medRxiv preprint doi: https://doi.org/10.1101/2020.08.01.20165142; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

A variational model for computing the effective reproduction number of SARS-CoV-2

Luis Alvarez[1], Miguel Colom[2] and Jean-Michel Morel[2]

[1] CTIM. Departamento de Informática y Sistemas, Universidad de Las Palmas de Gran Canaria. Spain

[2] Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli, F-94235, Cachan, France.

Abstract

We propose a variational model for computing the temporal effective reproduction number, $R(t)$, of SARS-CoV-2 from the daily count of incident cases and the serial interval. The $R(t)$ estimate is made through the minimization of a functional that enforces: (i) the ability to reproduce the incidence curve from $R(t)$ through a renewal equation, (ii) the regularity of $R(t)$ and (iii) the adjustment of the initial value to an initial estimate of R_0 obtained from the initial exponential growth of the epidemic. The model does not assume any statistical distribution for $R(t)$ and does not require truncating the serial interval when its distribution contains negative days. A comparative study of the solution is carried out with the standard EpiEstim method. For a particular choice of the parameters of the variational model, a good agreement is found between the estimate provided by the variational model and an estimate obtained by EpiEstim shiftef backward more than 8 days. This backward shift suggests that our model finds values for $R(t)$ that are more than 8 days closer to present. We also examine how to extrapolate $R(t)$ and the form of the incidence curve $i(t)$ in the short term. An implementation and comparison of both methods, applied every day on each country, is available at `www.ipol.im/ern`.

Keywords: COVID-19, Effective Reproduction number; reproductive rate; R0; Rt; SARS-CoV-2; Serial Interval.

Abbreviations:

EpiEstim : Software to compute the effective reproduction number proposed by Cori et al. in the paper: *A new framework and software to estimate time-varying reproduction numbers during epidemics* published in the American Journal of Epidemiology.

$R(t)$: Effective Reproduction Number. To differentiate between the continuous and discrete cases, we use the notation $R(t)$ in the continuous case and R_t in the discrete case.

$i(t)$: incidence curve, the number of daily tested positive registered. To differentiate between the continuous and discrete cases, we use the notation $i(t)$ in the continuous case and i_t in the discrete case.

Φ: serial interval.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.01.20165142; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

A key epidemiological parameter to evaluate the time varying transmission rate of a disease is the effective reproduction number $R(t)$, defined as the expected number of secondary cases produced by a primary case at each time t. The computation of an effective, or instantaneous, reproduction number is much more problematic than its global estimate, $R0$, 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. This 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. We refer to [14] for a comparison of strategies to compute $R0$ and $R(t)$.

The key ingredient of the estimation of $R(t)$ is a *renewal equation* linking the incidence $i(t)$ to $R(t)$. Here a caveat must be formulated. We shall reason as though $i(t)$ denoted the total number of new cases. But, in practice, the detected infected are only a portion of $i(t)$. Hence all formulas below rely on the assumption that the daily count of detected infected is actually proportional to the (unavailable) daily count of real infected. This assumption is actually not true, as it is strongly influenced by detection strategies. But if the detection policies evolve slowly, the arguments and calculations below remain valid.

The renewal equation requires the knowledge of the *serial interval function* $\Phi(s)$, which gives the probability distribution of the time between the onset of symptoms in a primary case and the onset of symptoms in secondary cases. This formula linking $i(t)$, $R(t)$ and $\Phi(s)$ goes back at least to Nishiura 2007 [12]. It writes

$$i(t) = \int_0^t i(t-s)R(t-s)\Phi(s)ds. \tag{1}$$

The only assumption underlying the reproduction formula (1) is that the serial interval depends only on biological factors, which is reasonable. If we assume that $R(t-s)$ is locally constant and equal to $R(t)$, that is $R(t-s) = R(t)$ for s such that $\Phi(s) > 0$, then the above expression becomes

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

This expression has been used in the literature by several authors (see [4], [5]) to estimate $R(t)$. In its stochastic Poisson formulation, it is a widely used strategy to compute $R(t)$ using (2) (see [15], [10] or [7]). In that case, the formula is given in stochastic form, assuming $i(t)$ follows a Poisson model (see [7], [14] [15]). Then the second member of (2) is taken to be the expectation of the Poisson model.

In [1] the problem of estimating $R(t)$ by maximum likelihood estimation 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 $R(t)$. One of the most widely used methods to estimate $R(t)$ is the one proposed by Cori et al. in [7]. The authors show that if the expectation of $i(t)$ is given by $\mathbf{E}[i(t)] = R(t)\sum_{s=1}^{t} i(t-s)\Phi(s)$ and $R(t)$ is assumed to follow a gamma prior distribution $\Gamma(a,b)$, then the following analytical expression can be obtained for the posterior distribution of $R(t)$:

$$R_{t,\tau} = \frac{a + \sum_{s=t-\tau+1}^{t} i_s}{b^{-1} + \sum_{s=t-\tau+1}^{t}\sum_{k=1}^{f} i_{s-k}\Phi_k}, \tag{3}$$

medRxiv preprint doi: https://doi.org/10.1101/2020.08.01.20165142; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

where R_t is assumed to be locally constant in a time window of size τ ending at time t. This method is implemented in the EpiEstim R package.

In this paper we use the renewal equation (1) rather than the simplified version (2) used by EpiEstim. One important difference between both formulations is that the estimation of $R(t)$ using the simplified model is shifted with respect to the estimation using the model (1). The reason is that if we replace $R(t-s)$ in equation (1) by a constant value, it would be more accurate to replace $R(t-s)$ by a shifted back value $R(t-\mu)$ than by $R(t)$, as t is the end of the integration interval. In other words, it would be more accurate to replace equation (2) by

$$i(t) = R(t-\mu) \int_0^t i(t-s)\Phi(s)ds. \tag{4}$$

where μ is the center of mass of the serial distribution $\Phi(s)$. So the assumption about the expectation of $i(t)$ should be $\mathbf{E}[i(t)] = R(t-\mu)\sum_{s=1}^{t} i(t-s)\Phi(s)$. Moreover, in the case of the EpiEstim estimate, given by (3), an extra shift can be expected due the assumption that R_t is locally constant in $[t-\tau, t]$. It follows that we can expect to observe a significant shift between the R_t estimate using the original model (1) and the one obtained by EpiEstim. In fact, our experiments here reveal a shift going up to 9-11 days. This shift suggests that our estimate is is closer to present than the one proposed by EpiEstim.

