# 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 effective reproduction number (ERN) of SARS-CoV-2 from the daily count of incident cases and the serial interval. The ERN estimate is made through the minimization of a functional that includes: (i) the adjustment of the incidence curve using an epidemiological model, (ii) the regularity of the estimation of the ERN 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 the ERN and does not require truncating the serial interval when its distribution contains negative days. A comparative study has been carried out with the standard EpiEstim method. For a particular choice of the parameters of the variational model and of the serial interval, a good agreement has been obtained between the estimate provided by the variational model and a shifted estimate obtained by EpiEstim. This backward shift suggests that our estimate is closer to present than that of EpiEstim. We also examine how to forecast the value of the ERN and the number of infected in the short term. An implementation of the model 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.

ERN : Effective Reproduction Number.

A key epidemiological parameter to evaluate the time varying transmission rate of a disease is the effective reproduction number (ERN), (also denoted in this paper by $R_t$ or $R(t)$), 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 [4] 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 [12] for a comparison of strategies to compute $R0$ and $R_t$.

The key ingredient of the estimation of $R(t)$ is a reproduction formula 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 reproduction formula requires the 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. This formula linking $i(t)$, $R(t)$ and $\Phi(s)$ goes back at least to Nishiura 2007 [10]. 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 $R(t-s) \equiv R(t)$ 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 to estimate $R(t)$ [2], [3]. In

its stochastic Poisson formulation, it is a widely used strategy to compute $R(t)$ using (2) (see [13], [8] or [5]). In that case, the formula is given in stochastic form, assuming $i(t)$ follows a Poisson model (see [5], [12] [13]). 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 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 $R(t)$. One of the most widely used methods to estimate $R(t)$ is the one proposed by Cori et al. in [5]. 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, then an analytical expression can be obtained for the posterior distribution of $R(t)$. To obtain a more regular estimate they compute $R(t)$ in a time window of size $\tau$ ending at time $t$, assuming that $R(t)$ is locally constant is that window. This method is implemented in the EpiEstim R package and it is also provided as a Microsoft Excel spreadsheet.

In this paper we use the epidemiological model (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

$$i(t) = R(t-\mu)\int_0^t i(t-s)\Phi(s)ds. \tag{3}$$

where $\mu$ is the center of mass of the serial distribution $\Phi(s)$. So the assumption about the expectation of $i(t)$ should be $\mathbf{E}[i(t)] = R(t-\mu)\sum_{s=1}^{t} i(t-s)\Phi(s)$. It follows that we can expect to observe a shift between an $R_t$ estimate using the original model (1) and the one obtained using its simplification (2). Our experiments here reveal a clear shift between our estimation and the one obtained by EpiEstim, going up to seven days. This shift suggests that our estimate is more to date than the one proposed by EpiEstim.

We now discuss what serial interval functions $\Phi$ 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 [6] 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 [6] on 468 cases. The authors of this paper recall that *this quantity cannot be inferred from daily case count data alone [14].* Moreover, the observed serial distribution in [6] had a significant number of cases on negative days, meaning that the infectee had developed symptoms up to 10 days before the infector.

In [7], 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 the experiments presented in this paper we shall use three serial intervals: the one obtained by Du et al. in [6] using 468 cases, a serial interval obtained by Nishiura et al. in [11] using 28 cases which is approximated by a log-normal distribution, and a serial interval obtained by Ma et al. in [9] 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.

# 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

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

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 $\Phi$ is therefore contained in the interval $[f_0, f]$.

We shall use the straightforward discretization of Equation (4),

$$i_t = \sum_{s=f_0}^{f} i_{t-s}R_{t-s}\Phi_s \qquad \text{for} \quad t = 0, .., t_c, \tag{5}$$

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

$$E(\{R_t\}) = \sum_{t=0}^{t_c} \left( \frac{i_t - \sum_{s=f_0}^{f} i_{t-s} R_{t-s} \Phi_s}{p_{90}(i)} \right)^2 + \sum_{t=1}^{t_c} w_t (R_t - R_{t-1})^2 + \sum_{m=0}^{M} \beta_m (R_{t_m} - \bar{R}_{t_m})^2, \quad (6)$$

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 equation (5) 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, $\beta_m$ is a weight that determines the confidence we have in such initial estimate $\bar{R}_{t_m}$. The larger $\beta_m$, the greater this confidence. Typically, $M$ is equal to 0 or 1. In the case of $M = 0$, we use in the energy a prescribed value of $R_0$ and in the case of $M = 1$, we use the prescribed value of $R_0$ and a prescribed value of $R_{t_c}$ (the current time). Minimizing the energy $E$ leads to satisfy approximately the epidemiological model (5) with a reasonably smooth $R_t$ and prescribed initial value for $R_0$ and $R_{t_c}$. The parameters $w_t$ and $\beta_m$ determine the importance assigned to these constraints in the estimation.

Minimizing $E$ with respect to the sequence $\{R_t\}$ yields a linear system of equations that is easily solved, if it is complemented with 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 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 7 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

$$i(t) = \begin{cases} I(0)e^{at} - I(0)e^{a(t-1)} & if \quad t < 0; \\ \\ m_7 \cdot t + n_7 & if \quad t > t_c. \end{cases} \tag{7}$$

*Computation of the initial exponential growth a and* $\bar{R}_0$ We now naturally estimate $a$ by

$$a = median(\{log\left(\frac{I(t+1)}{I(t)}\right) \ : \ t = 0, .., 14\}). \tag{8}$$

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 (5) we obtain that

$$i_0 = I(0)(1 - e^{-a}) = I(0)R_0 \sum_{k=f_0}^{f} (e^{-ka} - e^{-(k+1)a})\Phi_k. \tag{9}$$

Hence, we can compute an approximation of $R_0$ as

$$\bar{R}_0 = \frac{1 - e^{-a}}{\sum_{k=f_0}^{f} (e^{-ka} - e^{-(k+1)a})\Phi_k}. \tag{10}$$

Note that this estimation strongly depends on the serial interval used. For instance for $a = 0.2$ we obtain that $\bar{R}_0 = 2.26$ for the Nishiura et al. serial interval, $\bar{R}_0 = 2.62$ for the Ma et al. serial interval and $\bar{R}_0 = 1.80$ 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* $w_t$

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

$$w_t = \hat{w}_0 \frac{G_{\sigma_w} * i(t)}{p_{90}(i)}, \tag{11}$$

where $G_{\sigma_w} * i(t)$ represents the convolution of $i_t$ with a Gaussian kernel of standard deviation $\sigma_w$ 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 use $\sigma_w = 3$, as standard deviation of the Gaussian kernel.

