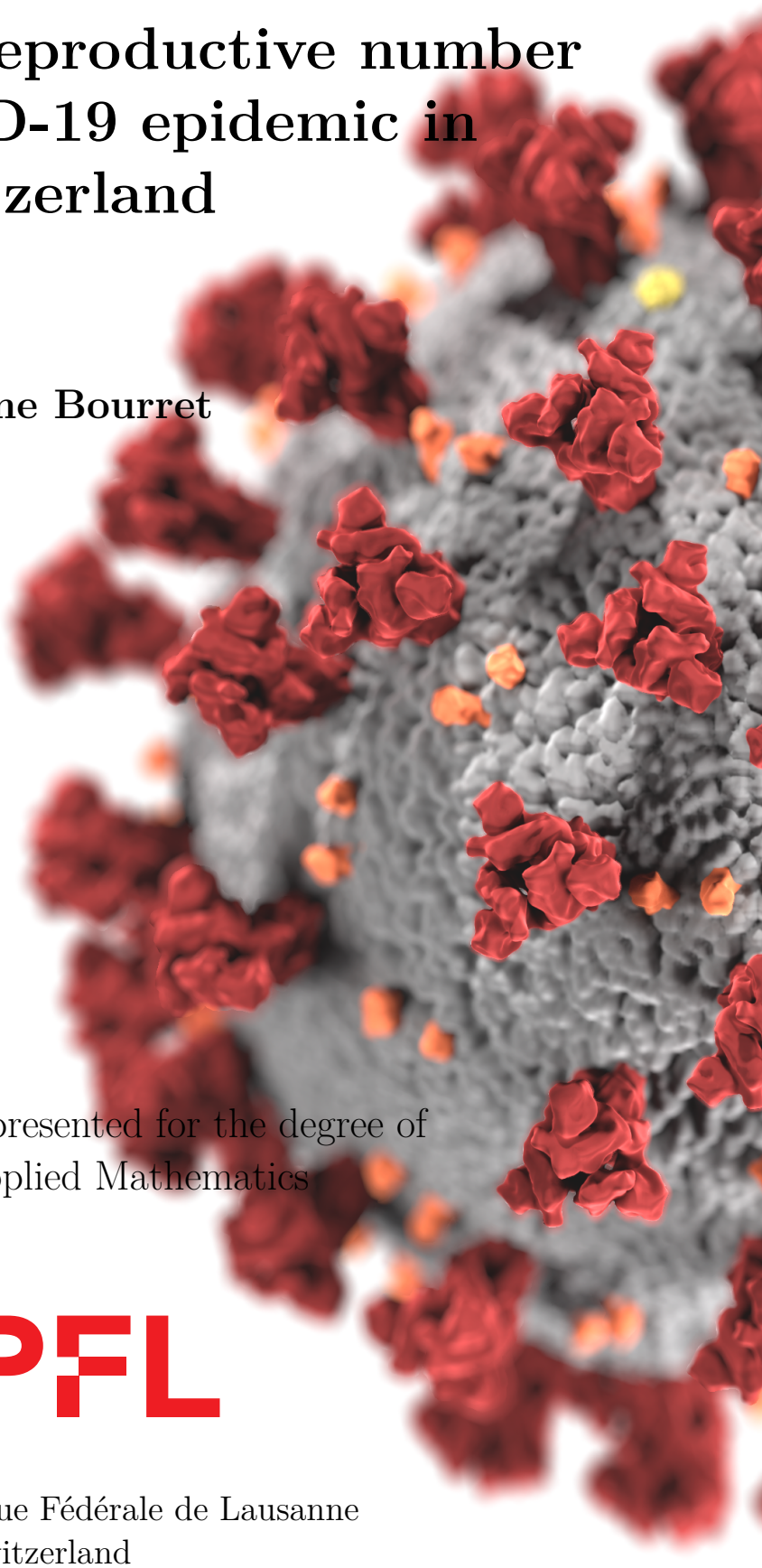
Estimating the reproductive number of the COVID-19 epidemic in Switzerland



Antoine Bourret

A semester project presented for the degree of
Master in Applied Mathematics

École Polytechnique Fédérale de Lausanne
Switzerland
January 17, 2021

Estimating the reproductive number of the COVID-19 epidemic in Switzerland

Antoine Bourret*

January 17, 2021

Abstract

Estimates of the reproduction number R_t are of great interest in monitoring the course of an epidemic and potential changes in disease transmission. The COVID-19 pandemic has evolved rapidly and has been monitored by state authorities to understand how health measures such as home quarantine and social distancing can contain the outbreak. The use of Cori et al. estimation method of R_t has proven to be very effective, and we propose and discuss in this report two alternative approaches, one based on the generation interval, the other on the epidemiological SIR model. In both methods, R_t is modeled as a smooth function of time through the use of generalized additive models. We also provide methods to retrieve the actual incidence events from the observed curve of new daily cases, given an incubation period. These methods prove their value both on smooth data and on data with strong weekly patterns. We estimate R_t for CODID-19 in Switzerland from March to the end of December 2020, and evaluate the effectiveness of health measures.

1 Introduction

The evolution of an epidemic can be tracked using its reproductive number, denoted R_t, which represents the average number of secondary cases caused by the infection of an individual at time t. In order for sanitary measures to effectively contain the spread of the epidemic, a good estimate of the reproduction number is required, as well as the uncertainties for the estimates. Monitoring R_t can help detect changes in the transmission of the disease. Estimates of R_t are mostly derived from reported quantities that are daily available, such as the incidence, death and recovered case curves. As a results, the symptom onset and reporting delay distributions must be taken into account in order to avoid inaccuracies and lagging errors. Real-time estimation of the reproduction number has been approached by the method of Cori et al. [2013], whom has proved to be statistically robust and very simple to use and implement. Other methods, such as those of Wallinga and Teunis [2004] and Bettencourt and Ribeiro [2008] are also suited for a retrospective analysis and the estimation of the cohort reproductive number, which is the expected number of secondary cases that an individual infected at time t will infect as a result of the progression of his or her infection [Fraser, 2007].

In this report, we review and extend some of the proposed models, and evaluate their effectiveness on simulated data. The first approach uses a generalized additive model (GAM) to model the reproduction number as a smooth function of time with spline functions. It requires the curve of incidence events and the generation interval. The second approach is based on the epidemiological SIR model, and interprets it as a GAM through the equations of infection and

*Electronic address: antoine.bourret@epfl.ch

recovery updates. This model does not need estimates for the generation interval, which might be hard to obtain in practice, but the curve of deceased and recovered individuals. Estimates of the transmission and removal rates are used to estimate the reproductive number. We also propose multiple deconvolution methods in order to deal with delays for the incubation and recovery periods, and to recover the true incidence curve as precisely as possible. These methods will be tested on smooth and noisy data, with a strong weekly pattern characterized by a drop in the number of new cases during weekends.

Finally, we apply these methods for the analysis of the COVID-19 epidemic in Switzerland, from March to December 2020. We evaluate the effectiveness of health measures and compare the outcome of the two proposed estimation methods of R_t.

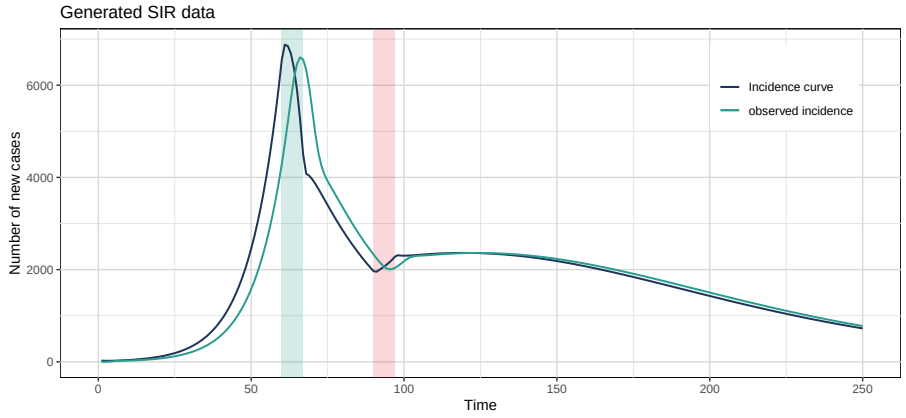
2 Simulated dataset

Through out this report, we will test the effectiveness of the proposed and reviewed methods on an artificial dataset that is proposed in Gostic et al. [2020], and is generated from a susceptible-exposed-infected-recovered (SEIR) model. Figure 1 shows the true and observed incidence curves, as well as the reproductive number. The simulation is performed on a time window of 300 days, but only the first 250 are shown for clarity. The residence times in compartments E and I of the SEIR model follow a Poisson distribution with a mean of 4 days. The generation interval is the sum of these two residence times, and thus follow a Gamma distribution with scale 4 and shape 2.