We now discuss what serial interval functions Φ are available for SARS-CoV-2. As we saw, the *serial interval* in epidemiology refers to the time between successive observed cases in a chain of transmission. Du et al. in [8] define this interval as follows:

> The serial interval 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 by Du et al. in [8] on 468 cases. The authors of this paper recall that *this quantity cannot be inferred from daily case count data alone [16]*. Moreover, the observed serial distribution in [8] had a significant number of cases on negative days, meaning that the infectee had developed symptoms up to 10 days before the infector.

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.

In this work, we have studied three serial intervals: the one obtained by Du et al. in [8] using 468 cases, a serial interval obtained by Nishiura et al. in [13] using 28 cases which is approximated by a log-normal distribution, and a serial interval obtained by Ma et al. in [11] using 1155 cases. As proposed by the authors this serial interval has been approximated by a shifted log-normal to take into account the cases in the negative days. In Fig. 1 we show the profile of these serial intervals. Of course, more accurate estimates of the serial interval for the SARS-CoV-2 can be expected in the future. In the online interface (`www.ipol.im/ern`) the users can, optionally, upload their own distribution for the serial interval.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.01.20165142; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

METHODS

A new variational model to compute R_t.

Equation (1) was originally formulated for serial interval functions $\Phi(s)$ satisfying $\Phi(s) = 0$ for $s \leq 0$, but this is definitely not true for the SARS-CoV-2. Hence, to avoid an artificial truncation of the serial interval function, we adopt the obvious generalization of this equation as

$$i(t) = \int_{-\infty}^{\infty} i(t-s)R(t-s)\Phi(s)ds. \tag{5}$$

In that way, the integration is performed on the whole support of the serial interval $\Phi(s)$. In practice, we deal with an incidence curve observed itself on a limited interval, up to present. Hence, boundary conditions including days in the future will be requested to apply the above formula to any point of the incidence curve. Our method requires the observation of:

- the incidence curve, namely the daily count of new detected cases of SARS-CoV-2 infections, denoted as $i_t = i(t)$ on day t.
- an empirical probability distribution $\Phi = (\Phi_{f_0}, \cdots, \Phi_f)$ for the serial interval. We assume that a patient can show symptoms up to f_0 days before the person who contaminated him/her shows symptoms himself/herself. So we have $f_0 = -4$ for the Ma et al. serial interval, $f_0 = 0$ for Nishiura et al. and $f_0 = -10$ for Du et al. The discrete support of Φ is therefore contained in the interval $[f_0, f]$.

We shall use the straightforward discretization of Equation (5),

$$i_t = F(i, R, \Phi, t) \equiv \sum_{s=f_0}^{f} i_{t-s}R_{t-s}\Phi_s \qquad \text{for} \quad t = 0, .., t_c, \tag{6}$$

where R_t represents the discrete version of $R(t)$, $t = 0$ is the time where the infection number starts to grow and t_c the current time. This equation is inserted in a variational model to estimate R_t by minimizing the energy

$$E(\{R_t\}) = \sum_{t=0}^{t_c} \left(\frac{i_t - \sum_{s=f_0}^{f} i_{t-s}R_{t-s}\Phi_s}{p_{90}(i)} \right)^2 + \sum_{t=1}^{t_c} w_t(R_t - R_{t-1})^2 + \sum_{m=0}^{M} \beta_m(R_{t_m} - \bar{R}_{t_m})^2, \tag{7}$$

where $p_{90}(i)$ is the 90th percentile of $\{i_t\}_{t=0,..,t_c}$ used to normalize the energy with respect to the size of i_t. The first term of E is a data adjustment term which forces the renewal equation (6) to be satisfied as much as possible. The second term forces R_t to be a smooth curve ; $w_t \geq 0$ represents the weight of the regularization at each time t. The higher the value of w_t the smoother R_t. The last term of E forces R_{t_m} to be close to an initial estimate given by $\bar{R}_{t_m}$ for some particular times t_m. Finally, β_m is a weight that determines the confidence we have in such initial estimate $\bar{R}_{t_m}$. The larger β_m, the greater this confidence. For instance, we can use any "a priori" estimate of R_0 as the prescribed value $\bar{R}_{t_0}$ (with $t_0 = 0$). Minimizing the energy E leads to satisfy approximately the renewal equation (6) with a reasonably smooth R_t and, optionally, prescribed initial values for some particular times t_m. The parameters w_t and β_m determine the importance assigned to these constraints in the estimation.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.01.20165142; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Minimizing E with respect to the sequence $\{R_t\}$ yields a linear system of equations that is easily solved. Yet, it requires the complement of adequate boundary conditions for R_t and i_t on both ends of the observation interval.

Definition of the boundary conditions.

For R_t, we will always assume that $R_t = R_0$ for $t < 0$ and $R_t = R_{t_c}$ for $t > t_c$ (the current time). Concerning i_t, when $t > t_c$ we use a linear regression to extrapolate the values of i_t beyond t_c. To compute the regression line ($i = m_7 \cdot t + n_7$) we use the last seven values of i_t. For $t < 0$ we will assume that the cumulative number of infected detected $I_t \equiv \sum_{k=0}^{t} i_k$ follows an exponential growth for $t < 0$, that is $I_t = I_0 e^{at}$, where a represents the initial exponential growth rate of I_t at the beginning of the infection spread. In summary, the extension of i_t beyond the observed interval $[0, t_c]$ is defined by

$$i_t = \begin{cases} I_0 e^{at} - I_0 e^{a(t-1)} & if \quad t < 0; \\ m_7 \cdot t + n_7 & if \quad t > t_c. \end{cases} \tag{8}$$

Computation of the initial exponential growth a and $\bar{R}_0$

We now naturally estimate a by

$$a = median(\{log\left(\frac{I_{t+1}}{I_t}\right) \ : \ t = 0, .., 14\}). \tag{9}$$

If we assume that $I_t = I_0 e^{at}$ follows initially an exponential growth and that R_t is initially constant, then using equation (6) we obtain that

$$i_0 = I_0(1 - e^{-a}) = I_0 R_0 \sum_{k=f_0}^{f} (e^{-ka} - e^{-(k+1)a})\Phi_k. \tag{10}$$

Hence, we can compute an approximation of R_0 as

$$\bar{R}_0 = \frac{1 - e^{-a}}{\sum_{k=f_0}^{f} (e^{-ka} - e^{-(k+1)a})\Phi_k}. \tag{11}$$

Note that this estimation strongly depends on the serial interval used. For instance, if we assume that $a = 0.250737$ (the exponential growth rate obtained in [2] when the coronavirus is in free circulation), we obtain that $\bar{R}_0 = 2.700635$ for the Nishiura et al. serial interval, $\bar{R}_0 = 3.084528$ for the Ma et al. serial interval and $\bar{R}_0 = 1.839132$ for the Du et al. serial interval.

