A pitfall in estimating the effective reproductive number R_t for COVID-19

Markus Petermann[a] and Daniel Wyler[b]

[a]Zentralstrasse 47, CH-8003 Zürich, Switzerland
[b]Institut für Theoretische Physik, Universität Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

June 15, 2020

Abstract

The effective reproductive number R_t of COVID-19 is determined indirectly from data that are only incompletely known. Approaches based on reconstructing these data by sampling time lags from suitable distributions introduce noise effects that can result in distorted estimates of R_t. This, in turn, may lead to misleading interpretations of the efficacy of the various measures taken to limit COVID-19 transmission. We discuss in some detail a study used for real time monitoring of the reproductive number in Switzerland; see https://ncs-tf.ch/en/situation-report.

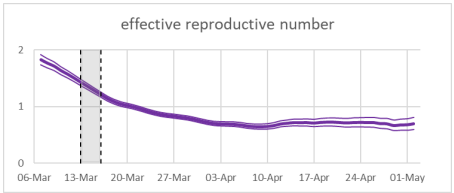

Figure 1: Effective reproductive numbers in Switzerland as shown in the real time monitoring on May 13. Mean and 95% uncertainty intervals are estimated on confirmed cases [1]. The shaded region is the onset of strong public health measures.

We argue that the method used to derive the above curve is systematically flawed and leads to an underestimation of the efficacy of the lockdown. The method adopted by the Robert Koch Institute suffers from similar deficiencies, their impact is however smaller.

Introduction

The daily varying effective reproductive number R_t is often used to monitor the spread of epidemic diseases such as COVID-19. It measures the expected number of secondary infections on day t due to a single infected individual and is given by

$$R_t = \frac{I_t}{\sum_{n>0} I_{t-n} w_n},$$

where I_t is the number of new infections on day t (and equally I_{t-n} the number of infections on day $t-n$ and so on), and the w_n are the infection intensities, i.e. w_n is the probability that a secondary infection was contracted from a person who got infected n days earlier. While the infection intensities can be fitted to available data, the I_t are not observable directly, unless representative proportions of the population were tested on a daily basis. Therefore, they need to be inferred indirectly from some other data.

There exist different schemes to reconstruct the I_t from the data, like the classical statistical inference methods. In the present article we concentrate on schemes that are based on the idea of a "mechanical" reconstruction of the infection data by sampling time lags of observed data from suitable distributions. In the context of the current COVID-19 pandemic, such schemes have been implemented in different ways by various groups in different countries.

In the following section, we present these schemes and show how they systematically introduce noise into the true data. In the remaining sections we examine the impact of the noise on the reproductive numbers calculated from these data.

The reconstruction scheme and its smoothing effect

We exemplify the scheme and how it introduces noise into the data by the version implemented by Scire et al [2], that is very similar to the implementation by Abbott s, Hellewell J, Thompson RNN et al [3]. For the sake of clarity, we limit our exposition of the scheme to the observables C_t, the number of confirmed cases on day t. The other observables used analogously by the group are hospitalizations and deaths. Moreover, as our focus is on the days around March 17, when the lockdown started, and as more than two months have passed since then, we restrict our exposition to those days of infection where all infections can be assumed to be confirmed by the actual date of the monitoring. According to the parameters used by the group, 95% of the cases are confirmed within 20 days after infection. For their method of extending the reconstruction to later days we refer again to their article.

Now, let X denote the incubation time of a randomly drawn case, i.e. the time between infection and symptom onset, and analogously Y the time between symptom onset and confirmation. The distributions of X and Y result from fitting to available data; see [2] and references therein. Then for every confirmed case a one samples independently a x'_a from the distribution of X and a y'_a from the distribution of Y.

The *reconstructed* infection day i'_a of this case a is then simply the day when the virus infection was confirmed minus $(x'_a + y'_a)$. Counting the number of cases that fall now on day t gives the reconstructed I_t, that we denote by I'_t. The reproductive numbers calculated from these I'_t are denoted by R'_t.