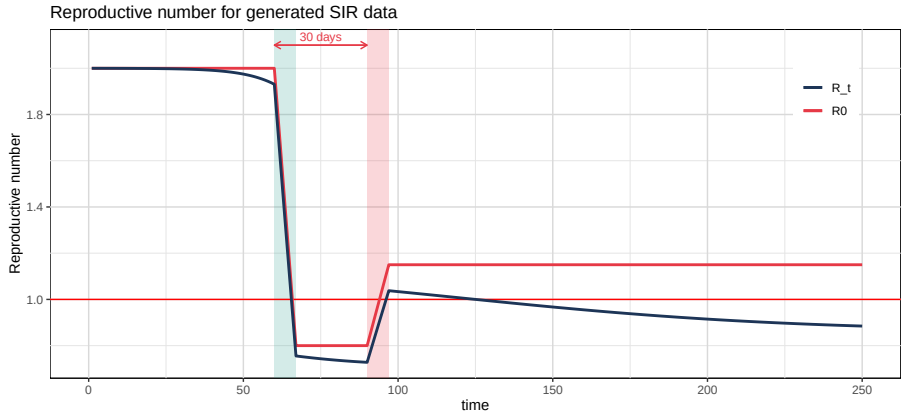
The basic reproduction number R_0 is the expected number of cases generated by one infected individual in a population that only have susceptible individuals. The true reproductive number R_t is obtained by multiplying R_0 by the proportion of susceptible people in the model, that is, when considering a SEIR model, by S_t/N where N is the population size. The R_0 used to generate the incidence curve is characterized by a large drop from 2 to 0.8 starting from day 60, and an increase from 0.8 to 1.15 from day 90, with a small linear transition period of seven days between them. Figure 1b shows R_0 and R_t, and these two transition periods. The critical threshold of 1 is shown in red. When R_t is above 1, the epidemic is growing at an exponential speed, while when R_t is below 1, the disease is exponentially declining.

The observed incidence curve is obtained by convolving the true incidence curve with the distribution for the incubation period, which is assumed to be the probability function of a discretized gamma with shape and scale parameters respectively set to 10 and 0.5, giving a mean of 5 days between the infection event and the time it is reported (symptom onset).

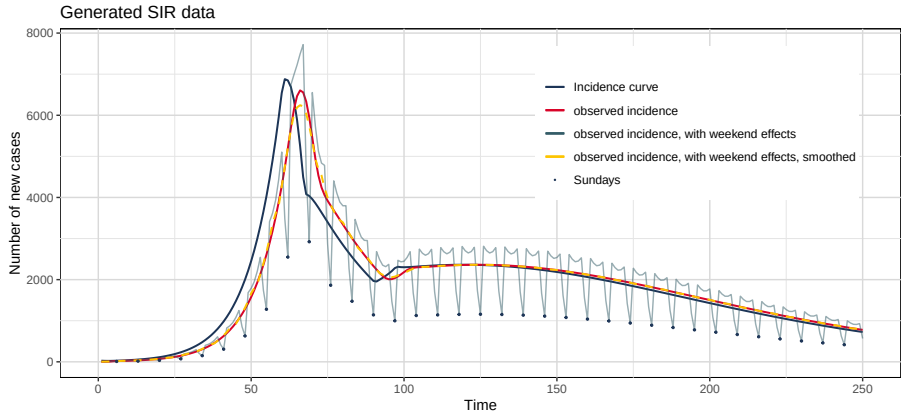
As real epidemic data are often much more noisy and can exhibit pattern in time, we decided to also compare our approaches on an incidence curve that shows weekly pattern. In particular, we constructed a reported curve such that patients are less likely to be reported during weekends than on the other days of the week. This is visible on the incidence curve with large drops for the new number of reported cases on weekends, as in Figure 1c. Smoothing the noisy incidence curve with a 7-day moving average might reduce this weekly pattern, but can lower the number of new cases during critical periods, compared to the curve with no weekly pattern. This is shown in Figure 1c, with the yellow dashed curve being smoother than the solid red one.



(a)



(b)



(c)

Figure 1: Synthetic dataset. (a) Incidence curves and observed incidence curves for a period of 200 days. (b) Reproductive number R_t (solid blue curve) and basic reproduction number R_0 (solid red curve) used to generate the data.(c) Weekends effects were also added in order to verify the efficiency of the method on more noisy data. The blue dotes correspond to the number of new cases in Sundays.

3 Deconvolution methods

3.1 Introduction

The spread of an epidemic is generally carried along with a series of measurement tools, that include the symptoms, report or death curves, and can provide estimates of the reproductive number on a daily basis. However, these curves are imperfect, as some individuals might be asymptomatic, and are thus not reported as infected. Thus, the reported number of cases at some point in time might not be representative of the number of individuals that were infected, due to delay between the infection event and the symptom onset, either as a result of lags in the development of detectable viral loads, or due to the presence of weekend effects. As such, estimating the reproductive number with these observed quantities can result in biased approximations, and an intermediate step is necessary in order to approximate the incidence curve I_t. Goldstein et al. [2009] proposed a method for reconstructing the incidence events by the deconvolution of the death curve, based on the Richardson–Lucy deconvolution scheme that is most commonly used in optics (Richardson [1972] and Lucy [1974]). As their method is based on the death curve, it is less prone to under-reporting and so provides very satisfying reconstructions. As discussed in Gostic et al. [2020], if one has access to an estimate for the distribution of the delays, the reproductive number R_t can be estimated by first inferring the incidence curve, and then using approximation tools for R_t on this inferred curve. In this context, Abbott et al. [2020b] proposed and developed the package EpiNow2 in `R` [Abbott et al., 2020a].

In this section, we review and propose several deconvolution methods that are based not on the death curve, but on the reported one. We denote the reported curve (also called the observed cases) by $\{Y_t\}_{t=1}^T$ and the incidence curve by $\{I_t\}_{t=1}^T$. These curves are assumed to be discretized in time, and are monitored and estimated on a daily basis. The distribution of the delay between the infection event and symptom onset (called the incubation period) is assumed to be known. As the main quantities of interest are computed daily, we suppose that the distribution for the incubation period is discretized, and denote it by $\{w_t\}_{t=1}^K$, where K is some upper bound on the number of days by which the infected individual will not develop any symptom. Lauer et al. [2020] estimated the incubation period of COVID-19 in the early phase of the epidemic, between January and February 2019 with 181 confirmed cases. Their estimate for the mean incubation period is of 5.1 days, with a 95% confidence interval between 4.1 and 5.8 days. They also provided an estimate of 0.01 for the proportion of individuals that will show symptoms after two weeks of infection. This incubation delay can be modeled with a discretized gamma or a negative binomial distribution. The latter has the advantage of being already discretized, and so does not suffer from a deviation of the mean and the variance that occur when discretizing the gamma distribution. As discussed in Gostic et al. [2020], wrong specification of the incubation period can results in biased estimates for the incidence curve, and thus for the reproductive number. They also discussed two simple methods for deconvolving the data: mean shift and convolution. The latter is discussed below.

3.2 Roll back deconvolution

The roll back deconvolution method redistributes the reported cases in previous days in order to get the incidence curve, using the distribution for the incubation period. Although this method is very simple, it adds an extra convolution step on top of the reported cases curve, which is undesirable when the incubation mean is large. Roll back deconvolution works as follows: for each day t, a sample of size Y_t is generated from the distribution of the incubation period X (that is, $F_X(s) = w_s$). We denote this sample by $x_1^{(t)}, \dots, x_{Y_t}^{(t)}$, and redistribute it in the past,

by producing the following estimate of the true infection event:

$$\hat{I}_t = \sum_{k=1}^{K} \left(\sum_{s=1}^{Y_{t+k}} \mathbb{1}_{(x_s^{t+k}=k)} \right), \quad t = 1, \ldots, T. \tag{3.1}$$

Figure 2 illustrates the procedure, in the case where the incubation period follows a discretized gamma distribution with mean 5 days. The confidence intervals can be easily obtained by repeating the procedure a certain number of times, using bootstrapping techniques for example.

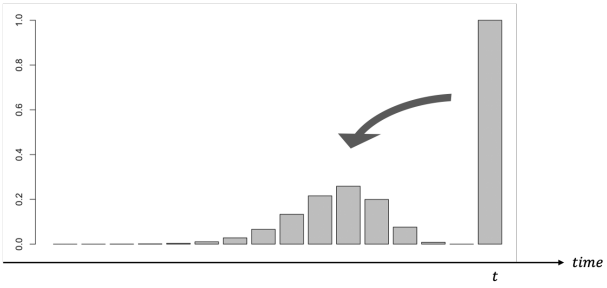
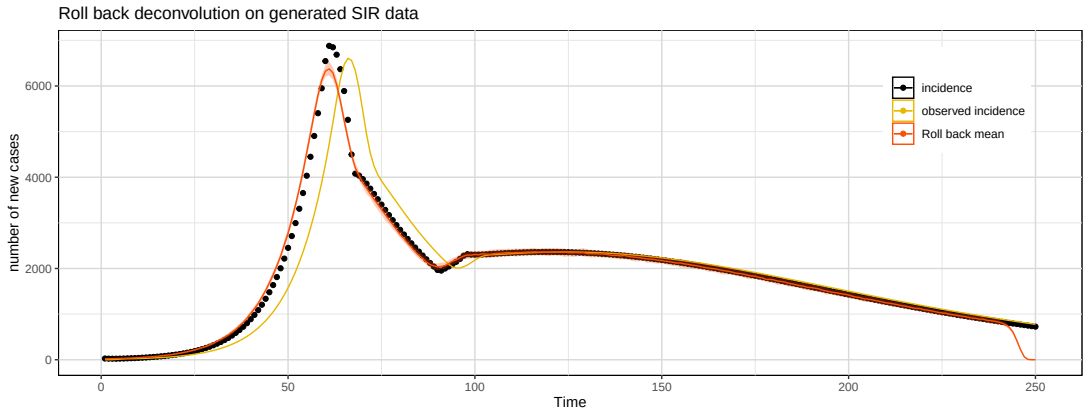


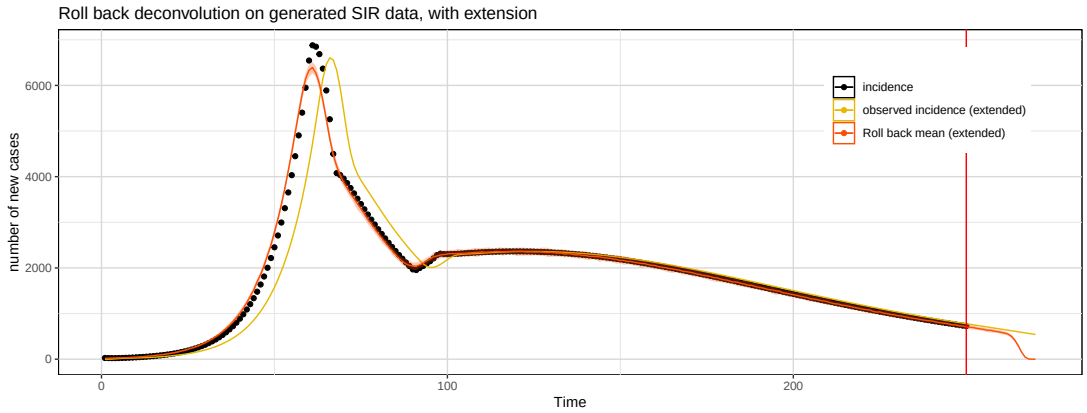
Figure 2: Illustration of the roll back method for estimating true infection events.