We assumed here that we would compute R_t from the beginning of the epidemic's spread to the present day. Since the epidemic is likely to be with us for a long period of time, note that it is also possible to use our model to start calculating R_t from any ulterior time t_1, in which case the user can provide an initial value of R_{t_1}.

Management of the regularization weight w_t

medRxiv preprint doi: https://doi.org/10.1101/2020.08.01.20165142; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Let $\hat{w}_0 > 0$ be a constant value, we define w_t as

$$w_t = \hat{w}_0 \frac{G_{\sigma_w} * i(t)}{p_{90}(i)}, \tag{12}$$

where $G_{\sigma_w} * i(t)$ represents the convolution of i_t with a Gaussian kernel of standard deviation σ_w and $p_{90}(i)$ is the 90th percentile of $\{i_t\}_{t=0,..,t_c}$. Therefore, at each time t, the regularization weight is proportional to the number of cases in t (filtered using a Gaussian convolution). By default, we fix $\sigma_w = 3$ as standard deviation of the Gaussian kernel.

We noticed that when $\hat{w}_0$ is very small, minimizing the energy (7) can lead to obtain negative values R_t for some t. In such cases, we increase iteratively the value of the regularization weights at such points and at their neighbors and we compute again the minimum of (7). More precisely for any t_k such that $R_{t_k} < 0$, we update $w_{t_k} = 10w_{t_k}$, $w_{t_k+1} = 10w_{t_k+1}$, $w_{t_k-1} = 10w_{t_k-1}$ and then recompute the minimum of the energy (7). This operation is performed until all R_t's become positive. We observed experimentally that this objective is reached in a few iterations.

Filtering "administrative noise"

The raw data curve i_n is extraordinarily noisy, and the *administrative noise* has unfortunately little to do with the Poisson noise used in most 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 cases statistics in France (`https://en.wikipedia.org/wiki/COVID-19_pandemic_in_France`):

- 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.

These recording delays and consecutive rash corrections make a peculiar feature of such time series that we call *administrative noise*. They result in strong impulse noise, together with a "week-end" 7-periodic noise. These noises clearly dominate the alleged Poisson noise inherent in any counting procedure. In the web appendix, we include a discussion on different strategies to deal with this administrative noise.

Summary of the algorithm computing R_t.

- Step 1: Pre-processing of the input data (detected infected) to reduce the administrative noise and to take into account that, some countries are not providing new data during the week-end.
- Step 2: Estimation of the time t_{init} where the accumulated data, I_t, starts to grow in an exponential way. We use a very basic algorithm where we impose that $I_{t+1} > 1.1 I_t$ in two consecutive days. More formally, we set

$$t_{init} := \min_{t>2} \{t : I_t > 10 \text{ and } I_t > 1.1 I_{t-1} \text{ and } I_{t-1} > 1.1 I_{t-2}\}.$$

medRxiv preprint doi: https://doi.org/10.1101/2020.08.01.20165142; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Once t_{init} is computed the previous data sequence is removed so that t_{init} becomes 0.

- Step 3: Compute the initial exponential growth rate a and an initial estimation of R_0 using (9) and (11).
- Step 4: Computation of R_t by minimizing the energy (7). If any of the computed R_t is non-positive we increase locally the regularization weight w_t and we iterate the minimization of (7) as explained above. We always use as prescribed value for R_0 the estimation given by (11).
- Step 5: Optionally, to obtain a more robust estimate of R_t and its variability in the last three days, steps 1-4 are repeated three times: first, the complete data sequence is used, second, the last day of the sequence is eliminated and third, the last two days are eliminated. From these three estimates of R_t, a first estimate of R_{t_c} is obtained, then we recompute $R(t)$ using the full sequence but adding as prescribed value for R_{t_c} the one obtained combining the information of the last three days. We highlight that this final step 5 is optional and, it is controlled by the algorithm accordingly to the value of the parameter β_1. If $\beta_1 = 0$, then step 5 is not applied and the larger β_1, the more confidence we have in the estimate of R_{t_c} using the information of the last 3 days. (see the Web appendix for more details).

Short time extrapolation of R_t and i_t.

Since the variational method provides a point estimate of R_t up to the current time t_c, by extrapolating the value of R_t into the future, it is possible, using the renewal equation (6), to obtain also an extrapolation of the number of infected i_t in the near future. To fix reasonable future values of R_t we use a basic procedure based on Hermite interpolation polynomials and a conservative estimation of the value of R_t at $t = t_c + N_d$ in the near future. The values $R_1 \approx R(t_c + N_d)$ and N_d can be provided manually by the user or can be automatically fixed by the algorithm. By default, we fix $N_d = 7$ and $R(t_c + N_d)$ equal to the median of the R_{t_c} values estimated in the last 7 days. The interest of this extrapolation is to show to policy makers the future evolution of the incidence curve "if the reproduction conditions remain as they are currently" (see the Web appendix for technical details).

RESULTS

All of the experiments made here can be reproduced with the online interface available at `www.ipol.im/ern`. In the web appendix we include more details about the use of the online interface. This online interface allows one to assess the performance of the method applied to any country and any state in the USA, with the last date updated to the current date.

We shall pay particular attention to the comparison with EpiEstim, the method proposed by Cori et al. in [7] that we have explained briefly in the introduction. It is one of the most widespread methods to estimate $R(t)$. In our comparative experiments, we use the following parameters for the EpiEstim method: a 7-day time interval, that is $\tau = 7$, and $a = 1$, $b = 5$ for the prior Gamma distribution $\Gamma(a, b)$. EpiEstim does not allow for a serial interval distribution

medRxiv preprint doi: https://doi.org/10.1101/2020.08.01.20165142; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

with positive values at 0 or on negative days, therefore to compute the EpiEstim estimate given by (3), all the values Φ_s for $s < 1$ are accumulated in the value Φ_1.

We compared the results obtained by our variational method and the ones obtained by EpiEstim for four countries: France, the United Kingdom, Spain and the United States. We used the incidence curves reported by the countries up to October 30, 2020 that we obtained from the European Centre for Disease Prevention and Control service. For several weeks in this period, France and Spain did not provide data in the weekends and gave instead a cumulative count of three consecutive days on Mondays. To avoid the artificial noise generated, the accumulated value of the three days divided by 3 was assigned to Saturday, Sunday and Monday.

