Rt2: computing and visualising COVID-19 epidemics temporal reproduction number

Bastien Reyné, Gonché Danesh, Samuel Alizon, Mircea T. Sofonea

Abstract Analysing the spread of COVID-19 epidemics in a timely manner is essential for public health authorities. However, raw numbers may be misleading because of spatial and temporal variations. We introduce Rt2, an R-program with a shiny interface, which uses incidence data, i.e. number of new cases per day, to compute variations in the temporal reproduction number ($\mathcal{R}_t$), which corresponds to the average number of secondary infections caused by an infected person. This number is computed with the **R0** package, which better captures past variations, and the **EpiEstim** package, which provides a more accurate estimate of current values. $\mathcal{R}_t$ can be computed in different countries using either the daily number of new cases or of deaths. For France, these numbers can also be computed at the regional and departmental level using also daily numbers of hospital and ICU admissions. Finally, in addition to $\mathcal{R}_t$, we represent the incidence using a one-week sliding window to buffer daily variations. Overall, Rt2 provides an accurate and timely overview of the state and speed of spread of COVID-19 epidemics at different scales, using different metrics.

Context

Monitoring the state and speed of spread of COVID-19 epidemics at the national and regional levels is crucial to implement non-pharmaceutical interventions (Flaxman et al., 2020). Every day, public health agencies communicate key figures to monitor the epidemic, especially incidences, which correspond to the number of new cases detected. These incidences are typically related to four variables, which are PCR-based detection, deaths, hospitalisations, and ICU admissions. The statistical analysis of time variations in these time series can inform us about epidemiological dynamics.

The purpose of the Rt2 application is to visualize the trends present in COVID-19 data through the temporal reproduction number, noted $\mathcal{R}(t)$ (Wallinga and Lipsitch, 2007). This is done at different levels (country and, for France, region, and department) using different types of data (incidence in number of cases detected, in number of deaths and, for France, in number of hospitalisations, and ICU admissions).

Reproduction numbers

Overview

One of the key parameters in an epidemic is the reproduction number, $\mathcal{R}$, which reads as the average number of people infected by a contagious person during his or her infection. At the beginning of an epidemic, when the whole population is susceptible (*i.e.* not immune) and no interventions are implemented, this number has a particular value noted $\mathcal{R}_0$ and called the basic reproduction number. During the course of the epidemic, when the proportion of immunized people becomes large enough to slow the transmission of the virus (by an effect comparable to a dilution of the individuals still susceptible), this number is called the effective, or temporal reproduction number, and denoted $\mathcal{R}(t)$.

Intuitively, if $\mathcal{R}(t) > 1$, then one person is infecting more than one person on average and the epidemic is growing, as shown in textbooks such as the one by Anderson and May (1991). As the COVID-19 epidemic spreads, $\mathcal{R}(t)$ decreases as an increasing proportion of the population becomes immune. When the group immunity threshold is exceeded, $\mathcal{R}(t)$ falls below the threshold of 1, an epidemic peak is reached, and the epidemic declines. Public health control measures can also decrease $\mathcal{R}(t)$. Therefore, the epidemic peak can be reached before the threshold for herd immunity is.

At a given point in time t, knowing the value of $\mathcal{R}(t)$ is therefore essential to determine the status of the epidemic.

Formal definitions

In practice, as discussed by Wallinga and Lipsitch (2007), we need two pieces of information to calculate the reproduction number:

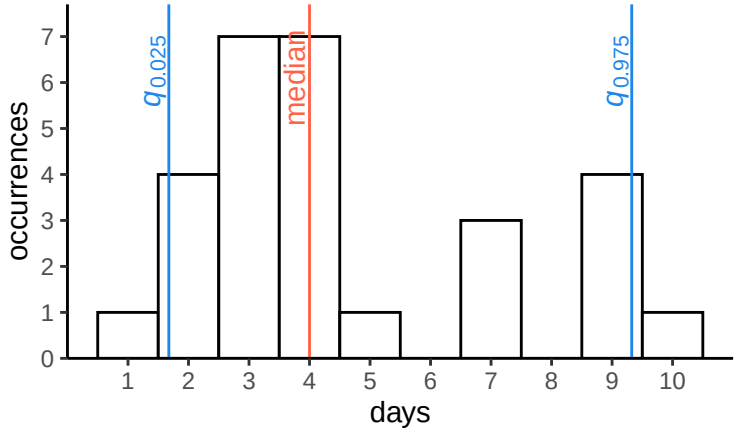

Figure 1: Serial interval for COVID-19 epidemics computed by Nishiura et al. (2020). The blue lines show the 95% confidence interval and the red line shows the median.