3.2.1 Results on artificial dataset

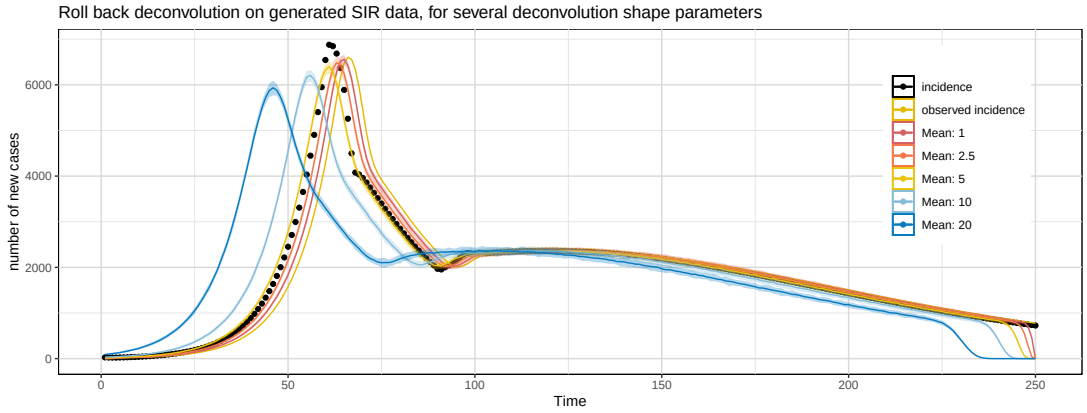
Results of this method are shown in Figure 3a. Although this method is simple and requires low computational resources, it does not catch the important parts of the epidemic curve. In particular, the large spike between days 50 and 70 is not well recovered, and the confidence bands are too narrow to include the observations. As the reported cases on a given day are spread out to the previous days in order to get the incidence for these days, the resulting estimate tends to be smoother than the true incidence curve. As noted in Gostic et al. [2020], this method can help to deal with observations that are very noisy and have a strong weekend effect. However, one of the reasons why this method is not used in practice for real-time monitoring of R_t is that it produces a large bias at the end of the reported case curve, with a large drop in the estimated incidence curve. This is because the last values rely on few observations, and are sampled from the tail of the distribution for the incubation period, which can be ill-specified. This effect is smaller as the mean incubation period is low, but does not become negligible in applications. One can tackle this issue by predicting the observed number of new cases in the near future, and including this in the roll back method. This is shown in Figure 3b, where a prediction step is done for the following 20 days using spline functions. To account for the uncertainty, one can use the confidence interval for the prediction of the next days of the reported curve in the bootstrapping method that is used to get the confidence intervals for $\hat{I}_t$. Although this solves the issue, one might question its effectiveness on less smooth data, for which the prediction step might not be representative of the future number of new reported cases. this motivates the use of more advanced and robust deconvolution methods.



(a)



(b)



(c)

Figure 3: Results of the roll back deconvolution method. Roll back deconvolution on the original dataset (a) and on the extended dataset with a predictive step at the end (b). In (c), we used incubation periods with different mean. The greater the mean, the smoother the estimates are and the larger is the drop of the estimates at the end. This implies that for incubation periods with a large mean, the extension step need to be done on a larger time window.

3.3 Deconvolution based on the optimization of a Poisson response

This methods provides a much more rigorous definition of the roll back deconvolution, by assuming that the observed event $Y_t + 1$ follows, conditionally on the incidence events $I_1, \dots, I_t$, a Poisson distribution with a weighted sum of these incidence events. More formally, we assume that $Y_{t+1}|I_1, \dots, I_t \sim \text{Pois}(\mu_{t+1})$, with $\mu_{t+1} = \sum_{s=1}^{K} w_s I_{t+1-s}$. The distribution F of the incubation period is taken as known ($F(s) = w_s$), so the likelihood for the true infection events can be written as

$$L(\{I_t\}_{t=1}^T) = \prod_{t=1}^{T} \frac{\mu_t^{Y_t} \exp(-\mu_t)}{Y_t!}$$

The log-likelihood is then

$$l(\{I_t\}_{t=1}^T) = \sum_{t=1}^{T} Y_t \log(\mu_t) - \mu_t - \log Y_t! = \sum_{t=1}^{T} \left[Y_t \log\left(\sum_{s=1}^{K} w_s I_{t-s} \right) - \sum_{s=1}^{K} w_s I_{t-s} - \log Y_t! \right],$$

which can be maximized with respect to the I_t, using conjugate gradient descent or the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm. To enforce $\hat{I}_t > 0$, one can set $I_t = \exp(\lambda_t)$, and then do the optimization with respect to the $\lambda_1, \dots, \lambda_T$. The variance for the maximum likelihood estimator can be computed with the information matrix. Indeed, if we let $\{I_t\}_{t=1}^T = \Theta$ be the set of parameters that we want to estimate, then the large-sample variance matrix for the parameters is

$$\text{var}(\Theta) = \mathcal{I}(\Theta) = \left[-E\left(\frac{\partial^2 l(\Theta)}{\partial \Theta \partial \Theta^T} \right) \right]^{-1}.$$

So as not to depend on an approximation of the hessian in the optimization algorithm, the derivatives and the hessian for the log-likelihood are the following:

$$\begin{aligned}
\frac{\partial l}{\partial I_i} &= \sum_{t=i+1}^{K+i} \left(\frac{Y_t}{\Lambda_t} - 1 \right) w_{t-i} \\
\frac{\partial^2 l}{\partial I_i^2} &= -\sum_{t=1}^{K} \frac{Y_t w_t^2}{\Lambda_{t+i}} \\
\frac{\partial^2 l}{\partial I_j \partial I_i} &= -\sum_{t=1}^{K-i+j} \frac{Y_{t+i} w_t w_{t+i-j}}{\Lambda_{t+i}^2} \quad \text{with } j < i
\end{aligned}$$

Note that one can write μ_t in compact vectorized form. Let $I = \{I_t\}_{t=1}^T \in \mathbb{R}^T$ be the vector of incidence events, and $\mu = [\mu_1, \dots, \mu_T]^T \in \mathbb{R}^T$ the vector of mean of the Poisson random vector. According to the definition of the model, we have that $BI = \mu$, with

$$B = \begin{bmatrix} w_1 & & & & \\ \vdots & \ddots & & 0 & \\ w_K & \dots & w_1 & & \\ & \ddots & & \ddots & \\ 0 & & w_K & \dots & w_1 \end{bmatrix} \in \mathbb{R}^{T \times T}.$$

This implies that the derivatives of the log likelihood can be written as

$$\frac{\partial l}{\partial I} = B\mu \odot 1/BI, \quad \frac{\partial^2 l}{\partial I \partial I^T} = -\mu B B^T 1/I^2,$$

where the inverse and $(\cdot)^2$ operations are taken element-wise for the vectors, and $\odot$ corresponds to the element-wise product of two vectors.

3.3.1 Results

Results of the method are shown in Figure 4, for two different optimization algorithms. Unlike the previous method, this one correctly catches the period where the number of new cases was maximal. The optimization is performed using the conjugate gradient and BFGS algorithms, with 100 iterations. The confidence intervals were not obtainable as the information matrix gives negative and positive values in the diagonal elements. This can be due to the fact that the optimization process did not converge to a local minimum. However, increasing the number of iterations seems not to solve the issue. Nevertheless, the true incidence curve is well recovered, except at the beginning where there are still some variations. As for the last days, the estimate might differ from the true curve, depending on the specification of the incubation profile. Indeed, a low probability of being reported a day after the infection will result in a low gradient for the penultimate day. Since the initial values given to the optimization algorithm correspond to the observed events, the last values are likely to be close to the initial values at the end of the optimization process. As in Section 3.2, one can make a prediction for some days beyond the last observed cases, and then use this method on the extended reported curve. One might however need to account for the uncertainty of the predictions in the optimization process. Although this method gives satisfying results, its effectiveness heavily relies on the quality of the observed data, as very noisy observations tend to produce wiggly estimates of I_t, when the optimization converges properly. To tackle this problem, one needs to smooth the data before estimating the incidence curve, but at the cost of having biased estimates, notably when the reported curve shows abrupt increases or decreases in the number of new cases. As the number of parameters equals the length of the observations, this method is more computationally expensive than the previous one. This is discussed in the next section, where we model the incidence curve as a smooth function of time, and add regularization. One final observation can be addressed to the optimization algorithm used, which might give different results and estimates, as shown in Figure 4.

The results of this deconvolution approach on more noisy data are shown in Section 3.5, and compared with the results of the two other methods.