For the parameters of the variational model, we used the serial interval proposed by Ma et al. in [11] and we set $\hat{w}_0 = \sqrt{10}$ (the regularization weight), $\beta_0 = 10^5$ (the confidence assigned to the value R_0 estimated as explained above) and $\beta_1 = \sqrt{10}$ (the confidence assigned to the value R_{t_c} estimated from the last three days).

In Fig. 2 the R_t estimate obtained by both methods are compared. We shifted back the EpiEstim results by 9 days to fit the results obtained by the variational model. Surprisingly, despite the fact that both methods are quite different, a good fit of both estimates in the four countries was observed. To measure the displacement, $\tilde{t}$, between both estimates we fit both curves by minimizing their RMSE,

$$\tilde{t} = \arg\min_{t\in[0,20]} D(R, R^e, t) \equiv \sqrt{\frac{\sum_{k=t_c-T+1}^{t_c}(R(k-t) - R^e(k))^2}{T}} \tag{13}$$

where R^e is the EpiEstim estimate and $T = 120$. Notice that, in the above expression, t is not, in general, an integer value. So to evaluate $R(k-t)$ we use linear interpolation. In table 1 we present the results for several countries. It is observed that the displacement between both estimates is between 9 and 11 days and that the fit between both curves is quite good with a distance between 0.019 and 0.092. The shift between 9 and 11 days is not surprising. Indeed, as stated in equation (4), we can expect a first shift with a magnitude around the mean of the serial interval (in the case of the used Ma et al. serial interval, the mean is 6.7). Moreover, using a time window of 7-days adds an additional 3.5 days of delay, which gives a global shift around 10 days with respect to the present. Hence the observed time shift comprised between 9 and 11 days is explainable.

In Fig. 3 we present the sequence of the daily number of detected infected, i_t, used in the experiments as well as its expected value using the renewal equation $F(i, R, \Phi, t)$ defined if (6) using the R_t estimate obtained by the variational model. It suggests that $F(i, R, \Phi, t)$ is an excellent smooth approximation of i_t.

In Fig. 4 we show the results, in the case of France, of the automatic procedure to obtain a 7-day extrapolation of the value of R_t and i_t. Once R_t is extrapolated, the subsequent "conditional forecast" of i_t is obtained using the renewal equation $F(i, R, \Phi, t)$. To apply the formula to obtain $F(i, R, \Phi, t)$, the required values of i_t beyond the current time are obtained in an iterative way using the extrapolation procedure defined in (8).

medRxiv preprint doi: https://doi.org/10.1101/2020.08.01.20165142; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

DISCUSSION

In this paper we proposed a variational model given by the expression (7) for computing the effective reproduction number R_t of SARS-CoV-2 using the daily registered infected and the serial interval. The main advantages we found with this method are:

- It is based on a known epidemiological model (given by the renewal equation (1)) which establishes how $R(t)$ and the serial interval intervene in the evolution of the number of incident cases. The method does not involve the "naive" simplification which assumes $R(t)$ to be locally constant. Our method does not assume that the observation noise is Poisson, because it plainly isn't.
- The method can use serial intervals with distributions containing negative days (as it is the case for the SARS-CoV-2). Thus, it avoids an artificial truncation of the distribution.
- The method computes a point estimate of R_t up to the current date. It seems to provide a more to date (by more than 8 days) estimate of R_t than EpiEstim, which is based on the model (2) and a time interval estimation.
- The method does not assume any statistical distribution for R_t. The main assumptions are that the R_t estimate should follow the renewal equation (1) but keeping R_t regular enough. We include this regularity hypothesis in the model using standard techniques of calculus of variations.

We have included an automatic procedure in the algorithm to deal with countries where no data is provided during the week-end, and, optionally, a pre-processing step of the data using classic filters.

While our method and the standard Cori et al. (EpiEstim) method are quite different, we found experimentally that for a particular choice of the parameters of the variational method, a good agreement can been obtained between the estimate of R_t provided by the variational model and a back shifted estimate of R_t obtained by EpiEstim.

Since the point estimation of R_t is obtained up to the current date, it allows us to extrapolate the value of R_t in the short term. In countries that introduce mild social distancing measures, it is hard to obtain an accurate forecast because the situation of the epidemic can change rapidly in any direction in a few days. Therefore, the extrapolation technique proposed in this paper is just an example to show how a recent R_t estimation can be used to predict the number of infected in the next days. Using Hermite interpolation polynomials we proposed two techniques to extrapolate the value of R_t. In the first one, the user supplies a future target value for R_t and the number of days needed to reach that value. This target value can depend on the actions that a country is taking to control the epidemic. The second technique is automatic and it provides an extrapolation in the very short term (one week). In this case we fix the expected value of R_t in one week to be the median of the R_{t_c} values obtained the last 7 days. This gives policy makers an evaluation of the incidence curve if everything remains equal.

We finally notice how dependent any estimation or forecast of R_t is on governement policies, both for gathering data and for recording them. The main hindrances to a precise estimate of R_t are :

medRxiv preprint doi: https://doi.org/10.1101/2020.08.01.20165142; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

a) the constant changes of detection policies, which can go from a minimal count of serious cases confirmed at hospitals to wide ranging random testing;

b) the incredible incapacity of administrations to record cases on a daily base, in the era of internet and instant communication.

An online implementation of the method is available at `www.ipol.im/ern` where the users can perform their own experiments using official registered data of infected or uploading their own data of daily infected and/or the serial interval distribution.

References

[1] P. Abry, N. Pustelnik, S. Roux, P. Jensen, P. Flandrin, R. Gribonval, C.-G. Lucas, E. Guichard, P. Borgnat, N. Garnier, and B. Audit, *Spatial and temporal regularization to estimate COVID-19 reproduction number R(t): Promoting piecewise smoothness via convex optimization.* medRxiv, 2020.

[2] L. Alvarez, M. Colom, and J.-M. 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/`.

[3] L. Alvarez and J.-M. Morel, *Removing weekly administrative noise in the daily number of COVID-19 new cases and applications to the computation of Rt*, MedRxiv, (2020).

[4] T. Boulmezaoud, *Un modèle de prédiction de l'épidémie Covid-19 et une stratégie zigzag pour la contrôler*, (2020).

[5] T. Z. Boulmezaoud, L. Alvarez, M. Colom, and J.-M. Morel, *A daily measure of the SARS-CoV-2 daily reproduction number for all countries*, IPOL Journal. Image Processing On Line, submitted, (2020).