We noticed that in the case $\hat{w}_0$ is small, by minimizing the energy (6) one can obtain negative values $R_t$ for some $t$. In such cases, we increase iteratively the value of the regularization weights at such points and their neighbors and we compute again the minimum of (6). 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 (6). 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.

To remove the bulk of such impulses, we (optionally) filter the raw time series of infected with a sliding median filter with window size radius $w_R$ and a linear Gaussian filter with standard deviation $\sigma$. The application of the optional median filter is especially important in cases where singular data appear due to strong adjustments of the infected values from one day to the next. This is the case, for example, in France, the United Kingdom or Spain, but it is less necessary for larger countries such as the United States or India. A more local linear filter is also optionnally applied. It is less aggressive and can be used to smooth the data a bit. This linear filtering can be a Gaussian filtering or a moving average.

*Summary of the algorithm computing $R_t$.*

- Step 1: Pre-processing of the input data (detected infected): filtering of the data to reduce the administrative noise and to take into account that currently, some countries are not providing new data during the week-end.

- Step 2: Estimation of the time $t_0$ where the accumulated data, $I_t$, starts to grow in an exponential way. We use a very basic algorithm where we impose that $I_{t+1} > 1.1I_t$ in two consecutive days. More formally, we set

$$t_0 := \min_{t>2}\{t : I_t > 10 \text{ and } I_t > 1.1I_{t-1} \text{ and } I_{t-1} > 1.1I_{t-2}\}.$$

  Once $t_0$ is computed the previous data sequence is removed so that $t_0$ becomes 0.

- Step 3: Compute the initial exponential growth rate $a$ and an initial estimation of $R_0$ using (8) and (10).

- Step 4: Initial computation of $R_t$ by minimizing the energy (6). If any of the computed $R_t$ is non-positive we increase locally the regularization weight $w_t$ and we iterate the minimization of (6) as explained above.

### Short time forecasting 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 $R$ into the future, it is possible, using equation (5), to obtain a forecast of the number of infected $i_t$ in the near future. We considered two approaches to extrapolate $R_t$. In the first one we extrapolate $R_t$ beyond the current time $t_c$ assuming that the evolution of $R_t$ is going to be similar to that of the last days and we use a basic harmonic oscillator model to fit the evolution in the last days and to extrapolate $R_t$. In the second approach we allow the user to provide the expected constant new value of $R_t$ in the future and how many days are required to reach such new value.

*Extrapolation of $R_t$ using the harmonic oscillator model.*

We use the following damped harmonic oscillator :

$$R''(t) + cR'(t) + d(R(t) - \tilde{R}_1) = 0 \quad t \geq t_c, \tag{12}$$

the harmonic oscillator parameters $c$, $d$ and $\tilde{R}_1$ are computed by fitting $R(t)$ to the solution of the harmonic oscillator (see the Web appendix for technical details).

*User interactive extrapolation of $R_t$*

In this case the user provides the expected new value $R_1 = R(t_c + N_d)$, where $N_d$ is the number of days required to reach this new value. To obtain a smooth transition between the current value $R(t_c)$ and $R_1$ we use a basic Hermite interpolation polynomial (see the Web appendix for technical details).

## RESULTS

All of the experiments made here can be reproduced with the online demo available at `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 3 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 (6):
  - $\hat{w}_0$: regularization weight. The default value is $\hat{w}_0 = 10$.
  - $\beta_0$: weight for the initial estimation of $R_0$ computed using (10). 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 $R(t)$.
  - $\beta_1$: weight in the energy (6) for the initial estimation of $R(t_c)$ computed using the estimate in the last 3 days. The default value is $\beta_1 = 10^{0.2}$.
- Optional Data Pre-filtering:
  - $r_W$: 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.
  - $\sigma$: 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 $R_1$.

We shall pay particular attention to the comparison with EpiEstim, the method proposed by Cori et al. in [5] that we have explained briefly in the introduction. It is one of the most widely used methods to estimate $R(t)$. For the EpiEstim method we used, initially, the default parameters proposed in the Microsoft Excel spreadsheet implementation of the method

provided by the authors. In particular; they use a 7 day time interval size to estimate $R(t)$. 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 data of infected reported by the countries up to July 23, 2020 that we obtained from the European Centre for Disease Prevention and Control service. Since EpiEstim does not allow for negative values we replaced any negative value by zero in the data sequence. For several weeks France and Spain have not provided data during the weekends and have given instead on Mondays a cumulative count of three consecutive days. To avoid the artificial noise generated in the event that no median filter is applied to the sequence, the accumulated value of the 3 days divided by 3 was assigned to Saturday, Sunday and Monday.

We compared the results of our method and EpiEstim after applying (for both methods) a median filter and a Gaussian filter to the data, and we also compared them without applying this pre-filtering step. In Fig. 2 we show both data sequences for the four countries.

EpiEstim does not allow for a serial interval distribution with positive values at 0 or on negative days, so for comparison purposes we used the serial interval of Nishiura et al., as it is the one requiring a minimal truncation when removing the non-positive days from the serial interval distribution. In Fig. 3 the results obtained by both methods are presented when a median and Gaussian pre-filter were applied. In the case of the USA we divided by 10 the number of all infected $i_t$ before using EpiEstim because otherwise the method did not work (likely the values of $i_t$ for USA are too high for the particular implementation in the Excel spreadsheet). The estimate of $R(t)$ by the variational method is invariant under this kind of data transformation. Surprisingly, despite the fact that both methods are quite different, a very good fit of both estimates in the four countries was observed, after the EpiEstim estimate has been shifted 7 days.

