cori's R formula.

input : t, day index
input : c, daily cases (infections or deaths)
input : Φ, serial interval function of SARS-CoV-2 transmission
output: Denominator in Cori's R formula at day t

$b, \tau = 5, 7$ *// Constants*
return $1/b + \sum_{s=t-\tau+1}^{t} \sum_{k=0}^{|\Phi|-1} c[s-k] \times \Phi[k]$

Algorithm 8: get_R_Cori(t, c, Φ) - Evaluate R at day n with Cori's method.

input : t, day index
input : c, daily cases (infections or deaths)
input : Φ, serial interval function of SARS-CoV-2 transmission
output: R evaluated at day t

num = cori_numerator(t, c)
den = cori_denominator(t, c)
if $den < 1e-9$ **then**
  return 0 *// Denominator almost zero*
$R \leftarrow$ num / (den + 5/(1+den)) *// Tampers values of den close to zero*
$R \leftarrow \min(R, 3.5)$ *// Avoid excessive R*
return R

Algorithm 9: get_Rn_r_Cori(t, d, Φ) - Evaluate R_{n-r} at day n with Cori's method.

input : t, day index
input : d, daily deaths
input : Φ, serial interval function of SARS-CoV-2 transmission
output: R_{n-r} (deaths) evaluated at day t

return get_R_Cori(t, d, Φ)

Algorithm 10: get_Rn_Cori(t, i, Φ) - Evaluate R_n at day t with Cori's method.

input : t, day index
input : i, daily infections
input : Φ, serial interval function of SARS-CoV-2 transmission
output: R_n (infections), evaluated at day t

return get_R_Cori(t, i, Φ)

Algorithm 11: get_conf_interval_Cori(i) - Computation of the confidence interval of Cori's method.

input : c, input time series
output: E, confidence interval

$E \leftarrow []$
$E[0, \ldots, \tau + |\Phi| - 1) \leftarrow 0$
for $t = \tau + |\Phi| - 1, \ldots, |c| - 1$ **do**
    num = cori_numerator(t, c)
    den = cori_denominator(t, c)
    $E[t] \leftarrow \sqrt{\text{num}}/\text{den}$ *// Standard deviation as confidence interval*
return E

References

[1] *COVID-19 Projections*, 2020 (accessed June 17, 2020). https://covid19.healthdata.org/united-states-of-america.

[2] Patrice Abry, Nelly Pustelnik, Stéphane Roux, Pablo Jensen, Patrick Flandrin, Rémi Gribonval, Charles-Gérard Lucas, Eric Guichard, Pierre Borgnat, Nicolas Garnier, and Benjamin Audit, *Spatial and temporal regularization to estimate covid-19 reproduction number r(t): Promoting piecewise smoothness via convex optimization*, (2020).

[3] Luis Alvarez, Miguel Colom, and Jean-Michel Morel, *An empirical algorithm to forecast the evolution of the number of covid-19 symptomatic patients after social distancing interventions*. IPOL Journal · Image Processing On Line (preprint), 2020, https://www.ipol.im/pub/pre/301/.

[4] Tahar Boulmezaoud, *Un modèle de prédiction de l'épidémie covid-19 et une stratégie zig-zag pour la contrôler*, (2020).

[5] Fred Brauer and Gerardo Chowell, *On epidemic growth rates and the estimation of the basic reproduction number*, Notes on modeling and numerical methods. Computational modeling of biological systems, MA Morales Vazquez and S. Botello Rionda (eds.), CIMAT, (2012).

[6] Gerardo Chowell, Hiroshi Nishiura, and Luis MA Bettencourt, *Comparative estimation of the reproduction number for pandemic influenza from daily case notification data*, Journal of the Royal Society Interface, 4 (2007), pp. 155–166.

[7] Anne Cori, Neil M Ferguson, Christophe Fraser, and Simon Cauchemez, *A new framework and software to estimate time-varying reproduction numbers during epidemics*, American journal of epidemiology, 178 (2013), pp. 1505–1512.

[8] Groupe de modélisation de l'équipe ETE, *Modélisation de l'épidémie de COVID-19*, 2020 (accessed June 17).

[9] ——, *Estimation du nombre de reproduction temporel*, 2020 (accessed May 30, 2020). http://bioinfo-shiny.ird.fr:3838/Rt/.

[10] Zhanwei Du, Xiaoke Xu, Ye Wu, Lin Wang, Benjamin J Cowling, and Lauren Ancel Meyers, *The serial interval of covid-19 from publicly reported confirmed cases*, medRxiv, (2020).

[11] Youyang Gu, *COVID-19 Projections Using Machine Learning*, 2020 (accessed June 17, 2020). https://covid19-projections.com/.

[12] Quan-Hui Liu, Marco Ajelli, Alberto Aleta, Stefano Merler, Yamir Moreno, and Alessandro Vespignani, *Measurability of the epidemic reproduction number in data-driven contact networks*, Proceedings of the National Academy of Sciences, 115 (2018), pp. 12680–12685.

[13] Shujuan Ma, Jiayue Zhang, Minyan Zeng, Qingping Yun, Wei Guo, Yixiang Zheng, Shi Zhao, Maggie H Wang, and Zuyao Yang, *Epidemiological parameters of coronavirus disease 2019: a pooled analysis of publicly reported individual data of 1155 cases from seven countries*, Medrxiv, (2020).

[14] Benjamin F Maier and Dirk Brockmann, *Effective containment explains subexponential growth in recent confirmed covid-19 cases in china*, Science, 368 (2020), pp. 742–746.

[15] Gary A Mamon, *Fit of french covid-19 hospital data with different evolutionary models: regional measures of r0 before and during lockdown*, arXiv preprint arXiv:2005.06552, (2020).

[16] Hiroshi Nishiura, *Time variations in the transmissibility of pandemic influenza in prussia, germany, from 1918–19*, Theoretical Biology and Medical Modelling, 4 (2007), p. 20.

[17] Hiroshi Nishiura, Natalie M Linton, and Andrei R Akhmetzhanov, *Serial interval of novel coronavirus (covid-19) infections*, International journal of infectious diseases, (2020).

[18] Thomas Obadia, Romana Haneef, and Pierre-Yves Boëlle, *The r0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks*, BMC medical informatics and decision making, 12 (2012), p. 147.

[19] RN Thompson, JE Stockwin, Rolina D van Gaalen, JA Polonsky, ZN Kamvar, PA Demarsh, Elisabeth Dahlqwist, Siyang Li, Eve Miguel, Thibaut Jombart, et al., *Improved inference of time-varying reproduction numbers during infectious disease outbreaks*, Epidemics, 29 (2019), p. 100356.

[20] Jacco Wallinga and Peter Teunis, *Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures*, American Journal of epidemiology, 160 (2004), pp. 509–516.