[6] G. Chowell, H. Nishiura, and L. M. 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] A. Cori, N. M. Ferguson, C. Fraser, and S. Cauchemez, *A new framework and software to estimate time-varying reproduction numbers during epidemics*, American journal of epidemiology, 178 (2013), pp. 1505–1512.

[8] Z. Du, X. Xu, Y. Wu, L. Wang, B. J. Cowling, and L. A. Meyers, *The serial interval of COVID-19 from publicly reported confirmed cases*, medRxiv, (2020).

[9] Groupe de modélisation de l'équipe ETE (Laboratoire MIVEGEC, CNRS, IRD, Université de Montpellier), *Estimation du nombre de reproduction temporel*, 2020 (accessed May 30, 2020). `https://bioinfo-shiny.ird.fr/Rt/`.

[10] Q.-H. Liu, M. Ajelli, A. Aleta, S. Merler, Y. Moreno, and A. Vespignani, *Measurability of the epidemic reproduction number in data-driven contact networks*, Proceedings of the National Academy of Sciences, 115 (2018), pp. 12680–12685.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.01.20165142; this version posted November 16, 2020. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

[11] S. Ma, J. Zhang, M. Zeng, Q. Yun, W. Guo, Y. Zheng, S. Zhao, M. H. Wang, and Z. Yang, *Epidemiological parameters of coronavirus disease 2019: a pooled analysis of publicly reported individual data of 1155 cases from seven countries*, Medrxiv, (2020).

[12] H. Nishiura, *Time variations in the transmissibility of pandemic influenza in Prussia, Germany, from 1918–19*, Theoretical Biology and Medical Modelling, 4 (2007), p. 20.

[13] H. Nishiura, N. M. Linton, and A. R. Akhmetzhanov, *Serial interval of novel coronavirus (COVID-19) infections*, International journal of infectious diseases, (2020).

[14] T. Obadia, R. Haneef, and P.-Y. Boëlle, *The r0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks*, BMC medical informatics and decision making, 12 (2012), p. 147.

[15] R. Thompson, J. Stockwin, R. D. van Gaalen, J. Polonsky, Z. Kamvar, P. Demarsh, E. Dahlqwist, S. Li, E. Miguel, T. Jombart, et al., *Improved inference of time-varying reproduction numbers during infectious disease outbreaks*, Epidemics, 29 (2019), p. 100356.

[16] M. A. Vink, M. C. J. Bootsma, and J. Wallinga, *Serial intervals of respiratory infectious diseases: a systematic review and analysis*, American journal of epidemiology, 180 (2014), pp. 865–875.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.01.20165142; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

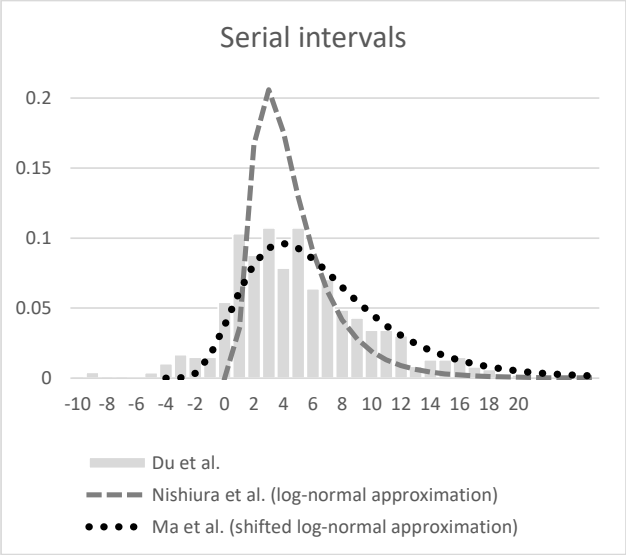

Figure 1: Serial intervals used in our experiments: the discrete one proposed by Du et al. in [8] (solid bars), a log-normal approximation of the serial interval proposed by Nishiura et al. in [13] (dashed line) and a shifted log-normal approximation of the serial interval proposed by Ma et al. in [11] (dotted line).

	France	United Kingdom	Spain	USA
shift ($\tilde{t}$)	9.60	9.46	10.18	9.33
$D(R, R^e, \tilde{t})$	0.073807	0.091506	0.081461	0.019639

Table 1: Shift between the R_t estimates using EpiEstim and the variational method obtained by minimizing the distance $D(R, R^e, t)$.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.01.20165142; this version posted November 16, 2020. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