It might be argued that if the data is pre-filtered, a smaller time interval could be used in EpiEstim. Fig. 4 shows the result obtained using EpiEstim with a time interval of size 1. It is observed, as expected, that the estimate of $R(t)$ is somewhat more irregular but due to the filtering, it is quite similar to those obtained for a time interval of size 7 and in this case the translation with respect to the results of the variational method is reduced to 4 days. In other

terms, as expected, the smaller the time interval, the smaller the backward shift of the estimate of $R_t$ with respect to the current date. In summary, two factors intervene in the shift between the estimation of $R_t$ by the variational model and by EpiEstim: a first shift, as explained in the introduction, occurs by assuming $R_t$ locally constant in the model (2) and a second shift occurs when estimating $R_t$ in a time interval.

In Fig. 5 we present the $R(t)$ estimates obtained by both methods when no pre-filtering was applied to the data. In that situation, the weight of the regularization was increased in our method to compensate for the lack of regularity of the data. The fit of both estimates is relatively good, but less exact than after applying our pre-filtering. This is a reasonable outcome. Indeed, the more irregular the curve $R(t)$, the less accurate the assumption that $R(t)$ is locally constant. Hence the model (1) used in our method and its simplified version (2) used in EpiEstim are less coherent with each other. Comparing Fig. 2 and 5, we notice that for the USA the original data $i(t)$ is more regular than in the other countries (likely because the numbers in USA are much larger than in the other countries; hence the administrative noise is smoothed out). So we find a better agreement between the estimate of $R_t$ by both methods.

In Fig. 6 we show the results of two different forecasting strategies using an extrapolation of $R(t)$. In the Web appendix we present additional experiments to illustrate the influence of the different parameters of the variational model.

# DISCUSSION

In this paper we proposed a variational model given by the expression (6) 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 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, we avoid 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 up to 7 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 epidemiological model (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 a median filter and a Gaussian filter. This pre-processing step removes most of the administrative noise of the data and eases the $R_t$ estimation. In any case, this pre-processing is optional and its application depends on the level of noise of the country data.

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 and the serial interval, a good agreement can been obtained between the estimate of $R_t$ provided by the variational model and a 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 forecast the value of $R_t$ in the short term. Currently, in countries that are relaxing 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 forecasting techniques proposed in this paper are just an example that shows how an $R_t$ estimation can be used to forecast the number of infected. We proposed two techniques to extrapolate the value of $R_t$. The first one is automatic and assumes that the behavior of $R_t$ in the next future will be similar to that of the recent past that is modeled using a harmonic oscillatory model. The second technique is interactive and allows the user to supply 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, but it is reasonable to assume that any country exerts itself to reach an ERN lower than 1.

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

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] T. Boulmezaoud, *Un modèle de prédiction de l'épidémie covid-19 et une stratégie zig-zag pour la contrôler*, (2020).

[3] 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).

[4] 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.

[5] 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.

[6] 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).

[7] 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/`.

[8] 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.

[9] 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).

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

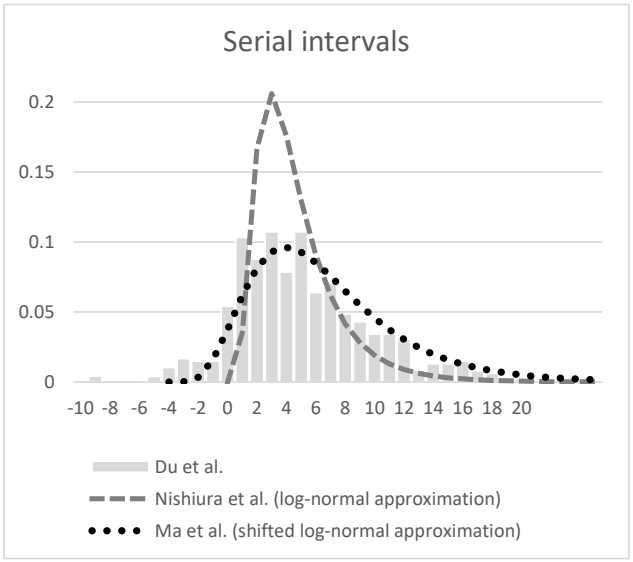


Figure 1: Serial intervals used in our experiments: the discrete one proposed by Du et al. in [6] (solid bars), a log-normal approximation of the serial interval proposed by Nishiura et al. in [11] (dotted line) and a shifted log-normal approximation of the serial interval proposed by Ma et al. in [9] (dashed line).