That this scheme introduces noise into the true data is seen as follows. We denote the true infection day by i_a, the true incubation period by x_a and the true time between symptom onset and confirmation by y_a. Then we have $i'_a = i_a + x_a + y_a - x'_a - y'_a$. As the sampled x'_a and y'_a are independent of the true x_a and y_a (because we don't know these; we just know that they are approximately distributed like X and Y, respectively), the reconstructed i'_a equals the true "signal" i_a plus some "noise" $d'_a = x_a + y_a - x'_a - y'_a$.

As we will see in the now following examples, this results in a smoothing of the infection number statistics[1], which in turn, under certain circumstances, has a significant impact on the reproductive numbers calculated from it.

An illustrative example

The following example illustrates the effect of this scheme on the reconstructed reproductive number. Assume that $I_1 = 128$ and $R_t = 2$ for $t \leq 6$ and $R_t = 0.8$ for $t \geq 7$, and that infectiousness is limited to the day after infection, i.e. $w_1 = 1$. This yields the 'true' infection numbers and reproductive numbers which are illustrated by the red curves in Fig. 2. For the *reconstructed* data we take X and Y both to be gaussian with mean 5 and standard deviation 1. Thus, the "noise" is also gaussian with mean 0 and standard deviation 2. With this "noise", the scheme results in the average in the corresponding blue curves.

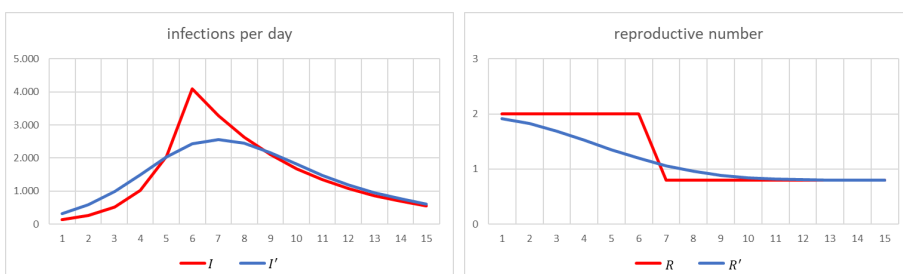

Figure 2: 'True' (red) and reconstructed (blue) infection and reproductive numbers

Of course, the blue curve R' is prone to lead to wrong decisions: If a lockdown caused the sharp decline in R from day 6 to day 7, then the blue curve may suggest that its impact was much less important and that most of the reduction

[1] From a mathematical point of view, this is clear: If all the I_t are large, then we have $C \approxeq I * p$, where $p_n = P(X + Y = n)$, and $I' \approxeq C * \tilde{p}$, where $\tilde{p}_n = p_{-n}$, i.e. $I' \approxeq I * p * \tilde{p}$.

was achieved already before the lockdown. It might even lead to the conclusion that the lockdown was not needed at all and that softer measures in force already before day 6 have had a sufficient effect, where in reality they had no effect at all. This in turn could lead to the conclusion, that the pandemic can be kept under control by adhering to soft measures only.

The reproductive number in Switzerland around March 17

We now turn to the situation in Switzerland around March 17, 2020. A calculation based on distributions for the time lags X and Y and the infection intensity as described in [2] yields the following result:

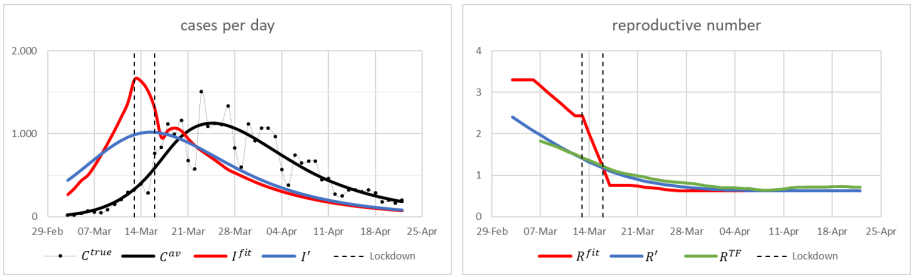

Figure 3: Cases and reproductive numbers in Switzerland around March 17. See text below for the definition of the curves.