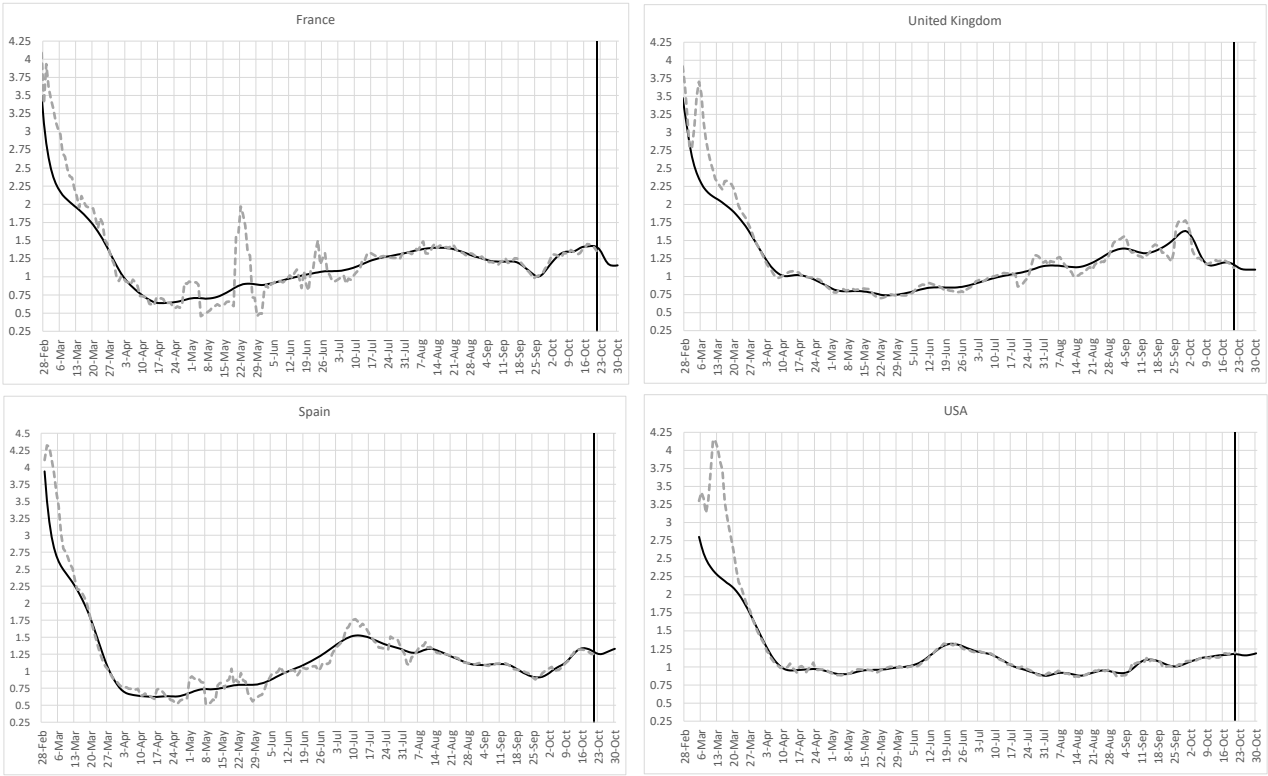

Figure 2: we show the R_t estimate obtained by EpiEstim (dotted line) and the variational model (solid line). The vertical line represents the 9-day shift applied to the EspiEstim estimate to fit the results of the variational technique.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.01.20165142; this version posted November 16, 2020. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

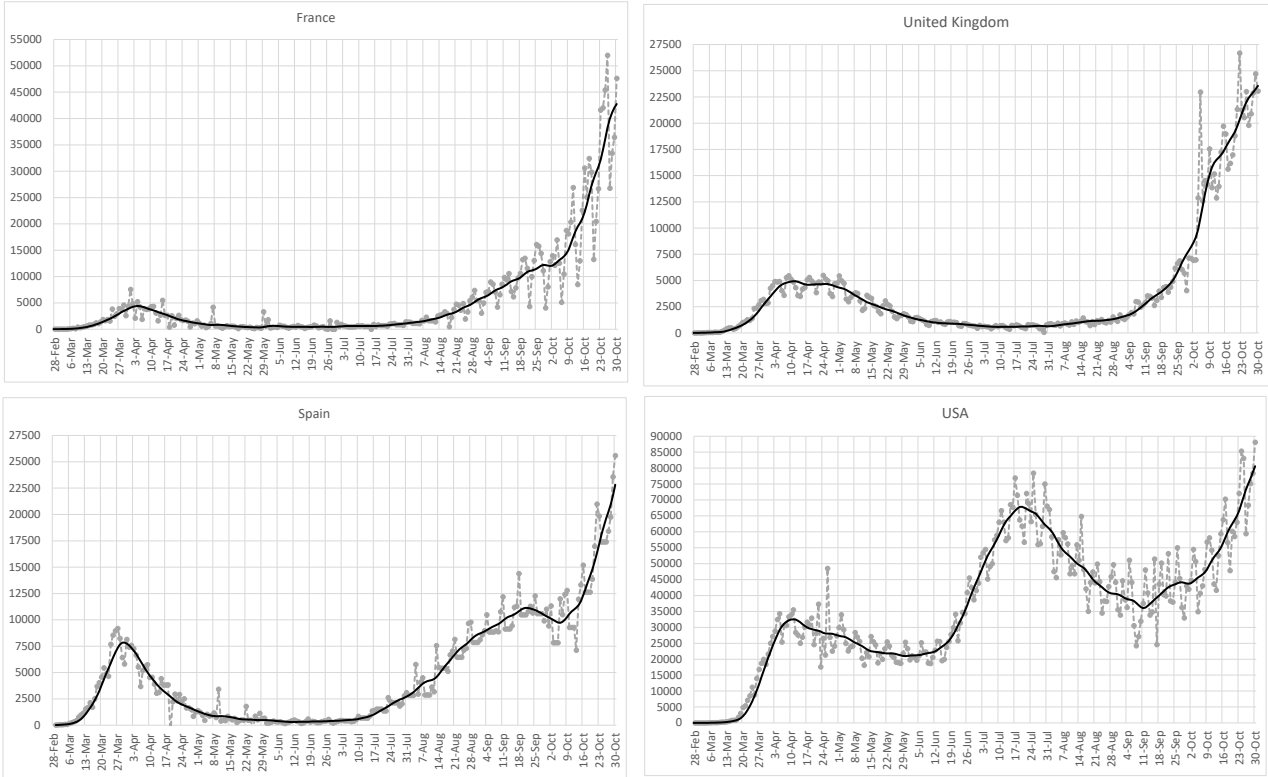

Figure 3: we show the registered daily number of new infected patients (dotted line) and its expected value using the renewal equation $F(i, R, \Phi, t)$ defined in (6) (solid line) using the R_t estimate obtained by the variational model.

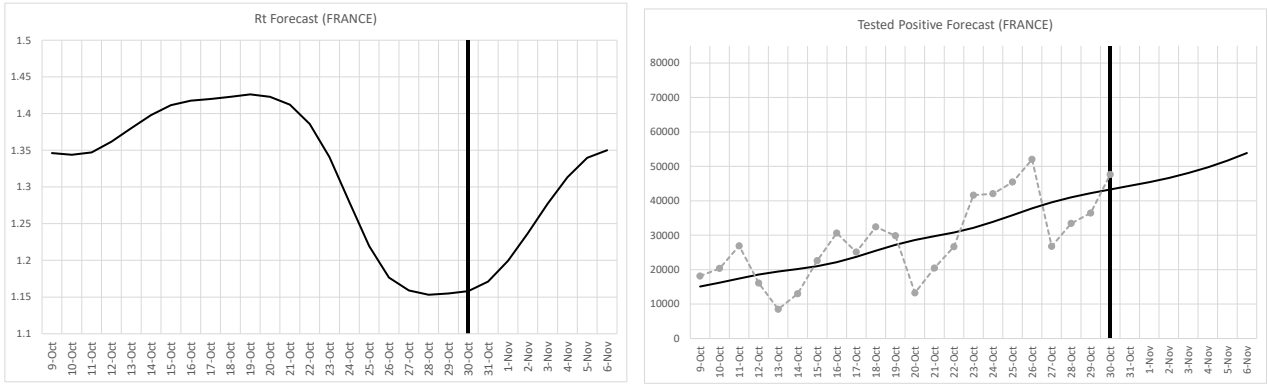

Figure 4: In the case of France, we show a 7-day forecast of R_t and i_t using as expected value of $R(t_c + 7)$ the median of the estimation of R_{t_c} in the last 7 days. The vertical line indicates the time when the forecast starts. On the left the plot the value of R_t and on the right we plot the value of i_t (dotted line) and its forecast (solid line) from the renewal equation $F(i, R, \Phi, t)$.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.01.20165142; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Web Appendix. Some technical issues about the variational method to compute R_t