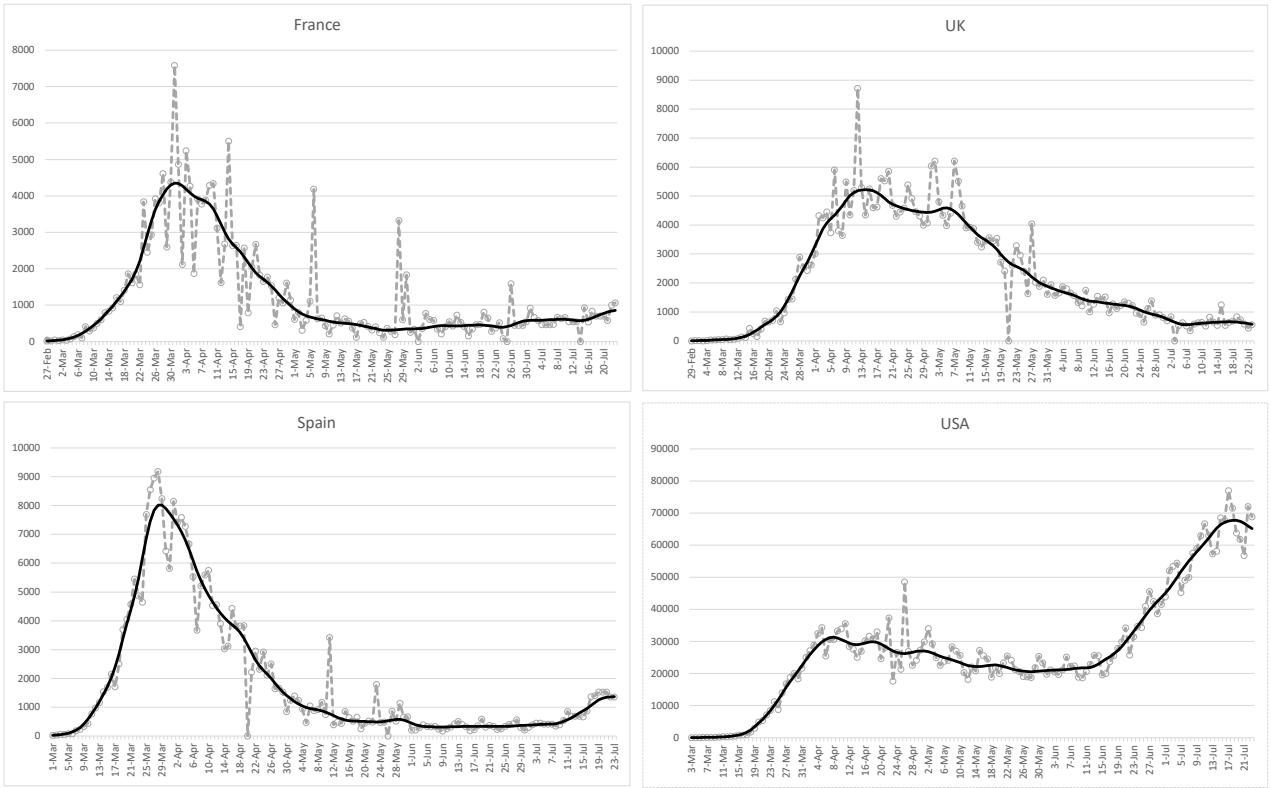


Figure 2: Data of daily infected patients used for comparison with EpiEstim: we show, for the four countries, the data we used when no filtering was applied (solid line) and the data after applying a median filter (dashed line) with $r_W = 3$ and a Gaussian filter with $\sigma = 2$.

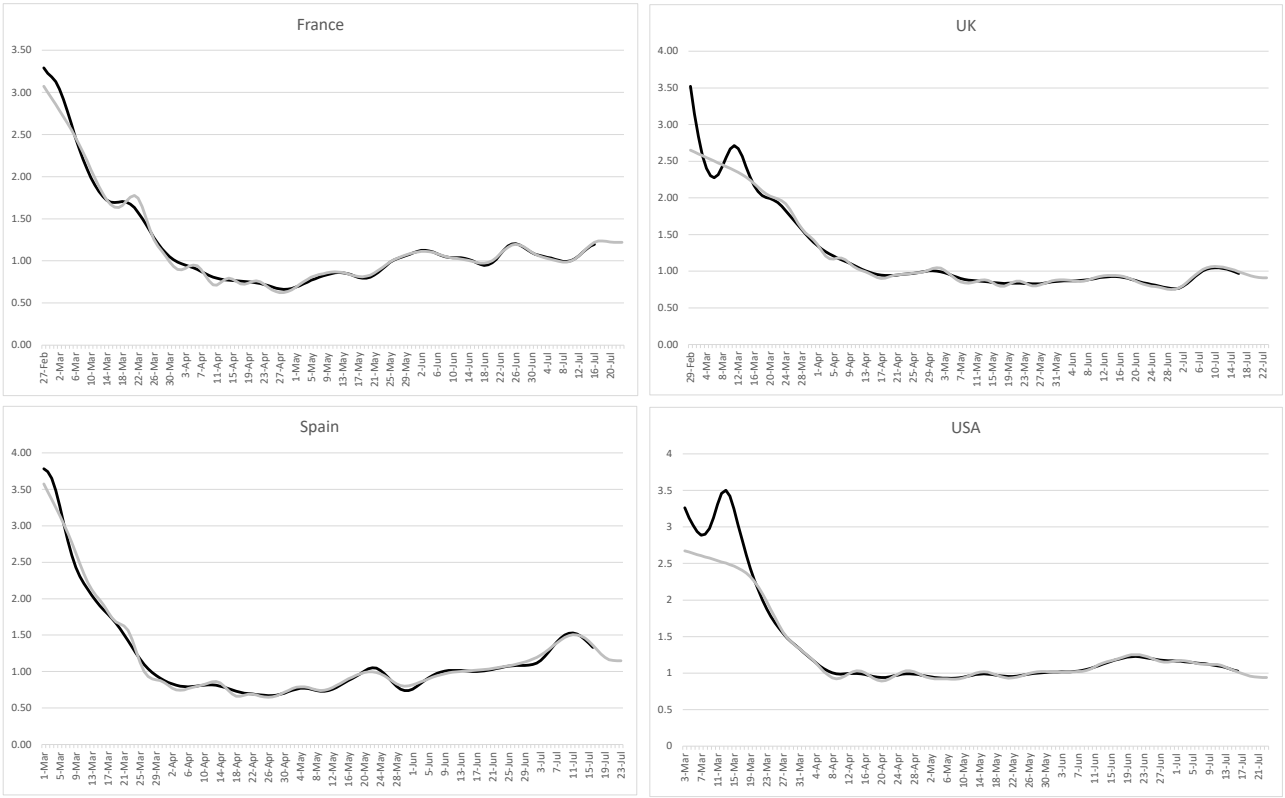


Figure 3: Comparison with EpiEstim including data filtering. We show $R(t)$ obtained using the proposed variational model (in grey) and EpiEstim (in black) using the data filtered with $r_W = 3$ and $\sigma = 2$. We used the Nishiura et al. serial interval. For the variational method we used $\hat{w}_0 = 10^{-1.2}$, $\beta_0 = 10^5$ and $\beta_1 = 0$. The results of EpiEstim are shifted 7 days to fit the ones obtained by the variational technique.