Unlike in the illustrative example above, we cannot start from 'true' numbers of new infections. We instead choose[2] the numbers of new infections, denoted by I^{fit} (red curve), in such a way that the resulting expected numbers of confirmed cases C^{av} (solid black curve) fit well the black dots C^{true} which show the actually reported data of confirmed cases [4]. Here, to get C^{av} from the numbers of new infections, we shift forward the infection day of each such case by sampling independently from X and Y.

Given I^{fit}, we proceed as in the illustrative example, but, of course, with "noise" according to these X and Y instead of the gaussians used there. This gives the *reconstructed* numbers of new infections I' (blue curve). Using the infection intensities w_n from [2], we calculate the corresponding reproductive numbers R^{fit} and R'. The latter matches well the green curve R^{TF} that shows the estimated mean reproductive numbers, as published on the website [1] on May 13.

The remarks made above on the illustrative example apply also here. We note also that our red curve of reproductive numbers is in quite good agreement with the results of the inference analysis reported in [5, 6].

[2]See appendix A for the algorithm used to infer I^{fit}.

An alternative approach to reconstruction

Contrary to Switzerland, where, as far as we know, the date of symptom onset of the single cases is not collected systematically, this information is available for the majority of the cases in Germany. For the sake of clarity, we assume here that it is known for all cases.[3] Then, the Robert Koch Institute applies the following simpler scheme [7]:[4] The reconstructed infection day i''_a of a case a is the day of its symptom onset minus m, where m is the average incubation time. Counting the number of cases that fall now on day t gives the reconstructed I_t that we denote by I''_t. The reproductive numbers calculated from these I''_t are denoted by R''_t.

Of course we can view the subtracted value m as sample from the distribution of the constant time lag $X = m$. Therefore, this scheme is at least formally very similar to the one adopted by [2].

From the above discussion it is now clear that the so reconstructed infection times $i''_a = i_a + d''_a$ with "noise" $d''_a = x_a - m$ lead also to a smoothing of the infection number statistics and thus to misleading reproductive numbers. But it is also intuitively clear, that the impact is significantly smaller. This is confirmed by the following calculation.

Assume that the above I^{fit} are the true new infections per day, and that the distribution of the incubation time X is as above. Denote by S^{av}_t the resulting expected number of cases with symptom onset on day t. Then $I''_t = S^{av}_{t+m}$. Assuming moreover that also the infection intensities are as above, we get the following result:

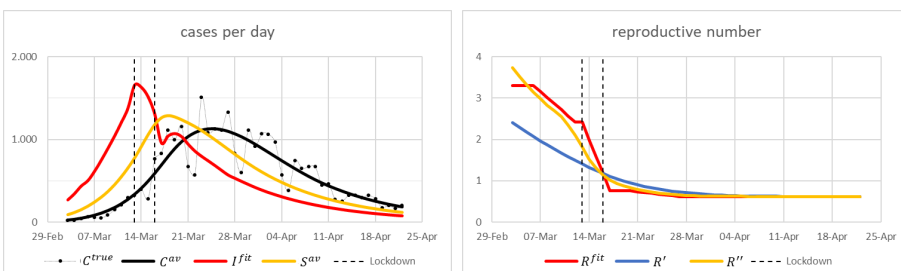

Figure 4: Left: 'True' infection numbers (red) and their corresponding expected numbers of cases with symptom onset per day (yellow) and of confirmed cases per day (black). Right: The corresponding 'true' reproductive numbers (red) and their reconstructions according to the schemes [7] (yellow) and [2](blue).

We stress that the knowledge of the dates of symptom onset is an advantage,

[3] We refer to [7] for the method applied to the cases with no known date of symptom onset.