Pre-processing of the input data

Weekly administrative noise

The algorithm first checks if the country did not provide data in some weekends during the period under analysis. In that case, from a given date in the past, there are zero values for the incident cases on Saturdays and Sundays, and cumulative values of three days on Mondays. To identify this pattern the algorithm checks if for some past time $t = M$ (the Monday) one has $i_{M-3} > 0$, $i_{M-2} = i_{M-1} = 0$ and $i_M > 1.5 \cdot i_{M-3}$. If that pattern is confirmed, are marked as Saturday and Sunday with no data all $t = M - 7 \cdot k - 1$, $t = M - 7 \cdot k - 2$ (with $k = 0, 1, 2..$) while $i_{M-7\cdot k-2} = i_{M-7\cdot k-1} = 0$. Let us denote by W_e the identified set of Saturdays and Sundays with no data. In the affected week-ends the incidence is set to $i_{M-7\cdot k-2} = i_{M-7\cdot k-1} = i_{M-7\cdot k} = i_{M-7\cdot k}/3$. In short, a third of the cumulative incident cases over the three days are assigned to Saturday, Sunday and Monday.

Another optional strategy proposed in [3] to reduce the weekly administrative noise is based on the correction of the value of the number of infected according to the day of the week to obtain a better fit between the number of infected and the one expected using the renewal equation $F(i, R, \Phi_s, t)$. In practice the method is very simple: one just has to multiply the number of infected by a factor which depends on the day of the week. See [3] for more details.

Some classic pre-processing filters

Several classic noise elimination filters are implemented in the online interface as optional pre-processing steps of the incidence curve. Yet, in general, the application of these filters is unnecessary as the variational model, together with the weekly administrative noise removal, is sufficient to obtain good results.

The first proposed filter is the median filter. Given a windows radius r_W we define the median filter in a time $t \leq t_c - r_W$ as

$$M(t) = median(\{i_n \ : \ |n - t| \leq r_W \ and \ n \notin W_e\})$$

We also implemented a linear Gaussian convolution filter. The general expression to compute the convolution with a symmetric kernel such as the Gaussian function is:

$$G * i(t) = g_0 \cdot i(t) + \sum_{s=1}^{N_G} g_s \cdot (i(t+s) + i(t-s)), \tag{14}$$

where the coefficients, g_s, computed from the Gaussian function, satisfy:

$$g_0 + 2 \sum_{s=1}^{N_G} g_s = 1. \tag{15}$$

medRxiv preprint doi: https://doi.org/10.1101/2020.08.01.20165142; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

In the expression (14), when $t+s > t_c$ we use a linear extrapolation to compute the value of $i(t+s)$. To approximate $i'(t_c)$, the derivative of $i(t)$ in the current time t_c, we use a weighted average of derivatives given by:

$$i'(t_c) = \frac{g_0 \cdot (i(t_c) - i(t_c - 1)) + \sum_{s=1}^{N_G} g_s \cdot (i(t_c - s) - i(t_c - s - 1))}{g_0 + \sum_{s=1}^{N_G} g_s}, \tag{16}$$

and then, for $t+s > t_c$ we define $i(t+s)$ as

$$i(t+s) = i(t_c) + i'(t_c)(t + s - t_c) \tag{17}$$

Finally we also used an optional moving average filter as a data pre-processing step.

Short time extrapolation of R_t and i_t

Extrapolating R_t.

To extrapolate the value of R_t we use a basic procedure based on Hermite interpolation polynomials and the knowledge of the expected value of R_t in a given day, $t_c + N_d$, in the near future. Assuming that the values N_d and $R_1 \approx R(t_c + N_d)$ are given, we use the following Hermite interpolation polynomial to extrapolate the value of R_t:

$$\tilde{R}(t) = \begin{cases} R_1 + R'(t_c)\frac{\gamma N_d}{8}(2\frac{t-t_c}{\gamma N_d} - 2)^2(2\frac{t-t_c}{N_d}) + & \\ (R(t_c) - R_1)\frac{1}{4}(2\frac{t-t_c}{N_d} - 2)^2(2\frac{t-t_c}{N_d} + 1) & if \quad t \in (t_c, t_c + \gamma N_d] \\ R_1 + (R(t_c) - R_1)\frac{1}{4}(2\frac{t-t_c}{N_d} - 2)^2(2\frac{t-t_c}{N_d} + 1) & if \quad t \in (t_c + \gamma N_d, t_c + N_d] \\ R_1 & t > t_c + N_d, \end{cases}$$

for any $\gamma \in (0, 1]$, $\tilde{R} \in C^1(t_c, \infty)$ satisfies that $\tilde{R}(t_c) = R(t_c)$, $\tilde{R}'(t_c) = R'(t_c)$, $\tilde{R}(t_c + N_d) = R_1$, $\tilde{R}'(t_c + N_d) = 0$. So, using this basic extrapolation procedure we get a smooth transition between $R(t_c)$ and R_1. By default, we initialize $\gamma = 1$ and then reduce the value of γ automatically to avoid that $\tilde{R}(t)$ has negative value using the following relation:

$$\tilde{R}(t) \geq \min\{R_1, R(t_c)\} + R'(t_c)\frac{4}{27}\gamma N_d \qquad \forall t \in [t_c, \infty)$$

therefore if $R_1, R(t_c) > 0$ and $\gamma \in (0, 1]$ is small enough, then $\tilde{R}(t) > 0$ $\forall t \in [t_c, \infty)$.

The values of R_1 and N_d can be fixed manually accordingly with the current and/or expected social distancing measures. A default automatic procedure for these parameters is to fix $N_d = 7$ and $R(t_c + N_d)$ equal to the median of the R_{t_c} values estimated in the last 7 days. To do so, we apply the variational model 7 times, removing each time the last value of the remaining data sequence (that is, we update each time t_c by $t_c - 1$), then we compute the median of the obtained R_{t_c} values.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.01.20165142; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Extrapolation of the incidence curve i_t.