1. the growth rate of the epidemic at a given time, $r(t)$, which can be easily calculated from incidence time series (daily number of new cases detected, new hospitalizations, new deaths...),
2. the time during which a person is contagious, D.

A well-known and intuitive relationship between all these terms is as follows:

$$\mathcal{R}(t) = D\, r(t) + 1 \tag{1}$$

However, this apparent simplicity masks significant problems.

First, even if we assume a Markovian setting, *i.e.* in the absence of memory discussed for COVID-19 epidemics by Sofonea et al. (2020), this equation remains an approximation: the $\mathcal{R}(t)$ from equation 2 is underestimated by a quantity $\mathcal{R}_0\ p\,(t)$, where $p\,(t)$ refers to the proportion of cases at time t, a quantity that is not always negligible, especially near the epidemic peak.

Second, the duration of contagiousness D is difficult to capture. It is tempting to use a fixed duration but this implicitly assumes that individuals become contagious immediately after infection and lose contagiousness at a constant rate (Markovian regime); two conditions that are rarely satisfied. On the one hand, there is often a latency period during which the individual is not infectious (in the case of COVID-19, this latency is less than the incubation period, hence the difficulty of measurement, as shown by He et al. (2020)). This imposes a positive corrective term to the estimated growth rate. On the other hand, the loss of contagiousness seems to show a non-Markovian pattern (He et al., 2020).

To visualize the neither exponential nor uniform aspect of the distribution of contagiousness over time, we show below the COVID-19 serial interval compiled from 28 infector-infectee pairs by Nishiura et al. (2020) in Figure 1. Clearly, the median (in red) and even 95% confidence interval (between the blue bounds) are both not very representative of the totality of the values.

To circumvent the contagiousness duration problem and its modelling in a classical SIR model, while relying on the empirical distribution, an elegant solution is provided by the Euler-Lotka equation. As detailed by Wallinga and Lipsitch (2007), provided that the epidemic growth is exponential, we can write

$$\mathcal{R}(t) = \left(\int_{a=0}^{\infty} e^{-r\,a}\, g(a)\, \mathrm{d}a \right)^{-1}, \tag{2}$$

where a is the 'age' of an infection, $g(a)$ is the distribution of contagiousness durations, and r is the growth rate, assumed to be constant over the time interval. In practice, $g(a)$ is approximated by the serial interval, noted here as $w(a)$, described above and which corresponds to the time between the onset of symptoms in an infection pair.

Outside the exponential regime and, in particular, in the vicinity of an epidemic peak, the continuous formulation in equation 2 does not apply. However, we can use a general formula introduced by Wallinga and Lipsitch (2007) that directly involves incidence data ($(y_1, \ldots, y_n)$, where y_k represents the number of new cases detected on day k):

$$\mathcal{R}\,(t) = \frac{y_t}{\sum\limits_{a \geq 0} y_{t-a}\, g(a)}. \tag{3}$$

The latter expression also has the advantage to work in discrete time, which is more appropriate for incidence data.

Implementation

Building on existing R-packages, we developed a shiny interface to show the temporal reproduction number and the incidence for a variety of data at different scales. We briefly present the data and the packages used, as well as the shiny implementation itself.

Data processing

Data is collected from two repositories: that from Max Roser and Hasell (2020), which itself mainly uses the data from the European Centre for Disease Prevention and Control (ECDC), and, for France, the national public health agency Santé Publique France. Both sources provide a daily-updated CSV file.

Since the packages used to compute an estimation of the temporal reproduction number require a two-columns data frame as an input (date and incidence data), raw data were pre-processed using **dplyr** and **tidyr** packages. More specifically, we created a function returning the desired data frame given the type of data and the location wanted.