[4] As our focus is again on the days around March 17, we restrict our exposition to those days of infection where all infections can be assumed to be confirmed by the actual date of the monitoring. For a method to extend the reconstruction to later days we refer to [7].

as adopting the same scheme but with dates of confirmation instead of symptom onset, would introduce the "noise" $x_a + y_a - m'$, where m' is the average time between infection and confirmation, into the true data, and this is clearly more "noise" than in the scheme based on dates of symptom onset.

Conclusion

In this note we have reexamined a type of schemes used to estimate the effective reproductive numbers R_t for COVID-19 by the example of two versions actually in use [2, 7]. These schemes are based on reconstruction of not directly observable data by sampling time lags of observed data from suitable distributions. Noise effects inherent in these schemes smooth the number statistics of the true data. The analysis of thus smoothed number statistics yields stable results and is easier to handle than classical inference methods as applied in [5, 6]. However, under certain circumstances like the current COVID-19 pandemic, the introduced noise effects dominate the information contained in the true data and lead to erroneous interpretations. The simpler approach adopted by [7] performs better than the one used in [2]. Moreover, we point out that adequate knowledge of the date of symptom onset is an advantage.

Appendix A: Fitting algorithm

Our algorithm leading to the curves I^{fit} and R^{fit} of Fig. 3 can be viewed as a deterministic or non-Bayesian version of the analysis performed in [6] to estimate the effects of non-pharmaceutical interventions on COVID-19 in Europe. We give in the following the details of our algorithm.

As in the main text, we denote by $w_n, n \geq 1$, the infection intensities and by R_t the effective reproductive number on day t.

We assume that no infections were present in Switzerland before $t_0 =$ February 24, 2020, i.e. for $t < t_0$ we set $I_t = 0$.

The period between t_0 and $t_1 =$ March 06 is taken as the burn-in phase, where we assume that each day λ infections are imported and that these already lead to secondary infections, i.e. for $t = t_0, t_0 + 1, \ldots, t_1 - 1$, we set

$$I_t = R_t \sum_{n \geq 1} I_{t-n} w_n + \lambda.$$

After t_1, we assume that the "internal" secondary infections dominate such that imported infections can be neglected, i.e. for $t \geq t_1$ the number of new infections I_t are given by

$$I_t = R_t \sum_{n \geq 1} I_{t-n} w_n.$$

In particular, we assume throughout that the number of new infections is a deterministic function of the imported infections and of the reproductive numbers, i.e. that random fluctuations in the number of infections can be neglected.

A posteriori, these choices turn out to be consistent. According to our fit of the new infections per day, we get approximately 500 new cases for March 06. Thus, the neglected fluctuations are with a fairly large probability within twice the square root of 500, i.e. within ± 45, while the fitted value for λ is approximately 47.

We denote by Z the (random) number of days between infection and confirmation by virus test. Then the number of cases that are tested on day t and confirmed positive is given by

$$C_t = \sum_{s \le t} \sum_{i=1}^{I_s} \delta_t(s + Z_{s,i}),$$

where $Z_{s,i}$ are independent and identically distributed (i.i.d.) copies of Z and $\delta_t(x) = 1$, if $x = t$, and 0 otherwise.

We denote the probability measure of the common probability space of all these random variables by P, and expected values with respect to P by E. For the expected number $C_t^{av} = E[C_t]$ of confirmed cases on day t we get

$$C_t^{av} = C_t^{av}(R, \lambda) = E\big[\sum_{s \le t} \sum_{i=1}^{I_s} \delta_t(s + Z_{s,i})\big] = \sum_{s \le t} I_s P(Z = t - s),$$

where we make explicit, that this series depends on the series of reproductive numbers, $R = (R_t)$, and on the number λ of daily imported infections during the burn-in phase.

Next, we denote by C_t^{true} the true number of cases confirmed positive on day t. For determining the best fit, we use the L^2-distance expression in the form[5]

$$\chi^2(R, \lambda) = \sum_{t \le T} \big(\sum_{s \le t} C_s^{av}(R, \lambda) - \sum_{s \le t} C_s^{true} \big)^2,$$

where the last information taken into account is that of $T =$ April 27, 2020, the last number of confirmed cases not affected by the first relaxation of measures of April 27.