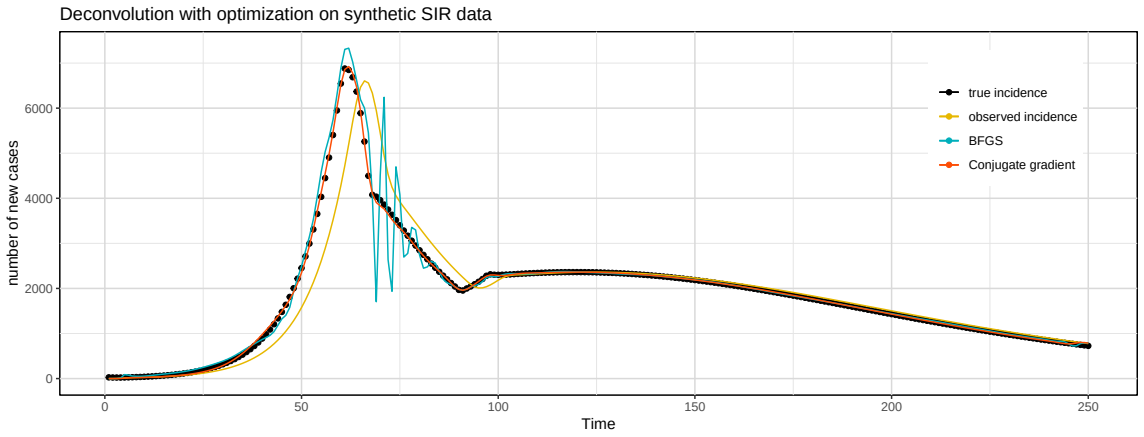


Figure 4: Results based on the optimization of a Poisson response, with conjugate gradient and BFGS algorithms. The BFGS optimization routine gives unrealistic estimates, with large variations when the transition period from $R_0 = 2$ to $R_0 = 0.8$ ended.

3.4 Modeling the incidence curve with splines

Here, we introduce the assumption that the incidence curve is a smooth function of time, that we model using spline functions. As in the previous section, we assume that $Y_{t+1}|I_1, \dots, I_t \sim \text{Pois}(\mu_{t+1})$, with $\mu_{t+1} = \sum_{s=1}^{K} w_s I_{t+1-s}$, and the distribution for the incubation period is known.

One can view this specification in terms of generalized linear models (GLM). For the notation, we refer to Davison [2003]. Let the reported cases Y_j follow a Poisson random variable with mean μ_j. Then the log-likelihood based on T samples $y_1, \dots, y_T$ can be written as

$$l(I) = \sum_{j=1}^{T} \log f_j(y_j; \mu_j) = \sum_{j=1}^{T} \{y_j \log(\mu_j) - \mu_j - \log y_j!\}$$

where we emphasized the fact that the log likelihood depends only on the unobserved events I, through the means for the Poisson distribution μ_j. Using the matrix notation introduced in the previous section, we assume that $\mu_j = [BI]_j = b_j^T I$, which implies that an identity link function is used. We want to maximize the log-likelihood, but the number of parameters exactly matches the number of observations, and so one might be tempted to restrict ourselves to a linear combination of splines basis for the incidence events I. More formally, we write $I = Z\gamma$, with Z the spline basis matrix in $\mathbb{R}^{T\times q}$ and γ the parameters associated with the spline basis, in $\mathbb{R}^q$. The log likelihood to be maximized is now $l(Z\gamma) - \lambda\gamma^T\Delta\gamma/2$, with a penalization matrix, usually of diagonal form. The estimation of λ is carried through cross-validation, and we use the Penalized-Iterative Weighted Least Square (P-IWLS) algorithm to estimate the parameters. In particular, starting from an initial vector of parameters γ_0, one writes the update as

$$\gamma_\lambda^{k+1} = \left((BZ)^T W BZ + \lambda\Delta\right)^{-1} (BZ)^T W (BZ\gamma_\lambda^k + (Y - BZ\gamma_\lambda^k)/BZ\gamma^k)$$

with $W = \text{diag}(BZ\gamma^k)$ the diagonal matrix in $\mathbb{R}^{T\times T}$, which depends on the current estimate of the parameters. B corresponds to the incubation period matrix defined as (3.3). When $\lambda = 0$, the standard errors can be obtained by taking the diagonal elements of the square matrix $((BZ)^T W BZ)^{-1}$, and the confidence intervals for the prediction are obtained using the delta method.

Defined like this, the estimates for the incidence curve are not guaranteed to be positive, which can lead to problems as the response variable is of Poisson type. This can be avoided by replacing negative values for the mean with very small positive ones. Also, this method does not require a particular choice for the initial values, which can be initiated simply based on the observed cases Y_t. The Poisson model assumes that the variance is equal to the mean, and this assumption does not hold when the data show clear signs of overdispersion, which can be assessed with a QQ-plot or others residual plots. To account for the increase or decrease of the variance of the response variables Y, one can use a quasi-Poisson approach or a negative binomial model. The quasi-Poisson model introduces a dispersion parameter ϕ such that $\text{var}(Y_t) = \phi\mu_t$, making the variance a linear function of the mean. The parameter ϕ can be estimated by

$$\hat{\phi} = \frac{1}{T-q} \sum_{t=1}^{T} \frac{(Y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}.$$

One must be careful when the reported numbers of cases are very low for a large proportion of the observations (for example with zero-inflated data), as the standard asymptotic results can then be questioned.

The negative binomial model uses a negative binomial density for Y_t,

$$f(y; k, \mu) = \frac{\Gamma(y+k)}{\Gamma(k)y!}\left(\frac{k}{\mu+k}\right)^k\left(1-\frac{k}{\mu+k}\right)^y, \quad y = 0, 1, 2, \ldots, \tag{3.2}$$

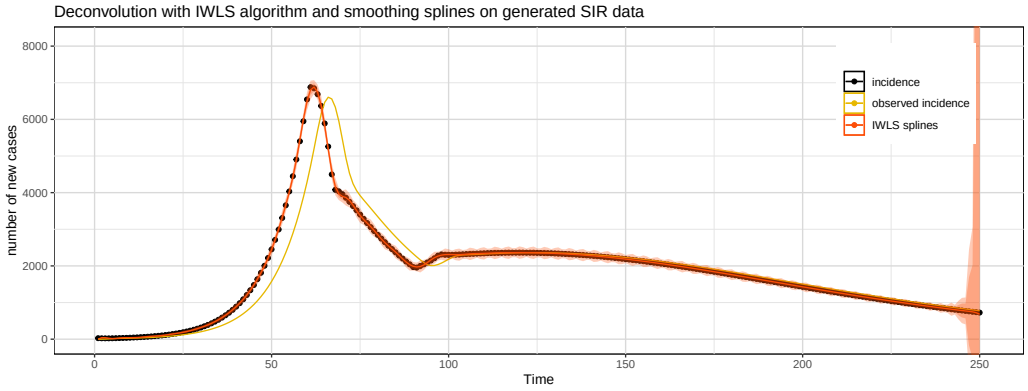
with $\mathbb{E}(Y) = \mu$ and $\text{var}(Y) = \mu(\mu/k + 1)$. In contrast with the quasi-Poisson model, the variance is a quadratic function of the mean, and as $k \to \infty$ we obtain a Poisson density, while when $k = 1$, we have a geometric density. The parameter k can be estimated with restricted maximum likelihood estimation (REML), for which more details can be found in Davison [spring 2020].

Although this deconvolution method seems computationally expensive, it is the fastest among the methods presented here, taking into account the computation time needed for the confidence intervals. It provides a smooth model of the incidence curve that does not convolve the reported curve, and allows for potentially overdispersed data. However, it requires the specification of the dimension q for the spline basis, and the regularization terms λ and Δ. The latter is often replaced by a diagonal matrix, or a tridiagonal matrix, with 2s on the main diagonal and -1s on the upper and lower diagonals. As for the dimension basis, one can set a sufficiently large value for q, and then readjust it with the equivalent degrees of freedom of the model. Another method consists in smoothing the reported curve with splines, computing the equivalent degrees of freedom, and then imputing this value for the choice of q. Both approaches give similar results on the artificial dataset and for the COVID-19 epidemic curves. The last degree of freedom that the model has is the parameter λ. Estimation for the smoothing parameter has been handled in the past through various approaches, including Bayesian inference (see Fahrmeir and Lang [2001] and Rue et al. [2009]), the maximization of the restricted maximum likelihood (REML) with Laplace's method [Wood, 2011], or using generalized cross validation (GCV) scores [Wood, 2008]. If the number of observations is not too large, one can use a grid search on the possible values of λ, and choose λ based on the performance achieved on some portion of the observed values. The choice for the validation part is not trivial, as the reported cases curve can have values that are particularly high for some period of time, and using this as a validation set can deprive us of data points that are important for estimation.

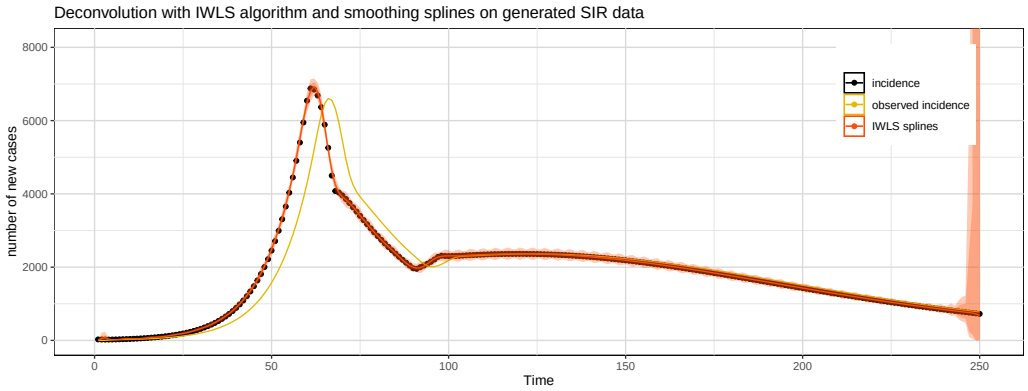
3.4.1 Results

Results of this method on the synthetic dataset are shown in Figure 5. As mentioned in the presentation of the method, the choice of the dimension of the spline basis q impact the effectiveness of the method. In particular, for $q = 100$ onward, the deconvolution procedure recovers closely the true incidence curve, and the confidence bands almost always include the true infection curve. However, one might observe some exploding estimates, with large confidence intervals (see Appendix A Figure 18). Indeed, choosing q too small can result in poor estimation of the incidence curve, with small confidence bands, and a choosing it too large produces good estimates for the central parts of the data, but might also produce an undesirable explosion at the end (which is the main point of focus, as we are mainly interested in estimating the reproductive number on a real time basis). One can solve this by again adding a pre-processing step where a prediction based on spline regression is done for the next 20 days, as shown in Figure 5c. This also avoids having very large confidence intervals at the end of the period of interest.