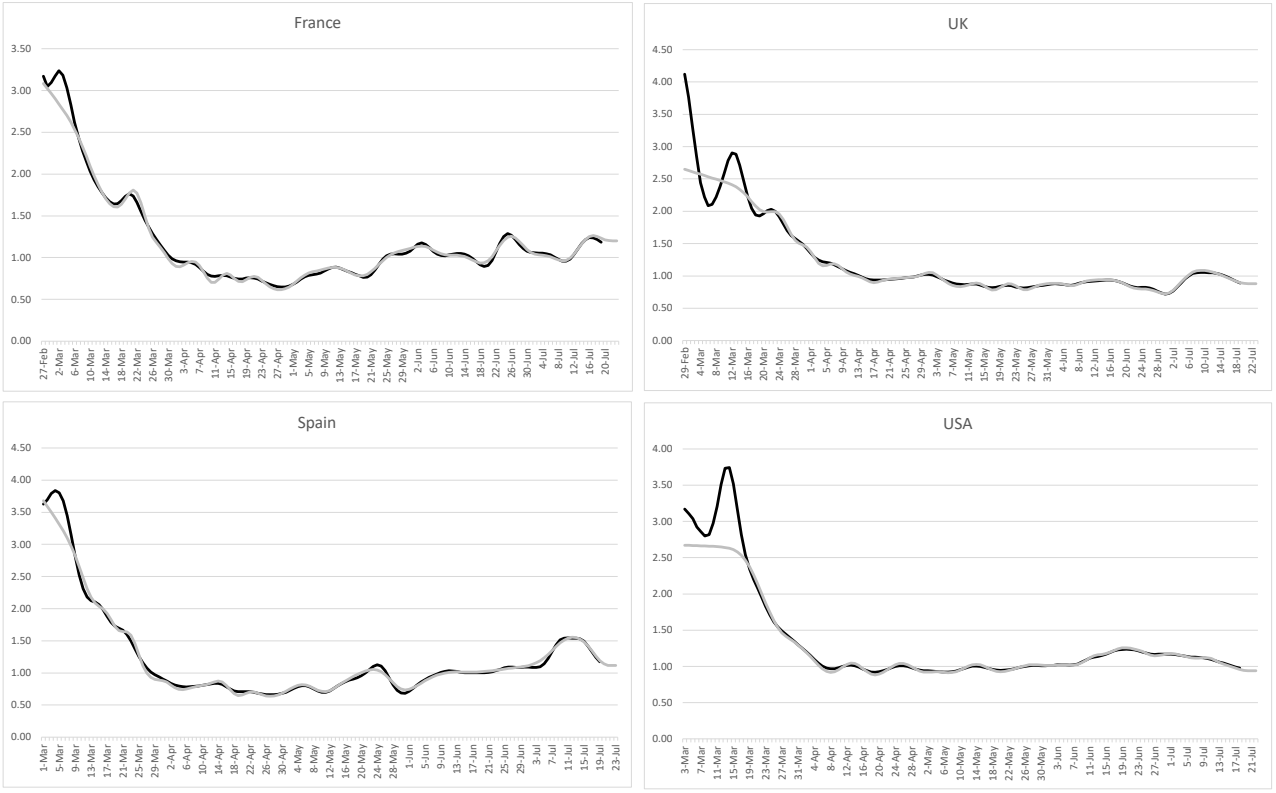


Figure 4: Comparison with EpiEstim applied with a smaller time interval. We show the same results as in Fig. 3 with the filtered data but we use a time interval of size 1 to apply EpiEstim. For the variational model we use $\hat{w}_0 = 10^{-1.6}$ (the estimation with EpiEstim is shifted 4 days).

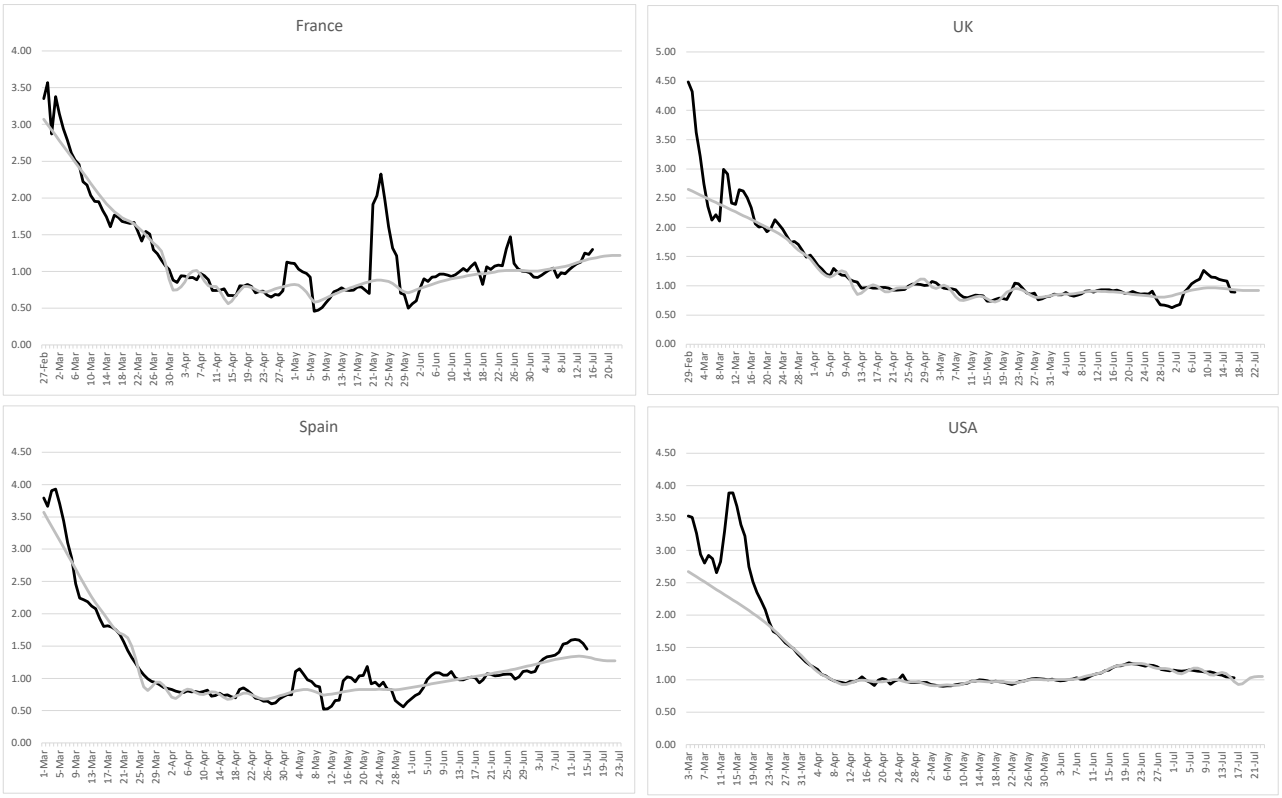


Figure 5: Comparison with EpiEstim without data filtering. We show the same results as in Fig. 3 but without data filtering and using $\hat{w}_0 = 1$ (the estimation of EpiEstim is shifted 7 days).