Then we set

$$(R^{fit}, \lambda^{fit}) = \operatorname*{argmin}_{(R,\lambda) \in A \times \mathbb{R}} \chi^2(R, \lambda),$$

where A is a suitable set of admissible R-curves, encoding a priori assumptions to reduce the number of parameters to be fitted, which is necessary, because minimization over all R-curves would lead to overfitting. The a priori assumptions made by us are as follows.

It is reasonable to assume that the R_t do not fluctuate largely from day to day, except possibly around a limited number of "change points", where a

[5]The reason, why we choose the L^2-distance between the cumulative counts and not between the daily counts, is that testing shows a weekly pattern due to non-corona related effects. Its impact shows less in the cumulative view. We refer to appendix B for a more direct way to take these effects into account.

change of measures is put in force. Therefore, we have chosen A to consist of piecewise linear R-curves. As we concentrate on the days around the lockdown, we limit ourselves to a minimal number of pieces. The natural choices for the change points are t_2 = March 12 (the last day of "normal life"), t_3 = March 13 (the press conference announcing first drastic measures), t_4 = March 17 (lockdown), t_5 = March 20 (ban of gatherings > 5 people), t_6 = March 27 (intermediate change point). After that, we assume that the reproductive number stays constant. (Here, the intermediate change point is introduced to test that after the "ban of gatherings > 5 people" the reproductive number stays effectively approximately constant. If this was not the case, then the choice of change points needed revision.) Moreover, the case numbers are too low to estimate the reproductive numbers in the burn-in phase. Therefore, we rely for this period on typical estimates found in the literature and set $R_t = 3.30$, for t between February 24 and March 05. Finally, as up to March 20 measures are strengthened, but never relaxed, we require that the admissible R-curves do not increase before t_5. Summarizing, we choose

$$\begin{aligned} A = \{R =&(R_t)_{t=\text{Feb 24},\ldots,\text{Apr 27}} : R \text{ is piecewise linear with change} \\ &\text{points at } t_1, \ldots, t_6 \text{ and } R_{\text{Feb 24}} = R_{\text{Mar 05}} = 3.30, \\ &R_{t_1} \geq R_{t_2} \geq \cdots \geq R_{t_5}, R_T = R_{t_6}\}, \end{aligned}$$

i.e. we fit 6 reproductive numbers, $R_{\text{Mar 06}}, R_{\text{Mar 12}}, R_{\text{Mar 13}}, R_{\text{Mar 17}}, R_{\text{Mar 20}}$, $R_{\text{Mar 27}}$, plus λ.

We would like to emphasize that the reproductive numbers resulting from this fitting procedure are quite unstable, in particular with respect to variations of the distributions of Z and of the infection intensities w_n that are not inferred here but estimated elsewhere [2, 8]. Therefore, R^{fit} should at most be considered a likely picture of reality. To get more confidence in the results, a thorough sensitivity analysis would be required, confidence intervals should be calculated (see e.g. the above mentioned [6]), and improvements to the analysis like the ones described in appendix B schould be incorporated.

Appendix B: Suggested improvements to the analysis

The simple analysis described above only gives a first indication of the reproductive numbers around the lockdown. In order to more precisely quantify the effects of the measures on the reproductive number, this analysis should be extended to include at least the following elements.

Find change points in a systematic way

Instead of choosing the change points by a priori reasoning as above, one can infer them from the data by replacing the above A by

$$\begin{aligned} A = \{R =&(R_t)_{t=\text{Feb 24},\ldots,\text{Apr 27}} : \exists(\tau_i)_{i=1,\ldots,6} \text{ with } t_0 < \tau_1 < \cdots < \tau_6 \text{ such that} \\ & R \text{ is piecewise linear with change points at } \tau_1, \ldots, \tau_6 \text{ and} \\ & R_{\text{Feb 24}} = R_{\text{Mar 05}} = 3.30, R_{\tau_1} \geq R_{\tau_2} \geq \cdots \geq R_{\tau_5}, R_T = R_{\tau_6}\}. \end{aligned}$$