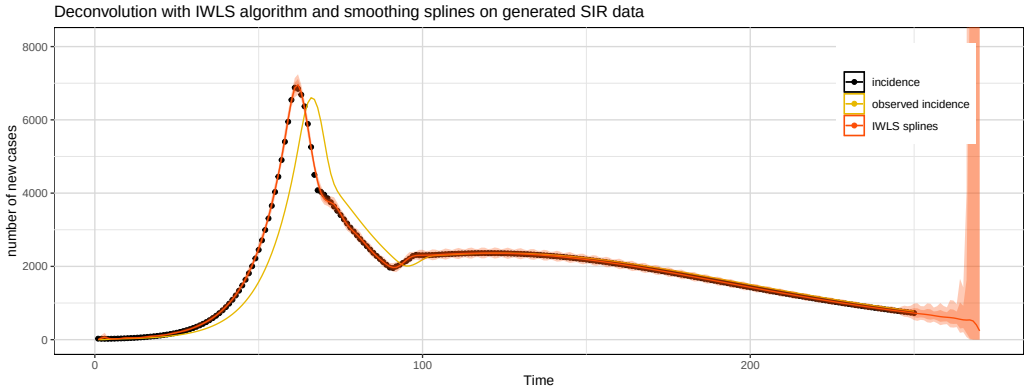
As shown in Figure 5, true incidence events are well estimated, and almost always fall inside the confidence bands. The transition periods are also well captured, in particular for day 90 and 97 (when the transition from $R_0 = 0.8$ to $R_0 = 1.15$ occurred).



(a)



(b)

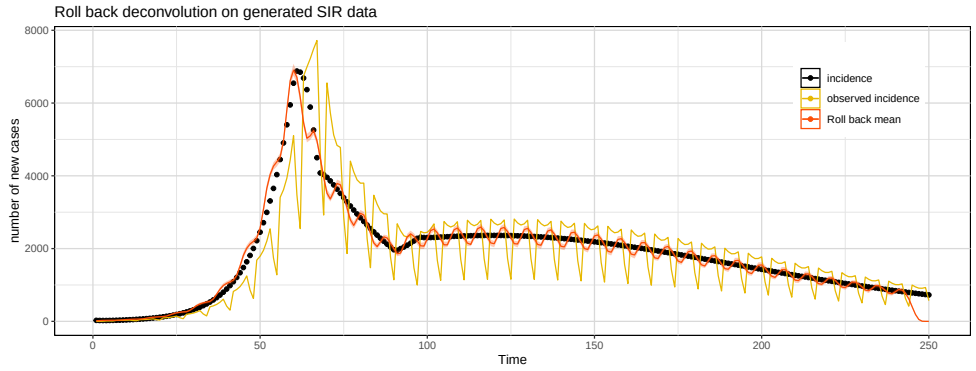


(c)

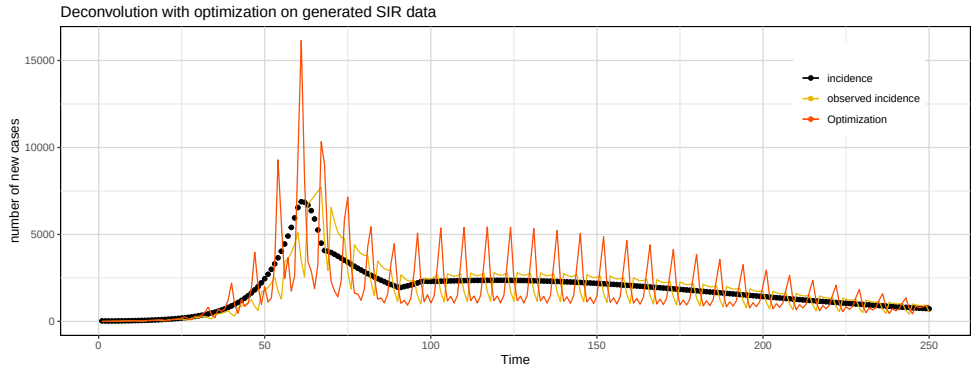
Figure 5: Results of the deconvolution with splines, with a basis dimension of size 70 with (a) identity link and (b) log link function. An extra prediction step is used in (c) with spline regression and log link function. A Poisson model is used, and the point-wise confidence intervals are shown in light red. The log link function avoids having negative values, but brings more instabilities during the first days, as the number of new cases is relatively low. Adding the extra prediction step ensures that the confidence bands are not exploding for the time interval of interest.

3.5 Deconvolution with noisy data

In this section, we evaluate the performance of the deconvolution methods presented in the previous three sections for data with strong weekly pattern. Details regarding how the incidence curve was generated can be found in Section 2. Deconvolving the observed incidence curve is crucial in order to get the true infection events that can be used to estimate the reproductive number R_t. Using this step, we avoid having estimates of R_t that would not be representative of the current epidemic situation. In the presence of highly variable data, as it is the case when dealing with weekly patterns, estimating R_t directly from observed events would bring as much variability in the estimates, making blurry and hardly interpretable the evolution of the epidemic. In the synthetic dataset, the true incidence curve is smooth compared to the observed one, and the deconvolution process should manage to recover it. From Figure 6, one can see that the first two deconvolution approaches presented in Section 3.2 and Section 3.3 hardly estimate the true incidence curve I_t. Using roll back, the weekly pattern is still present, but to a lesser extent compared to the observed curve. As for the method based on the optimization of the likelihood with gradient descent, the noise is even more amplified, as the method is overfitting the observed data. Smoothing the observations with a moving average before using the deconvolution method does not give better results, as shown in Figure 19 Appendix A. Even though the noise is no longer present, one must consider a different approach to deal of noisy data.



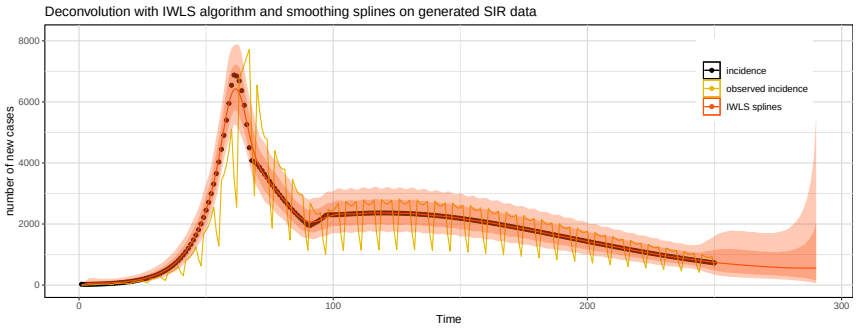
(a)



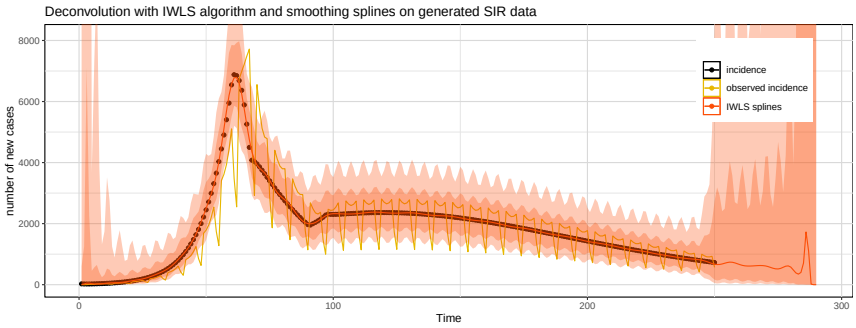
(b)

Figure 6: Results of the deconvolution methods for incidence curves with a strong weekly pattern: (a) Roll back deconvolution, (b) deconvolution with optimization of a Poisson response (Section 3.3). The observed incidence curve is represented by the solid yellow curve, and the true incidence by black dots. For the two deconvolutions, no extension was used.

The third deconvolution method, which is based on modeling the incidence curve with spline functions, has the advantage of benefiting from the mathematically robust definition of the true incidence curve as described in 3.3, with the possibility of controlling the smoothness of the resulting estimated incidence curve. This control can be achieved through the basis dimension q and the penalization term λ. Large values for q will result in an overfit of the data, as in Figure 6b, while too small values can result in a poor fit. The basis dimension q can be chosen from a range of potential candidates $q_1, \ldots, q_k$ by estimating the true incidence events for each q_j (without a penalization term) and compute the mean square error between the convolved estimated true events and a smoothed version of the observed infection curve. Then, the penalization term λ is chosen from a range of potential candidates similarly, based on the basis dimension which yielded the smallest mean square error. Results are shown in Figure 7. Compared to the data with no weekly pattern, the confidence bands are much larger, but does nevertheless include the true incidence curve I_t. Without a penalization term, estimated incidence events are very close to the true ones, with transition periods that are also well captured. However, the confidence bands are larger, especially for the first days of the epidemic, where these bands cover the range [0, 8000] for some days, which is highly unlikely and is partly linked to the behavior of spline functions near the extremities of the range of fitted values. Adding a penalization greatly reduces this effect, and avoids having wiggly behavior (for example between day 200 and 250 in Figure 7b). The confidence bands are then smaller, and the fitted values are smoother.