It also might worth to note the incidence data were smoothed out using a 7-days rolling average, in order to compensate the incidence reporting delays happening on week-ends. We used the `rollmean` function from the **zoo** package.

R0

The **R0** R-package developed by Obadia et al. (2012) implements maximum likelihood estimation methods for the basic reproduction number $\mathcal{R}_0$ using the approach from White and Pagano (2008).

It can also compute temporal reproduction numbers $\mathcal{R}(t)$ using a procedure introduced by Wallinga and Teunis (2004), which works in the following way. We assume an epidemic curve is a time series of incidences $(y_1, \ldots, y_n)$, where y_k represents the number of new cases detected on day k. Note that in the original model, detection corresponds to the onset of symptoms.

This time series is the result of a set of transmission events between all cases detected since the beginning of the epidemic. If the proportion of undetected cases remains constant over the time window studied, the estimate is not biased; otherwise, pre-processing of the data is necessary, as in White et al. (2009). The structure of this set of events is seen as a tree (more precisely a connex acyclic oriented graph) whose probability of underlying data (*i.e.* the likelihood) is decomposed into infector/infected pairs (transmissions are assumed to be independent).

If $p_{i,j}$ denotes the probability that an individual I_d detected on day t_d is the infector of an individual I_i detected on the day t_i, then, by definition of the serial interval the distribution of which is characterized thereafter $(w_k)_k \geq 0$, we have

$$p_{i,j} \propto w_{t_i - t_j} \tag{4}$$

Therefore, the relative likelihood $\ell_{i,j}$ of the pair (i, j) to be a transmission pair is given by the expression

$$\ell_{i,j} = \frac{p_{i,j}}{\sum_{k \neq i} p_{i,k}} \tag{5}$$

The individual reproduction number of case j (denoted R_j) is therefore the sum of all cases i she/he may have infected, weighted by the relative likelihood of each pair:

$$R_j = \sum_i \ell_{ij} \tag{6}$$

The temporal reproduction number is obtained by averaging the individual reproduction number over all cases detected on the same day j, where $t_j = t$:

$$\mathcal{R}(t) = \frac{1}{y_{t_j}} \sum_{j : t_j = t} R_j \tag{7}$$

Note that **R0** allows the user to enter raw values for the serial interval (which we implement in one of the options of Rt2).

EpiEstim

The **EpiEstim** software package, developed by Cori et al. (2013) and updated by Thompson et al. (2019), is motivated by the fact that, in situations where the studied epidemic is still ongoing, the total number of infections caused by the last detected cases is not yet known. This issue is particularly acute when it comes to evaluating the effectiveness of control measures rapidly (i.e. within few days).

In the approach used by Wallinga and Teunis (2004) and Obadia et al. (2012) (in the **R0** package), the reproduction number of cases (or of cohorts), is retrospective: its calculation is based on the number of secondary cases actually caused by a cohort of detected infectors from the date on which they were detected.

Conversely, the method used by Cori et al. (2013) and Thompson et al. (2019) in the **EpiEstim** package computes the instantaneous reproduction number in a prospective manner: its calculation is based on the potential number of secondary infections that a cohort of cases could have caused if the conditions of transmissibility had remained the same as at the time of their detection.

Formally **EpiEstim** maximizes the likelihood of incidence data (viewed as a Poisson count) observed over a time window in which the reproduction number is assumed to be constant. By noting y_k and y_k^+ respectively the numbers of new local and total (i.e. including imported) cases, and w_s the probability corresponding to a serial interval s, the temporal (instantaneous) reproduction number for the interval $[t; t-\tau]$ satisfies

$$\mathcal{R}_\tau(t) = \underset{R}{\operatorname{argmax}} \left\{ \prod_{k=t-\tau}^{t} y_k!^{-1} e^{-R\sum_{s=1}^{k} w_s y_{k-s}^+} \left(R \sum_{s=1}^{k} w_s y_{k-s}^+ \right)^{y_k} \right\} \tag{8}$$

Quoting Cori et al. (2013), this distinction is equivalent to that between the (retrospective) life expectancy of a cohort of individuals, calculated once they have all died, and the (prospective) life expectancy of the same cohort, estimated under the assumption that mortality will remain the same as that known at birth. Note that the additional calculations required by the **EpiEstim** software package to make inferences tend to increase the computation time.