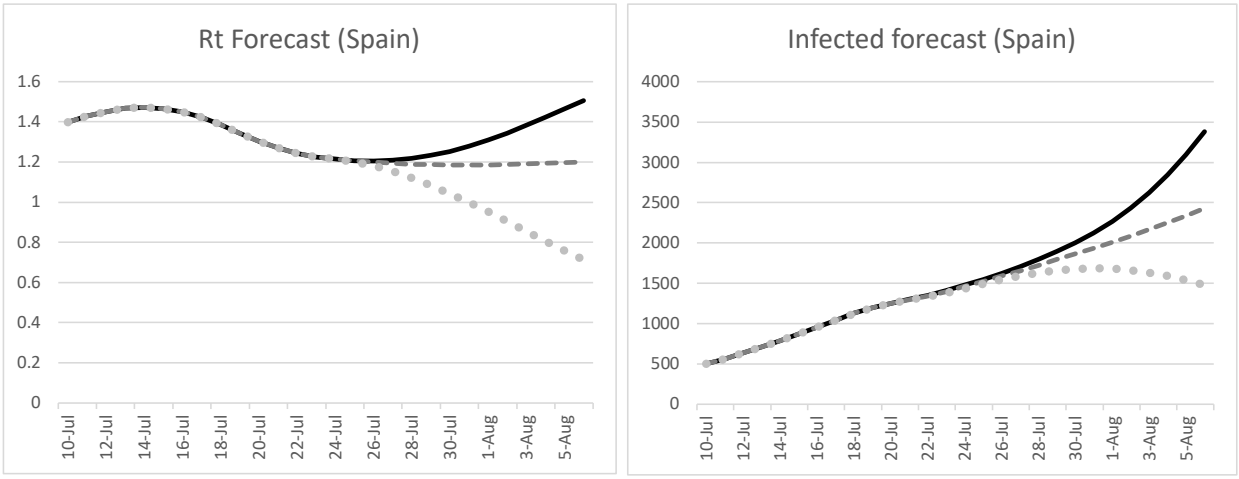


Figure 6: Forecasting. We show a 14 day forecast for $R(t)$ and $i(t)$ (we used the data up to July 23 and we forecast from July 24 to August 6). We show a forecast using the harmonic oscillatory model (solid line), a forecast with an objective $R(t)$ value given by $R_1 = R(t_c + 21) = 1.15$ (dashed line) and a forecast with $R_1 = R(t_c + 21) = 0.5$ (dotted line). In the last both cases the number of days to reach the objective value is 21. For the variational model, we used as parameters $r_W = 3$, $\sigma = 2$, $\beta_0 = 10^5$, $\beta_1 = 0$ and the Du et al. serial interval.

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

## Pre-processing of the input data

We check for each country if, 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 we look, in the past for a time $t = M$ (the Monday) where $i_{M-3} > 0$, $i_{M-2} = i_{M-1} = 0$ and $i_M > 1.5 \cdot i_{M-3}$. If we found that pattern we mark 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-end we update the value $i_{M-7 \cdot k-2} = i_{M-7 \cdot k-1} = i_{M-7 \cdot k} = i_{M-7 \cdot k}/3$. In short, we assign to Saturday, Sunday and Monday a third of the cumulative incident cases over the three days.

The next step in the data pre-processing is an optional median filter to remove administrative noise. 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 optionally perform a linear Gaussian convolution to the data. The general expression to compute the convolution with a symmetric kernel such as the Gaussian function is:

$$G * i(t) = g_0 \cdot i(t) + \sum_{s=1}^{N_G} g_s \cdot (i(t+s) + i(t-s)), \tag{13}$$

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

$$g_0 + 2 \sum_{s=1}^{N_G} g_s = 1. \tag{14}$$

In the expression (13), when $t + s > t_c$ we use a linear interpolation 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:

$$i'(t_c) = \frac{g_0 \cdot (i(t_c) - i(t_c - 1)) + \sum_{s=1}^{N_G} g_s \cdot (i(t_c - s) - i(t_c - s - 1))}{g_0 + \sum_{s=1}^{N_G} g_s}, \tag{15}$$

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

$$i(t + s) = i(t_c) + i'(t_c)(t + s - t_c) \tag{16}$$

Finally we also include an optional moving average filter as a data preprocessing step.

## Short time forecasting of $R_t$ and $i_t$

*Extrapolation of $R_t$ using the harmonic oscillator model.*

We fix the parameters $c$, $d$ and $\tilde{R}_1$ of the harmonic oscillator (12) by fitting $R(t)$ to the solution of the harmonic oscillator for $t \in [t_c - N_1, t_c]$. $\tilde{R}_1$ is the reference value in the last days given by

$$\tilde{R}_1 = median(\{i_t \ : \ t \in [t_c - N_1, t_c]\}).$$

In the experiments we fixed the value of $N_1$ to 14. We point out that the exponential decay of the amplitude of the oscillations is given by $e^{-ct}$. If $(4d - c^2 > 0)$, then the period $T$, of the oscillations is given by

$$T = \frac{4\pi}{\sqrt{4d - c^2}}$$

To avoid spurious oscillations and a strong increase of their amplitude we impose (when computing $c$ and $d$) that

$$c \geq -0.025$$

$$T \geq 7 \quad when \quad 4d - c^2 > 0,$$

once the parameters of the harmonic oscillatory model are fixed we forecast the evolution of $R(t)$ for $t > t_c$ using the solution of the harmonic oscillatory model with the initial conditions

$R(t_c)$ and $R'(t_c)$.