(a)



(b)

Figure 7: Results of the deconvolution with smoothing splines for noisy data. The basis dimension is 59 (so roughly one parameter every 4.2 days), and was selected automatically. (a) A penalization term is added, and λ was selected automatically, following the procedure described in Section 3.5. (b) No penalization term is used. The point-wise and simultaneous confidence bands are shown respectively in dark and light red.

4 Estimating the reproductive number R_t

Here, we provide two methods for estimating the reproductive number, one based on the incidence curve, the other based on a SIR approach. As mentioned in Section 1, the effective reproduction number R_t can give important insight as to how a disease outbreak is evolving. It is defined as the average number of secondary cases infected by an infected individual. When R_t is above the threshold of 1, the epidemic is growing at an exponential speed, while when R_t is below 1, the disease is exponentially declining. Mathematically, and using the notation from Fraser [2007], if $I(t)$ and $\beta(t, \tau)$ are respectively the incidence rate curve and the transmissibility function representing the change of I at time t with τ days since the infection, the following renewal equation is satisfied

$$I(t) = \int_0^\infty \beta(t, \tau) I(t - \tau) d\tau.$$

Moreover, the reproductive number R_t and case reproductive number R_t^c are given by

$$R_t = \int_0^\infty \beta(t, \tau) d\tau, \quad R_t^c = \int_0^\infty \beta(t + \tau, \tau) d\tau.$$

Fraser [2007] provides an illustration to distinguish R_t and R_t^c. For the case where $\beta(t, \tau)$ can be separated as a product of two functions that depend respectively on t and τ, the infectiousness simplifies to $\beta(t, \tau) = R_t w_\tau$, where w_t can be described as the generation interval distribution. The instantaneous reproduction number can then be written as

$$R_t = \frac{I(t)}{\int_0^\infty I(t - \tau) w_\tau d\tau},$$

which leads to the following estimator, considering that the reported incidence curve is discretized on a daily basis:

$$\hat{R}_{t_i} = \frac{I(t_i)}{\sum_{j=0}^n I(t_i - t_j) w_j},$$

These results are used by Cori et al. [2013] to estimate the instantaneous reproductive number R_t. Perasso [2018] provides a deeper insight of the reproduction number in mathematical epidemiology, and links R_t with compartmental SIR models that we will use later in this report.

4.1 Estimating R_t with smoothing splines

Analogously to what was done in Section 3.4, we here present a method to estimate the (instantaneous) reproductive number with spline functions, so that the resulting quantity is a smoothed function of time and can handle observations that could be very noisy. The model assumes that the mean number of newly infected individuals at time t is a weighted average of the new infected individuals at previous time step, multiplied by some quantity R_t that varies across time. More formally, let I_t be the incidence curve for time step $t = 1, \dots, T$, and let $\{\tilde{w}_j\}_{j=1}^K$ be the infectivity profile, which is different from the incubation period that we introduced in Section 3. We suppose that each I_t is Poisson distributed, with mean

$$\mu_t = R_t \sum_{j=1}^K \tilde{w}_j I_{t-j} = \exp\left[\log(R_t) + \log\left(\sum_{j=1}^K \tilde{w}_j I_{t-j}\right)\right], \tag{4.1}$$

where the generation interval is assumed to be known. As such, the second term in the exponential (the offset) can be explicitly computed, and the only unknown is the log of the reproductive

number. In order to estimate this, we suppose that $\log(R_t)$ is smoothed over the time, that is,

$$\log(R_t) = \sum_{i=1}^{q_1} \beta_j B_j(t),$$

where $B = \{B_1, \ldots, B_{q_1}\}$ is a spline basis of dimension q_1, defined over the interval $[0, T]$. The parameters $\{\beta\}_{j=1}^{q_1}$ are estimated through the minimization of the penalized sum of squares

$$\sum_{t=1}^{T} (y_t - \mu(x_t))^2 + \lambda \int_{\chi} \mu''(x)^2 dx$$

where $\chi = [0, T]$ and λ controls the weight of the roughness penalty [Davison, spring 2020], and is estimated by restricted maximum likelihood (REML).

Figure 8 shows the estimate of R_t based on this method for the synthetic dataset, along with its estimated first derivative, which can be obtained using finite difference methods. The spline dimension is chosen to be 100, which corresponds to one parameter every three days. The generation intervals $\{\tilde{w}_j\}_{j=1}^{K}$ are set according to the parameters used to generate the fake dataset (see Section 2). The equivalent degrees of freedom, as defined in Davison [spring 2020] are about 70, and so one can reduce the dimension for the basis accordingly.

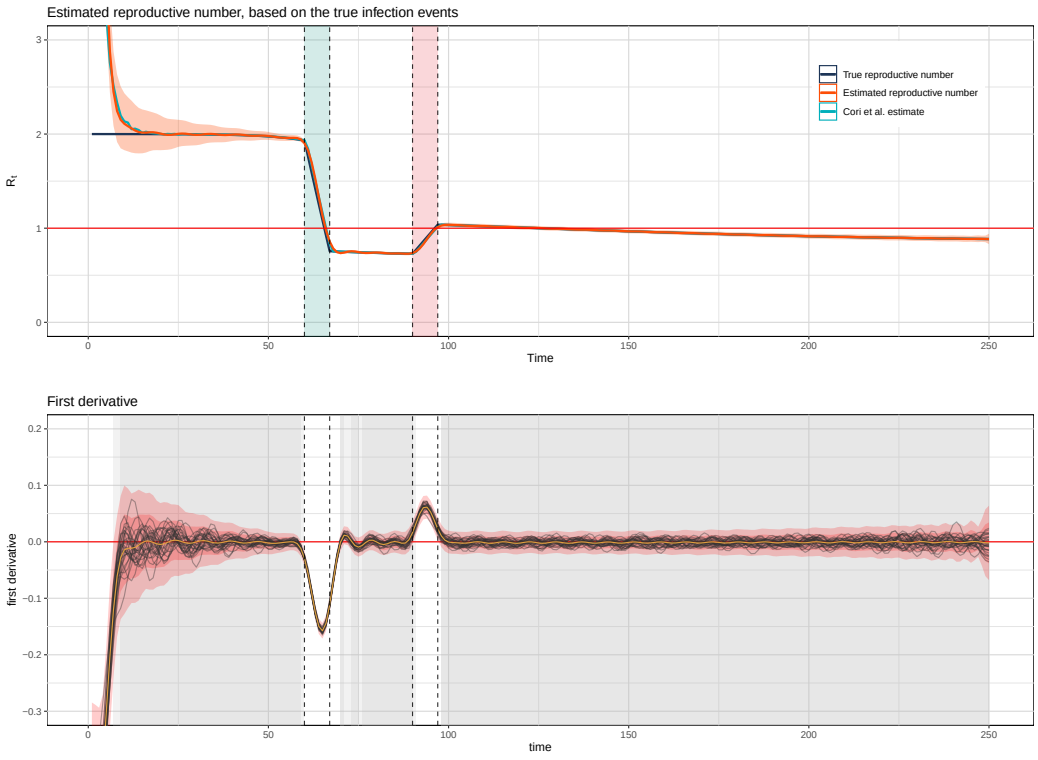


Figure 8: (top) Estimated R_t based on the true infection events I_t and the spline method. The two transition phases are shown in light blue and red, and the point-wise 95% confidence bands are shown in red. (bottom) Estimated first derivative of R_t, with in light red the point-wise (inner) and simultaneous (outer) confidence bands. The black curves are sampled curves from the parameters estimates, and the solid yellow curve is the estimated first derivative of R_t. The regions where the confidence bands include the threshold 0 are shown in grey. The estimates for the first derivative are obtained using a finite difference method.

The method manages to recover the true dynamic of the simulated epidemic, with the decrease of the reproductive number at 60 steps from 2 to 0.8 and smaller confidence bands. Until day 60, the confidence bands are much larger than for the rest of the time interval, and the first estimates of R_t are well above their true values. This can be caused by a lack of data, as there are few new infected people per day at the start of the epidemic. This effect does not seem to be linked to the method itself, as the Cori et al. estimates show similar behaviour.

The plot showing the estimated first derivative of R_t, denoted $\hat{R}'_t$, can be useful to detect the underlying changes in R_t. The confidence bands include the value 0 the time intervals [0, 60], [70, 90] and [100, 250], and can be used to detect the two transition periods. R_t exhibit a slight downtrend between day 100 and 250, but we can not exclude a neutral trend, though $\hat{R}'_t$ is slightly negative during this period.

For real data, one does not observe the true incidence curve, and an intermediate deconvolution step is needed to recover it. As discussed in Gostic et al. [2020], shifting the reproductive number curve does not add an extra convolution step to the estimates (or the observed data), but does neither properly estimate true incidence events, which can result in inaccurate estimates. One major advantage of the smoothing spline estimation of R_t is that the uncertainties in the observations can be relayed to the estimation process, by setting the inverse of the variance for each $\hat{I}_t$ as weight for the P-IWLS algorithm. Hence, observations with large variance will have their impact reduced, and those with smaller confidence intervals are judged as reliable. Figure 9 shows the effect of using a deconvolution step before the estimation of R_t, based on the observed curve without weekend effects. The confidence bands are wider, reflecting the uncertainties in the estimates for the deconvolution. As the deconvolution method is using spline functions, the resulting estimates of R_t can suffer from problems that have smoothing splines to model the data near the border of the domain definition (of the observed values), which is transcribed by wiggly behaviour during the initial days, as exposed in Figure 9. This is amplified for its derivative estimates $\hat{R}'_t$.