Interface

We used the **shiny** package in order to present the results through an interactive web app allowing users to choose key parameters. Unfortunately, the **EpiEstim** package proved to be too slow for such an interactive setting. To overcome this issue, we first generate a CSV file that contains every single parameter combination and can be read by the app. The CSV file is routinely updated on a day-to-day basis.

Visualisation

The visualisation was performed using the **dygraphs** library, integrated to the shiny app through the **htmltools** package. The **dygraphs** package allows for interactions, mainly zooming-in on the time axis (x-axis) and displaying the value of any mouse-selected day.

Figure 2 shows a typical screenshot of the visual interface.

Limitations and interpretations

There are a series of limitations inherent to the estimation of the temporal reproduction number ($\mathcal{R}t$), which we attempted to address in the development of Rt2 to limit misinterpretation risks.

Time delays

While the reproduction number quantifies a potential for the spread of the epidemic at a given date, its estimate is based on data that reflect the state of the epidemic several days earlier. Indeed, from the moment a person is infected, it takes a few days before the virus can be detected, and then usually another week before a potential hospitalization, with a possible admission in an intensive care unit

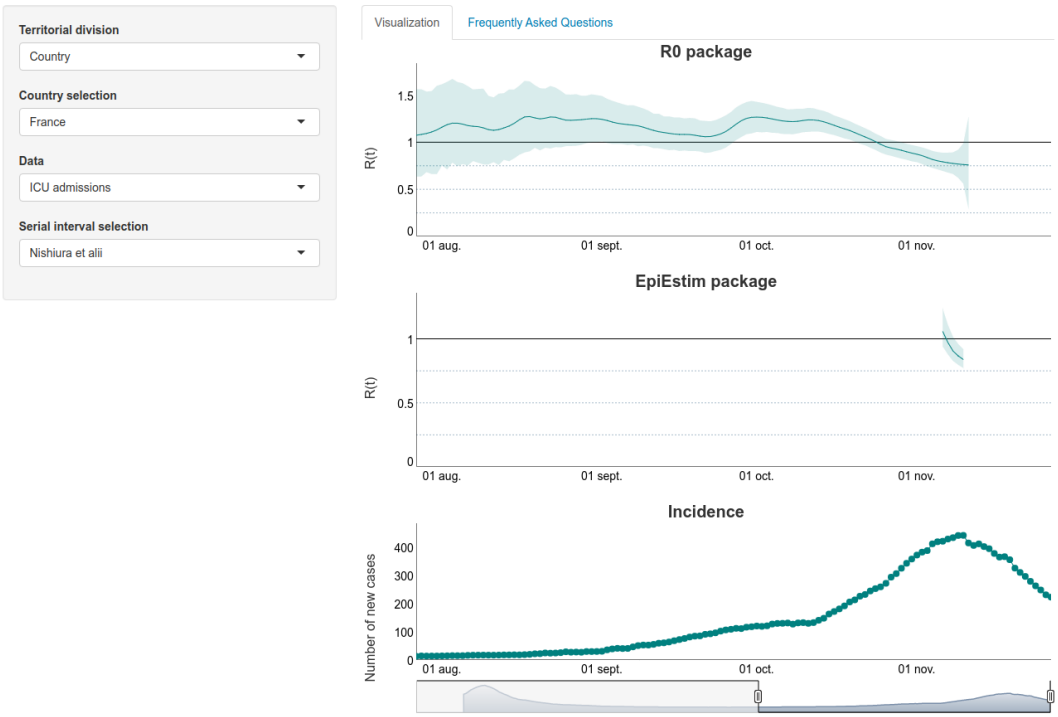

Figure 2: Screenshot of the Rt2 for ICU admissions incidence data from France.

(ICU) after another delay. Death can occur several days or weeks after ICU admission. In addition, time must be allowed for the case/admission/death to be identified and reported.

To account for this, and based on estimations of the COVID-19 epidemics in France (Sofonea et al., 2020), we impose a shift of 10, 12, 14, and 28 days between $\mathcal{R}(t)$ and the case, hospitalisation, ICU admission, and death daily incidence data respectively.