*User interactive extrapolation of $R_t$*

The Hermite interpolation polynomial we use to extrapolate $R_t$ is given by

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

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 $R_1$. The values of $R_1$ and $N_d$ can be fixed manually accordingly with the current and/or expected social distancing measures taking by the states. For instance, in the European countries, currently, we observe that the value of $R_t$ oscillates around 1. In the experiments, we fixed the default values of this parameters to $R_1 = 1$ and $N_d = 21$ days. By default, we initialize $\gamma = 1$ and then reduce the value of $\gamma$ 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 $\tilde{R}(t) > 0$ $\forall t \in [t_c, \infty)$.

*Forecasting of the daily infected $i_t$.*

Once we have forecast $R_t$ for $t = t_c + 1, t_c + 2, .., t_c + dT$ (where $dT$ is the number of new days to forecast), we can compute by iteration $i_t$ from $t_c + 1$ to $t_c + dT$ using equation (5) and the extrapolation of $i_t$ given by (7). Then to reduce the effect of the extrapolation of $i_t$ when $t > t_c$, we observe that, in fact, equation (5) can be interpreted as a fixed point equation, so we iterate equation (5) until convergence to update $i_t$ for $t = t_c - 9, ..., t_c + dT$. We start at

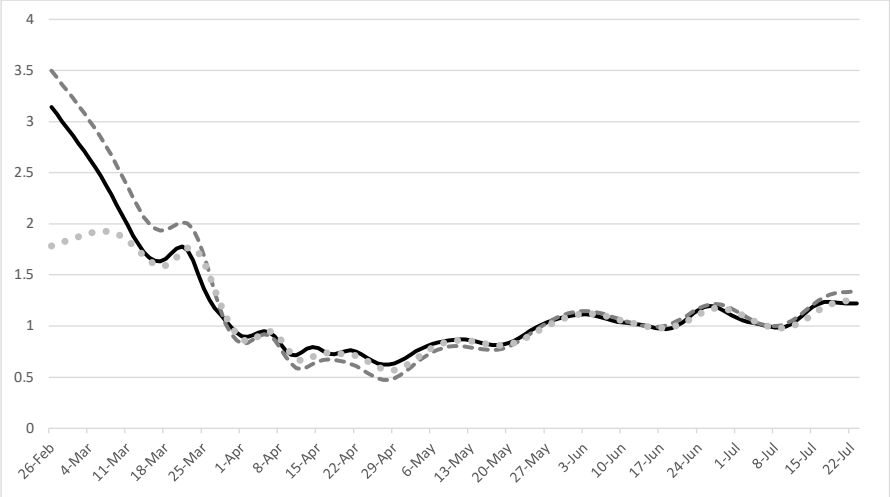


Figure 7: Influence of the serial intervals. We show, in the the case of France, the results of the $R(t)$ computation obtained by the proposed method using the three serial intervals Nishiura et al. (solid line), Ma et al. (dashed line) and Du et al. (dotted line). In all cases we used the filtered data with $r_W = 3$, $\sigma = 2$, $\beta_0 = 10^5$, $\beta_1 = 0$ and $\hat{w}_0 = 10^{-1.2}$.

$t_c - 9$ to smooth a bit the forecast and to reduce potential discontinuities of the forecast $i_t$ at the time $t_c$.

## Influence of the algorithm parameters

This section shows some experiments illustrating the influence of the different parameters of the variational model. In Fig. 7 we illustrate the influence of the serial interval. We see a big difference at first because the value of $R_0$ is highly dependent on the serial interval. At the end it is observed that the estimate with the Ma et al. serial interval is significantly higher than the other two and that the estimate with the Nishiura et al. serial interval tends to come to the end flatter than the other two. In Fig. 8 we illustrate the influence of the regularization parameter. As expected, we note that the higher the value of this parameter, the smoother the estimate of $R(t)$.

In Fig. 9 we illustrate the influence of the weight $\beta_0$ in the energy (6). The higher the value of this parameter, the more similar the value of $R(0)$ is to the initial estimate obtained from the initial exponential growth rate of the epidemic.

In Fig. 10 we show the influence of the data filtering. We notice that the median filter gets

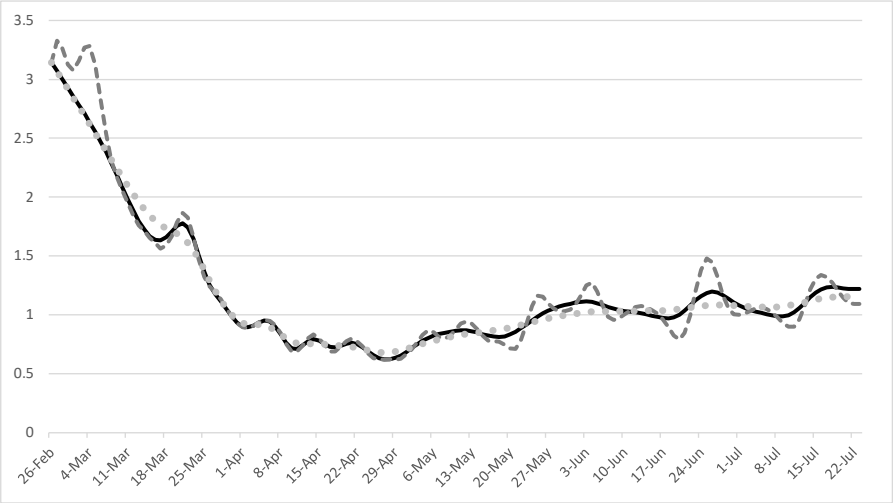


Figure 8: Influence of the regularization parameter. We show, in the the case of France, the results of the $R(t)$ computation obtained by the proposed method using the filtered data with the Nishiura et al. serial interval, $r_W = 3$, $\sigma = 2$, $\beta_0 = 10^5$, $\beta_1 = 0$ and $\hat{w}_0 = 10^{-1.2}$ (solid line), $\hat{w}_0 = 10^{-5}$ (dashed line), and $\hat{w}_0 = 1$ (dotted line).