In the presence of weekly pattern, using a penalization term for the deconvolution can greatly impact $\hat{R}_t$. Results are shown in Figure 10. Without a penalty, we know from Section 3.5 that true incidence events can be well estimated, but with large confidence intervals and last values that depends highly on the prediction step used to extend the original observed number of infected individuals. As such, transitions periods can be adequately captured given sufficiently observations around (typically for day 60 and 90), but can lead to estimates that deviate from the true values for recent observations. On the other side, using a penalty gives estimates of R_t that are smoother and more stabilized for recent times. Hence, the use of penalization is more suitable on a daily basis, while ignoring it helps detect the transition periods of the epidemic in the past and assess the effectiveness of sanitary and containment measures.

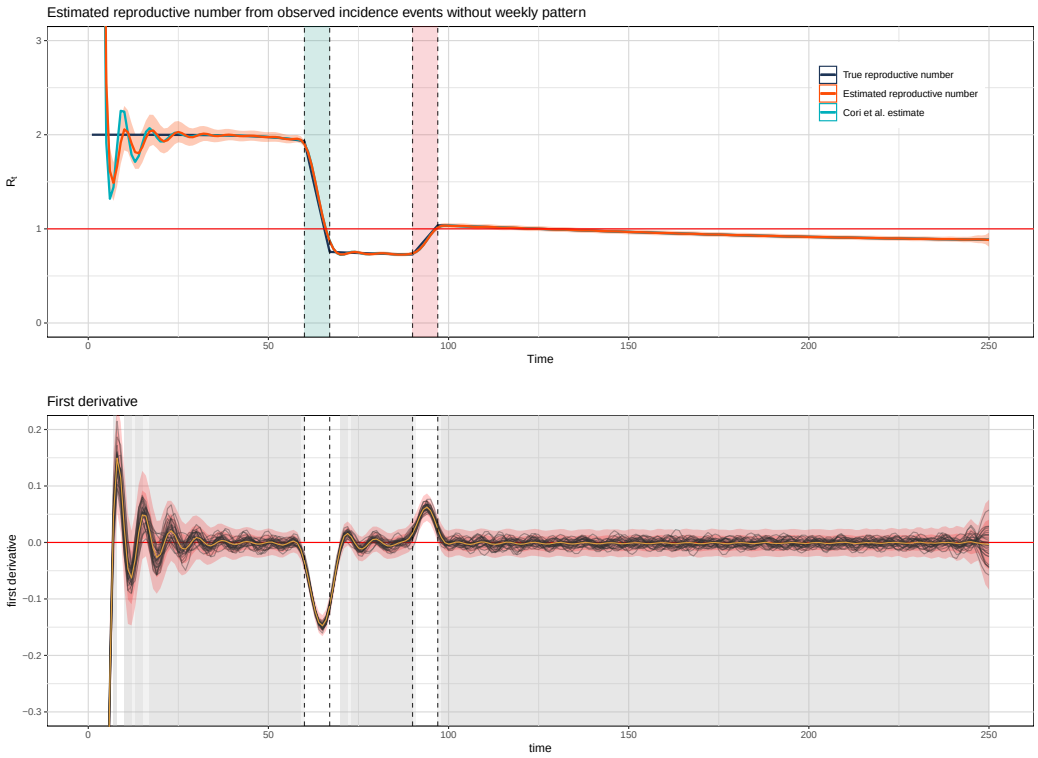


Figure 9: (top) Estimated R_t based on the estimates of true infection events I_t and the spline method. The two transition phases are shown in light blue and red, and the point-wise 95% confidence bands are shown in red. (bottom) Estimated first derivative of R_t, with in light red the point-wise (inner) and simultaneous (outer) confidence bands. The black curves are sampled curves from the parameters estimates, and the solid yellow curve is the estimated first derivative of R_t. The regions where the confidence bands include the threshold 0 are shown in grey.

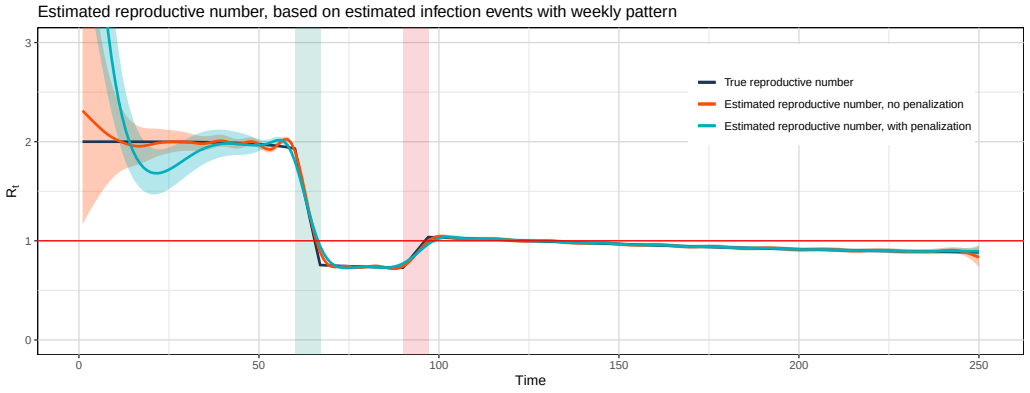
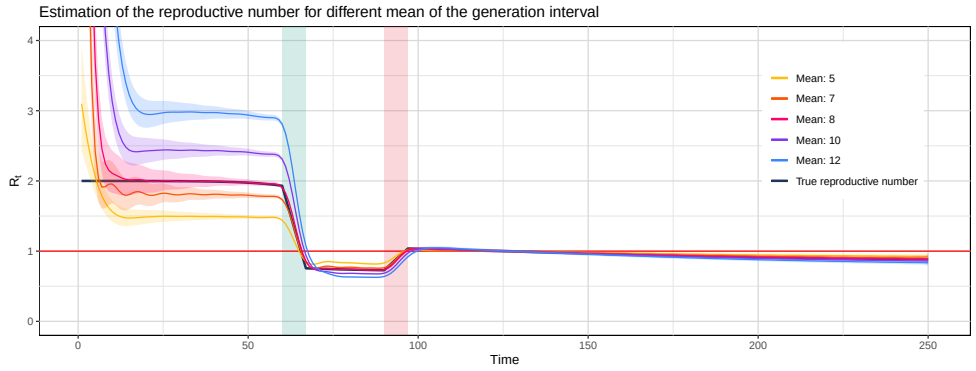


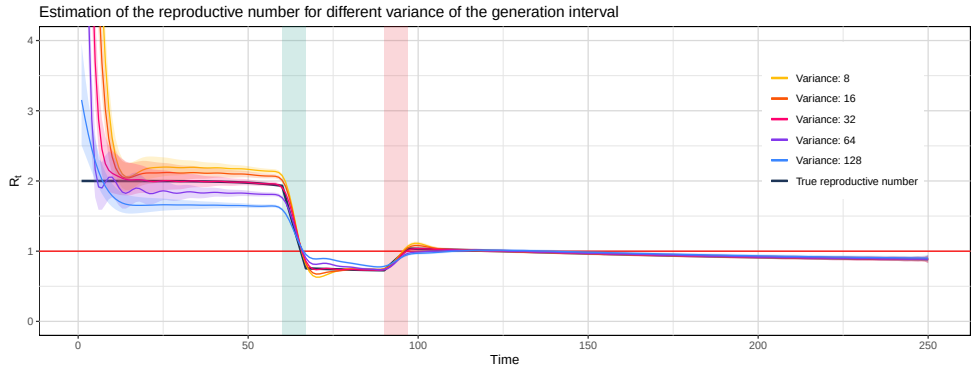
Figure 10: Estimated R_t from the curve of new cases with weekend effects. In red, no penalization is used, and in blue, a penalization is used (and chosen as described in Section 3.5. As the confidence intervals for the first days are particularly large when no penalty is used, the confidence intervals for $\hat{R}_t$ are wide here as well.

4.1.1 Sensitivity analysis and misspecification of the generation interval

Gostic et al. [2020] provide extensive discussion of the effects of misspecification of the generation interval $\{\tilde{w}_j\}_{j=1}^K$ for estimating R_t with the Cori et al. method. A similar analysis is carried here with our approach, by varying the mean and the variance for the generation interval. One can also check how a misspecification of the incubation period can affect the quality of $\hat{R}_t$. Figure 11 illustrates the effects of the misspecification of the mean and the variance. A higher mean results in higher estimates for the first phase of the epidemic (until $t = 60$), and vice versa for a lower mean. The transition phase is also slightly delayed in case of a higher mean. A change in the variance has similar effects, with exaggerated estimates when the variance is smaller, especially at the end of the two transition periods, as shown in Figure 11. As noted in Gostic et al. [2020], biases caused by a misspecification of the generation interval are small when R_t is close to the critical threshold 1, but can be substantially larger during the early phase of the epidemic, when the estimates of R_t are larger as a result of limited data. The generation interval is in practice harder to estimate, and is often replaced by the serial interval, which share the same mean but might have different variance, and take negative values. Thus, extra care must be taken in order not to have large bias in estimates of the reproduction number. The uncertainties for the generation interval can be taken into account by sampling different mean and variance values, and estimate R_t accordingly.



(a)



(b)

Figure 11: Estimated reproductive number for different (a) mean and (b) variance of the generation intervals. In (a), the variance was set to 32, and the mean is set to the values 4, 6, 7, 9, and 11. In (b), the mean is set to 8, and to the values 8, 16, 32, 64 and 128 for the variance.