Weekend effects

Weekend effects can be taken into account by substituting the above C_t by

$$C_t = \sum_{s \leq t} \sum_{i=1}^{I_s} \delta_t(s + Z_{s,i} + \zeta_{s+Z_{s,i},i}).$$

Here, $\zeta_{t,i}, i \geq 1$, are i.i.d. copies of $\zeta_{wd(t)}$, where $wd(t) \in \{\text{Mon, Tues}, \ldots, \text{Sun}\}$ denotes the weekday of t. $P(\zeta_{\text{Sat}} = 2)$ is then the probability, that a test that without weekend effects would have been performed on a Saturday was actually performed on the following Monday. These probabilities could be estimated directly from the confirmed cases, or they can be fitted together with the other parameters. For the expected number of confirmed cases on day t we get then

$$C_t^{av} = E\big[\sum_{s \leq t} \sum_{i=1}^{I_s} \delta_t(s + Z_{s,i} + \zeta_{s+Z_{s,i},i})\big] = \sum_s K_{t,s} I_s,$$

with

$$K_{t,s} = \sum_r P(Z = r - s) P(\zeta_{wd(r)} = t - r).$$

Impact of quarantining on infection intensities

A natural guess for the infection intensities w_n^{qu} of those that quarantine starting the day after the onset of symptoms is

$$w_n^{qu} = \frac{w_n P(X \geq n)}{\sum_{k \geq 1} w_k P(X \geq k)}, n \geq 1,$$

where X denotes the incubation time.

The average infection intensities at day t are then

$$\gamma_t w_n + (1 - \gamma_t) w_n^{qu},$$

where $\gamma_t \in [0, 1]$ is the proportion of infectives on day t that do not quarantine although symptomatic. Thus, quarantining can be taken into account by minimizing over $A \times \mathbb{R} \times G$, where G consists of an adequate set of admissible γ-curves.

Acknowledgements

We thank Nicola Kistler for asking the right question and Erik Böttger, Jürg Fröhlich and Emanuel Wyler for helpful discussions.

Disclosure statement

The authors have no financial support nor any other potential conflict of interest relevant to this article.

Correspondence

wyler@physik.uzh.ch

References

[1] https://bsse.ethz.ch/cevo/research/sars-cov-2/real-time-monitoring-in-switzerland.html
Site visited on May 13, 2020

[2] Scire J et al. Reproductive number of the COVID-19 epidemic in Switzerland with a focus on the Cantons of Basel-Stadt and Basel-Landschaft. Swiss Medical Weekly 2020;150:w20271
DOI: https://doi.org/10.4414/smw.2020.20271

[3] Abbott S, Hellewell J, Thompson RN et al. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts [version 1; peer review: awaiting peer review] Wellcome Open Research 2020, 5:112
DOI: https://doi.org/10.12688/wellcomeopenres.16006.1

[4] https://github.com/daenuprobst/covid19-cases-switzerland

[5] Lemaitre J et al. Assessing the impact of non-pharmaceutical interventions on SARS-CoV-2 transmission in Switzerland. Swiss Medical Weekly 2020;150:w20295
DOI: https://doi.org/10.4414/smw.2020.20295

[6] Flaxman S et al. Estimating the effets of non-pharmaceutical interventions on COVID-19 in Europe. Nature
DOI: https://doi.org/10.1038/s41586-020-2405-7 (2020)

[7] an der Heiden M, Hamouda O. Schätzung der aktuellen Entwicklung der SARS-CoV-2-Epidemie in Deutschland - Nowcasting.
Epid Bull 2020;17:10-15 DOI 10.25646/6692.2.
See https://edoc.rki.de/handle/176904/6650.4 for updated version

[8] Nishiura H, Linton NM, Akhmetzhanov AR. Serial interval of novel coronavirus (COVID19) infections. Int J Infect Dis. 2020 Apr 1;93:284–6.