Sampling effort variations

While it is not necessary for all cases to be recorded for the methods in **R0** and **EpiEstim** to work properly, it is necessary for the sampling rate to be as constant as possible over time. Indeed, an intensification of screening effort mechanically amplifies the number of cases detected, therefore mimicking a growing epidemic. In France, case incidence in can be particularly misleading because screening intensity was initially very low (limited to severe cases), but increased a lot during the epidemic. Furthermore, there are known weekly variations with lower incidences during week-ends for instance.

To limit variations in sampling efforts, we use a 7-days moving average in our calculations and when plotting the incidence data.

Serial interval

As explained above, a key input for estimating the reproduction number is the number of days during which a person is contagious. In practice, this 'generation time' is estimated by tracking contacts (*i.e.* infector-infectee pairs) and counting the number of days between the dates of onset of symptoms in the infecting and infected individuals respectively. This approximation via the symptom onset dates corresponds to the the serial interval.

Data for this serial interval in France and in Europe was initially (and still largely is) very limited. To account for this, our application allows the user to choose between several data sets for the serial interval in order to better understand the effect of this parameter on the results shown in Figure 3. Note that one of the distributions available corresponds to the raw data from Nishiura et al. (2020).

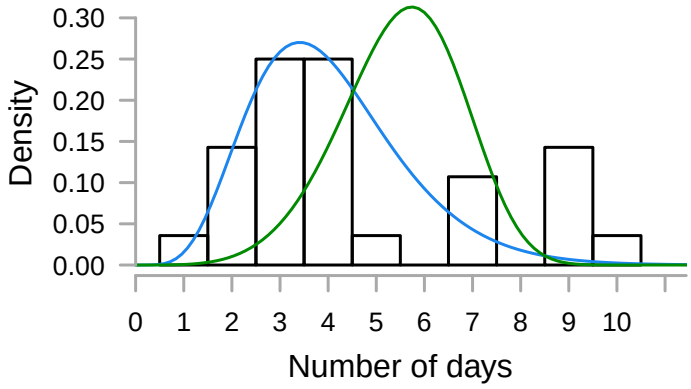

Figure 3: Serial intervals available in the Rt2 package. The blue curve is a Gamma distribution, $\text{Gamma}(6.5, 0.62)$, the green curve is a Weibull distribution, $\text{Weibull}(5, 6)$, and the histogram shows the raw data from Nishiura et al. (2020).

Summary

We developed an R-program with a shiny interface, which shows variation in incidence data and in the temporal reproduction numbers ($\mathcal{R}_t$). The originality is that we use different methods to compute the reproduction number, and allow the use to select incidence data from various geographical locations (country, region, department) or from different types (infections, deaths, hospitalisations, ICU admissions). Furthermore, the Rt2 interface is built to minimise the risk of misinterpretation related to variations in sampling effort, delays between incidence data and epidemic state, and serial interval. This tool can directly be used by public health authorities or the general public.

Acknowledgements

We are very grateful to the the itrop HPC (South Green Platform) of IRD Montpellier for hosting this application (more details on https://bioinfo.ird.fr/).

The ETE modeling team is composed of Samuel Alizon, Thomas Bénéteau, Marc Choisy, Gonché Danesh, Ramsès Djidjou-Demasse, Baptiste Elie, Yannis Michalakis, Bastien Reyné, Quentin Richard, Christian Selinger, Mircea T. Sofonea.

This work was partly supported by the Occitanie region and the ANR (PHYEPI project). GD is funded by the Fondation pour la Recherche Médicale (FRM grant number ECO20170637560).