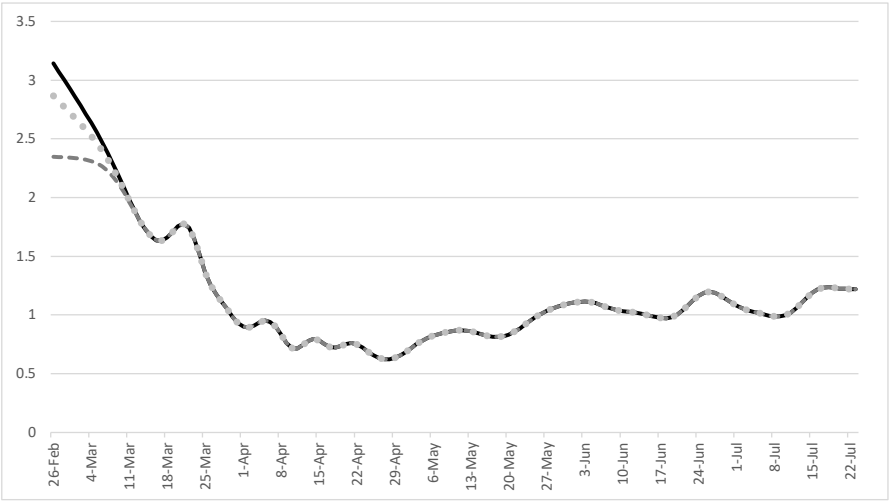


Figure 9: Influence of the weight $\beta_0$ in the energy (6). We show, in the the case of France, the results of the $R(t)$ computation obtained by the proposed method using the filtered data with the Nishiura et al. serial interval, $r_W = 3$, $\sigma = 2$, $\hat{w}_0 \equiv 10^{-1.2}$ and $\beta = 10^5$ (solid line), $\beta = 10^{-5}$ (dashed line), $\beta = 10^{-2}$ $\beta_1 = 0$ (dotted line).

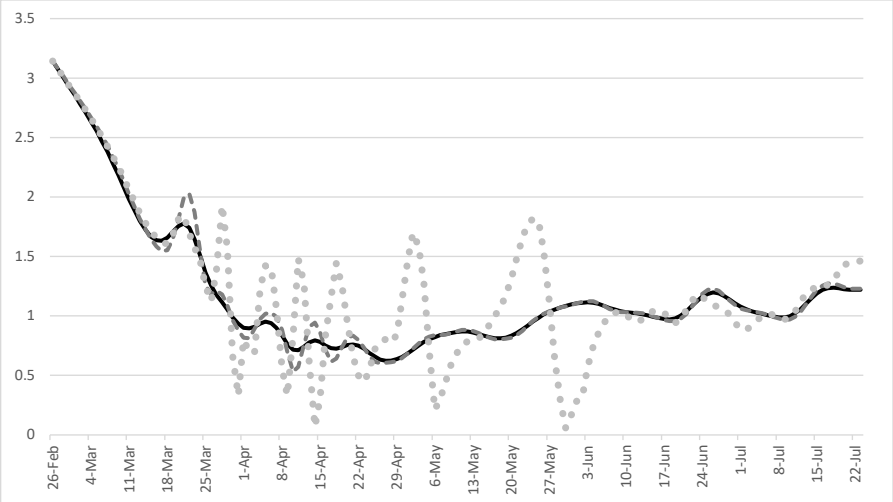


Figure 10: Influence of the data filtering. We show, in the the case of France, the results of the $R(t)$ computation obtained by the proposed method using the Nishiura et al. serial interval, $\hat{w}_0 = 10^{-1.2}$, $\beta_0 = 10^5$, $\beta_1 = 0$ and $r_W = 3$, $\sigma = 2$ (solid line), $r_W = 3$, $\sigma = 0$ (dashed line), and $r_W = 0$, $\sigma = 0$ (dotted line).

rid of much of the "administrative" noise, so the estimate of $R(t)$ is much smoother when the median filter is applied.

## Managing the variability of the $R_t$ estimate.

We emphasize that we do not assume any statistical model on the distribution of the $R_t$ values, therefore, we cannot provide confidence intervals for this estimate using such statistical model. 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. We compute $R^0(t)$ by minimizing (6) for $t \in \{t_0, .., t_c\}$ with $M = 0$ (that is we do not add any restriction on the value of $R(t_c)$).

2. We compute $R^{-1}(t)$ by minimizing (6) for $t \in \{t_0, .., t_c - 1\}$ with $M = 0$, that is we remove the last value of the data sequence.

3. We compute $R^{-2}(t)$ by minimizing (6) for $t \in \{t_0, .., t_c - 2\}$ with $M = 0$, that is we remove the last two values of the data sequence.

4. We fix $M = 1$ in (6), $t_1 = t_c$ 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. We compute $R(t)$ by minimizing (6) for $t \in \{t_0, .., t_c\}$ with $M = 1$ using $\bar{R}_1$ as initial estimate of $R(t_c)$.

This procedure stabilize the estimate of $R(t_c)$ with respect to the estimation in the last 3 days. In addition, it allows us to calculate a measure of the variation of $R(t)$ estimate in the last 3 days using the expression

$$sigma(t) = \sqrt{\frac{(R(t) - R^0(t))^2 + (R(t) - R^{-1}(t))^2 + (R(t) - R^{-2}(t))^2}{3}} \tag{17}$$

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 $[R_t - 2 \cdot sigma(t), R_t + 2 \cdot sigma(t)]$.

## Representation of the filtered data in the online software

In the online software we 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 of no pre-filtering is applied we plot the result of the application of the formula (5) to the data sequence $i_t$ with the estimated $R_t$. That is, we plot:

$$\tilde{i}_t = \sum_{s=f_0}^{f} i_{t-s} R_{t-s} \Phi_s \qquad \text{for } \; t = 0, .., t_c, \tag{18}$$

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