Once w R_t has been extrapolated for $t = t_c + 1, t_c + 2, .., t_c + dT$ (where dT is the number of future days to extrapolate), we can compute by iteration i_t from $t_c + 1$ to $t_c + dT$ using the renewal equation $F(i, R, \Phi, t)$, the required values of i_t beyond the current time are obtained using the extrapolation procedure defined in (8). Then to reduce the effect of the extrapolation of i_t when $t > t_c$, we apply a second time $F(i, R, \Phi, t)$ to the obtained extrapolated sequence for $t = t_c - 9, ..., t_c + dT$. We start at $t_c - 9$ to include an extra smooth of the first estimate of $F(i, R, \Phi, t)$ when t approaches t_c.

Managing the variability of the R_t estimate.

Since we did not assume any statistical model on the distribution of the R_t values, no confidence interval for this estimate is available. For certain choices of the variational model parameters, the good agreement with the results obtained by EpiEstim gives us an idea, by comparison, about the variability of our estimate. To reduce the variability of the R_t estimate in the last days we use the following procedure:

1. Compute $R^0(t)$ by minimizing (7) for $t \in \{t_0, .., t_c\}$ with $M = 0$ (that is without any restriction on the value of $R(t_c)$).
2. Compute $R^{-1}(t)$ by minimizing (7) for $t \in \{t_0, .., t_c - 1\}$ with $M = 0$, hence removing the last value of the data sequence.
3. Compute $R^{-2}(t)$ by minimizing (7) for $t \in \{t_0, .., t_c - 2\}$ with $M = 0$, hence removing the last two values of the data sequence.
4. Fix $M = 1$ in (7), $t_1 = t_c$ and
$$\bar{R}_1 = \frac{R^0(t_c) + R^{-1}(t_c) + R^{-2}(t_c)}{3}$$
to compute $R^{-1}(t_c)$ and $R^{-2}(t_c)$ we use linear extrapolation.
5. Compute $R(t)$ by minimizing (7) for $t \in \{t_0, .., t_c\}$ with $M = 1$ using $\bar{R}_1$ as initial estimate of $R(t_c)$.

This procedure stabilizes the estimate of $R(t_c)$ with respect to the estimation in the last three days. In addition, it allows us to calculate a measure of the variation of $R(t)$ estimate in the last three days using the expression

$$sigma(t) = \sqrt{\frac{(R(t) - R^0(t))^2 + (R(t) - R^{-1}(t))^2 + (R(t) - R^{-2}(t))^2}{3}} \tag{18}$$

To illustrate this variability, in the software available online at `www.ipol.im/ern`, we represent the estimate of R_t around an empirical interval of variability defined at each point as $[R_t - 2 \cdot sigma(t), R_t + 2 \cdot sigma(t)]$.

medRxiv preprint doi: https://doi.org/10.1101/2020.08.01.20165142; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Technical details of the EpiEstim estimate

As shown in [5], assuming, for $R_{t,\tau}$, a Gamma distributed prior, $\Gamma(a, b)$ with parameters a, b, over a time window of length τ, the EpiEstim estimate of R_t can be expressed as

$$R_{t,\tau} = \frac{\frac{a}{\tau} + \bar{i}_{t,\tau}}{\frac{b-1}{\tau} + \sum_{k=1}^{f} \bar{i}_{t-k,\tau}\Phi_k} \tag{19}$$

where $\bar{i}_{t,\tau}$ is the moving average of i_t in a time window of length τ, that is

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

therefore the EpiEstim estimation can be obtained by filtering the data using a moving average and then applying equation (19). Moreover, if we assume that

$$ab = \frac{\sum_{s=t-\tau+1}^{t} i_s}{\sum_{s=t-\tau+1}^{t} \sum_{k=1}^{f} i_{s-k}\Phi_k}, \tag{20}$$

we obtain (see [5]) that equation (19) becomes

$$R_{t,\tau} = \frac{\bar{i}_{t,\tau}}{\sum_{k=1}^{f} \bar{i}_{t-k,\tau}\Phi_k} \tag{21}$$

which corresponds to the usual R_t estimate obtained directly from equation (2) (but pre-processing the input data first using a moving average).

Using the online interface in `www.ipol.im/ern`

Summary of algorithm parameters.

- i_t: input data with the daily registered infected curve.
- Serial interval used: by default we propose three options: the serial intervals obtained by Ma et al., by Nishiura et al. and by Du et al.. The users can also upload their own serial interval.
- Parameters in the energy (7):
  - $\hat{w}_0$: regularization weight. The default value is $\hat{w}_0 = 10^{0.7}$.
  - β_0: weight for the initial estimation of R_0 computed using (11). The fixed value is $\beta_0 = 10^5$. This parameter is not in the online interface because it has not any significant influence in the last values of $R(t)$.
  - β_1: weight in the energy (7) for the initial estimation of $R(t_c)$ computed using the estimate in the last 3 days. The default value is $\beta_1 = 10^{0.5}$.
- Optional Data Pre-filtering:

medRxiv preprint doi: https://doi.org/10.1101/2020.08.01.20165142; this version posted November 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

  - r_W: radius of the median filter window for data filtering. If the value of this parameter is zero no median filter is applied. The default value is 0.
  - σ: standard deviation of the Gaussian linear filter for data filtering. If the value of this parameter is zero no Gaussian filter is applied. The default value is 0.
  - r_M: radius of the moving average window filter. If the value of this parameter is zero no moving average is applied. The default value is 0.

- Forecasting (in the case of user interactive forecasting):
  - R_1: expected value of $R(t_c + N_d)$ for forecasting.
  - N_d: Number of days to reach the value R_1.

In the online interface our variational R_t estimate is compared with the one obtained by EpiEstim using a 7-day time window and $a = 1$ and $b = 5$ for the prior Gamma distribution. The EpiEstim estimate is shifted backwards to fit our estimate. We also show a plot of the initial sequence i_t and, for comparison, in the case a pre-filtering is applied to the data, we plot the result of the filtered sequence. In the case where no pre-filtering is applied we plot the result of the application of the formula (6) to the data sequence i_t with the estimated R_t. That is, we plot:

$$\tilde{i}_t = \sum_{s=f_0}^{f} i_{t-s} R_{t-s} \Phi_s \qquad \text{for} \quad t = 0, .., t_c, \tag{22}$$

We observe that due to the regularization included in the estimation of R_t, $\tilde{i}_t$ is an smoothed version of i_t.

If the option "Remove weekly administrative noise" is activated, we use the method proposed in [3] to remove this administrative noise.