Bibliography

R. M. Anderson and R. M. May. *Infectious Diseases of Humans. Dynamics and Control*. Oxford University Press, Oxford, 1991. [p1]

A. Cori, N. M. Ferguson, C. Fraser, and S. Cauchemez. A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. *Am J Epidemiol*, 178(9):1505–1512, 2013. doi: 10.1093/aje/kwt133. [p4]

S. Flaxman, S. Mishra, A. Gandy, H. J. T. Unwin, T. A. Mellan, H. Coupland, C. Whittaker, H. Zhu, T. Berah, J. W. Eaton, M. Monod, A. C. Ghani, C. A. Donnelly, S. Riley, M. A. C. Vollmer, N. M. Ferguson, L. C. Okell, and S. Bhatt. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, 584(7820):257–261, Aug. 2020. doi: 10.1038/s41586-020-2405-7. [p1]

X. He, E. H. Y. Lau, P. Wu, X. Deng, J. Wang, X. Hao, Y. C. Lau, J. Y. Wong, Y. Guan, X. Tan, X. Mo, Y. Chen, B. Liao, W. Chen, F. Hu, Q. Zhang, M. Zhong, Y. Wu, L. Zhao, F. Zhang, B. J. Cowling, F. Li, and G. M. Leung. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat Med*, 26(5):672–675, 2020. doi: 10.1038/s41591-020-0869-5. [p2]

E. O.-O. Max Roser, Hannah Ritchie and J. Hasell. Coronavirus pandemic (COVID-19). *Our World in Data*, 2020. [p3]

H. Nishiura, N. M. Linton, and A. R. Akhmetzhanov. Serial interval of novel coronavirus (COVID-19) infections. *Int J Infect Dis*, 93:284–286, 2020. doi: 10.1016/j.ijid.2020.02.060. [p2, 5, 6]

T. Obadia, R. Haneef, and P.-Y. Boëlle. The R0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks. *BMC Med Inform Decis*, 12(1):147, 2012. doi: 10.1186/1472-6947-12-147. [p3, 4]

M. T. Sofonea, B. Reyné, B. Elie, R. Djidjou-Demasse, C. Selinger, Y. Michalakis, and S. Alizon. Epidemiological monitoring and control perspectives: application of a parsimonious modelling framework to the COVID-19 dynamics in France. *medRxiv*, page 2020.05.22.20110593, May 2020. doi: 10.1101/2020.05.22.20110593. [p2, 5]

R. N. Thompson, J. E. Stockwin, R. D. van Gaalen, J. A. Polonsky, Z. N. Kamvar, P. A. Demarsh, E. Dahlqwist, S. Li, E. Miguel, T. Jombart, J. Lessler, S. Cauchemez, and A. Cori. Improved inference of time-varying reproduction numbers during infectious disease outbreaks. *Epidemics*, 29:100356, Dec. 2019. doi: 10.1016/j.epidem.2019.100356. [p4]

J. Wallinga and M. Lipsitch. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences*, 274(1609):599–604, 2007. [p1, 2]

J. Wallinga and P. Teunis. Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures. *Am J Epidemiol*, 160(6):509–516, 2004. doi: 10.1093/aje/kwh255. [p3, 4]

L. F. White and M. Pagano. A likelihood-based method for real-time estimation of the serial interval and reproductive number of an epidemic. *Stat Med*, 27(16):2999–3016, 2008. doi: 10.1002/sim.3136. [p3]

L. F. White, J. Wallinga, L. Finelli, C. Reed, S. Riley, M. Lipsitch, and M. Pagano. Estimation of the reproductive number and the serial interval in early phase of the 2009 influenza A/H1N1 pandemic in the USA. *Influenza Other Resp.*, 3(6):267–276, 2009. doi: 10.1111/j.1750-2659.2009.00106.x. [p3]

Bastien Reyné
MIVEGEC, CNRS, IRD, Université de Montpellier
911 avenue Agropolis, 34394 Montpellier
France
ORCiD 0000-0002-4899-1240
bastien.reyne@ird.fr

Gonché Danesh
MIVEGEC, CNRS, IRD, Université de Montpellier
911 avenue Agropolis, 34394 Montpellier
France
ORCiD 0000-0001-8688-7666
gonche.danesh@ird.fr

Samuel Alizon
MIVEGEC, CNRS, IRD, Université de Montpellier
911 avenue Agropolis, 34394 Montpellier
France
ORCiD 0000-0002-0779-9543
samuel.alizon@cnrs.fr

Mircea T. Sofonea
MIVEGEC, Université de Montpellier, CNRS, IRD
911 avenue Agropolis, 34394 Montpellier
France
ORCiD 0000-0002-4499-0435
mircea.sofonea@umontpellier.fr