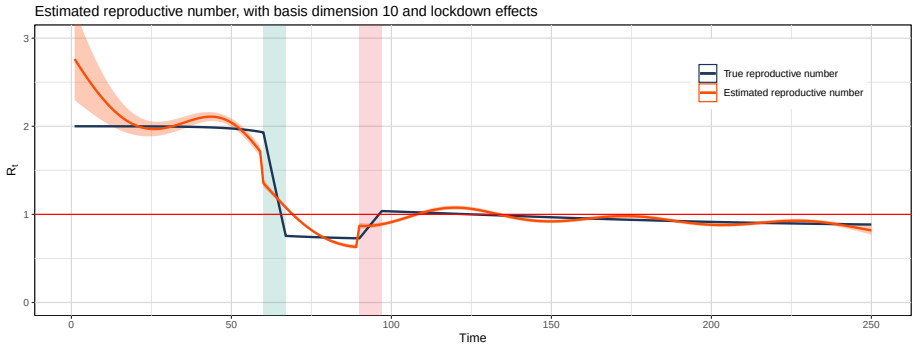
4.1.2 Lockdown effects

Estimating R_t is of huge interest as regards to tracking the effectiveness of measures to contain the spread of an epidemic. Pan et al. [2020] explored the impact of various interventions on the reproductive number for the COVID-19 epidemic, including traffic restriction, social distancing and home quarantine, and found evidence that a better control of the COVID-19 outbreak in Wuhan was associated with multiple public health measures.

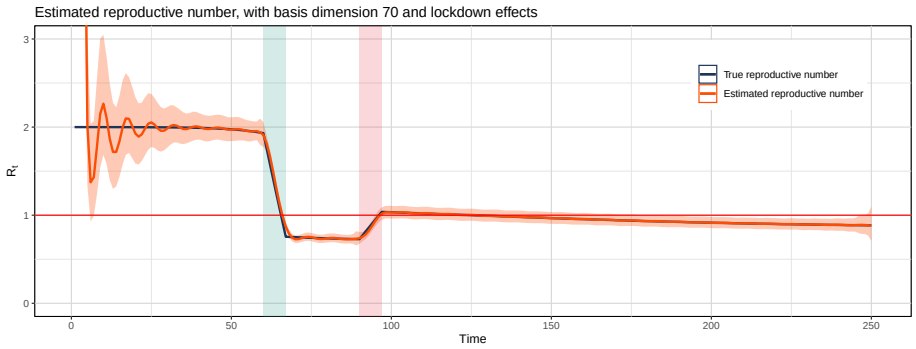
In our model, the effect of a measure can be assessed with the addition of a lockdown effect in the linear term of the generalized additive model. In particular, supposing that some measure comes into force at a time $b \in [0, T]$, then one can modify (4.1) to include this term as follows:

$$\mu_t = \exp\left[\log(R_t) + \log\left(\sum_{j=1}^{K} \tilde{w}_j I_{t-j}\right) + \alpha \mathbb{1}_{\{t \geq b\}}\right], \tag{4.2}$$

where the parameter α is estimated. One can also include a factor variable for each containment measure. Figure 12 shows $\hat{R}_t$ for the cases where the basis dimension is 10 or 150. Here, we suppose that the lockdown takes place between $t = 60$ and $t = 80$. As shown, when the basis dimension is sufficiently low, the lockdown effect is significant and visually detectable, but the estimates of R_t are not credible.



(a)



(b)

Figure 12: Estimated R_t based on true incidence events and with lockdown effects. The basis dimension are (a) 10 and (b) 70. In both cases, a quasi-Poisson model is used, with log link function. As the basis dimension increases, the effect of the lockdown variable is less significant (in (b), the p-values are below 0.05).

4.2 Estimation of R_t based on a SIR model

Here we present an approach for estimating R_t based on the susceptible-infectious-recovered (SIR) model. If we denote by $S(t)$ the number of susceptible individuals at time t, by $I(t)$ and by $R(t)$ the numbers of infected and recovered individuals at time t, then the dynamics of the SIR model can be described by the differential equations

$$\begin{cases} \dfrac{\partial S(t)}{\partial t} = -\beta(t)s(t)I(t), \\ \dfrac{\partial I(t)}{\partial t} = \beta(t)s(t)I(t) - \gamma(t)I(t), \\ \dfrac{\partial R(t)}{\partial t} = \gamma(t)I(t), \end{cases}$$

with $S(t) + I(t) + R(t) = N$ being the population size. The second equation can be simplified, giving

$$\frac{\partial I(t)}{\partial t} = \beta(t)I(t)\frac{[N - I(t) - R(t)]}{N} - \gamma(t)I(t). \tag{4.3}$$

As the number of cases and recovered cases are usually provided at a daily basis, one can discretize these equations, giving:

$$\begin{cases} I(t+1) = I(t) + \beta(t)I(t)[N - I(t) - R(t)]/N - \gamma(t)I(t) \\ R(t+1) = R(t) + \gamma(t)I(t) \end{cases} \tag{4.4}$$

The only unknown quantities here are $\beta(t)$ and $\gamma(t)$, for which estimates are needed to get the reproductive number $R_t = \beta(t)/\gamma(t)$ [see Perasso, 2018].

Hong and Li [2020] proposed an approach based on a Poisson model that estimates the time-varying transmission and removal rates $\beta(t)$ and $\gamma(t)$ with spline functions. Unlike methods that assume that the parameters are time-invariant, their estimates can provide the dynamics of the transmission of the epidemic, and can be used for assessing the effectiveness of various containment measures to stem the disease. More formally, if we let $B(t) = \{B_1(t), \ldots, B_{q_1}(t)\}$ and $\tilde{B}(t) = \{\tilde{B}_1(t), \ldots, \tilde{B}_{q_2}(t)\}$ be the cubic B-spline functions over the time interval $[0, T]$, with knots $0 = w_0 < \cdots < w_{q_1 - 1} = T$ and $0 = w_0 < \cdots < w_{q_2 - 1} = T$, then the transmission and removal rates are of the following form:

$$\log \beta(t) = \sum_{j=1}^{q_1} \beta_j B_j(t) \quad \text{and} \quad \log \gamma(t) = \sum_{j=1}^{q_2} \gamma_j \tilde{B}_j(t). \tag{4.5}$$

Let the observed number of infected and recovered cases $\tilde{I}(t)$ and $\tilde{R}(t)$ follow Poisson distributions with means $I(t)$ and $R(t)$ as defined by the discretized differential equations. Then one can maximize the log-likelihood $l(\beta, \gamma)$ with respect to the parameters $\beta_1, \ldots, \beta_{q_1}, \gamma_1, \ldots, \gamma_{q_2}$, and obtain an estimate for the reproductive number as

$$\hat{R}_t = \frac{\hat{\beta}(t)}{\hat{\gamma}(t)} = \exp\left(\sum_{j=1}^{q_1} \hat{\beta}_j B_j(t) - \sum_{j=1}^{q_2} \hat{\gamma}_j \tilde{B}_j(t) \right). \tag{4.6}$$

The optimization can be carried with any gradient-based algorithm, for example the conjugate gradient or BFGS algorithm. Also, at each iteration k, equation (4.4) is used to update the infected and recovered cases, based on the values of the parameters $\beta_1^k, \ldots, \beta_{q_1}^k, \gamma_1^k, \ldots, \gamma_{q_2}^k$. As such, the link between the successive I_t and R_t is preserved, and is used to generate an epidemic curve in its entirety. One key advantage of this method is that the estimate for the reproductive number is guaranteed to be positive. The confidence intervals for the parameters are obtained from the information matrix, and the confidence intervals for the reproductive number is obtained with the delta method.

4.2.1 Generalized additive model for SIR

Another approach, which is computationally faster, converts the previous model into generalized additive form. Velthuis et al. [2002] adapted the traditional SIR model with time-independent infection rate $\beta(t)$ with a GLM in order to estimate the transmission of *Actinobacillus pleuropneumoniae* in pigs, and to study whether it could be explained by the variation in infectivity of the pathogen. As explained latter [Velthuis et al., 2007], estimating the transmission parameters based on generalized linear models has the key advantage of being able to account for heterogeneity in the studied populations, by including factor variables for subgroups in the three compartments of the SIR model. However, and as mentioned previously, subsequent observations are assumed to be independent, which can be questionable in practice. Here, in order to allow for the inclusion of a linear term in the model, we write the transmission and removal rates as

$$\beta(t) = \sum_{j=1}^{q_1} \beta_j B_j(t), \quad \gamma(t) = \sum_{j=1}^{q_2} \gamma_j \tilde{B}_j(t) \ . \tag{4.7}$$

Using this definition, and by putting $\beta(t)$ and $\gamma(t)$ back in (4.4), we get the following infectious and recovery updates

$$\begin{cases} I(t+1) = I(t) + \sum_{j=1}^{q_1} \beta_j \left[B_j(t)I(t) - B_j(t)I(t)^2/N - B_j(t)I(t)R(t)/N \right] - \sum_{j=1}^{q_2} \gamma_j \tilde{B}_j(t)I(t), \\ R(t+1) = R(t) + \sum_{j=1}^{q_2} \gamma_j \tilde{B}_j(t)I(t). \end{cases} \tag{4.8}$$

These can be expressed in a generalized additive setting, with the vector of response variables Y that follow a Poisson distribution with mean $Q\theta$, with $Y = [I(1), \ldots, I(T), R(1), \ldots, R(T)]^T$, $\theta = [\beta_1, \ldots, \beta_{q_1}, \gamma_1, \ldots, \gamma_{q_2}]^T$ the parameters to be estimated and Q the $(2 \times T) \times (q_1 + q_2)$ matrix

$$Q = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix},$$

with

$$Q_{11} = \begin{bmatrix} \left[I(0) - I(0)^2/N - I(0)R(0)/N\right] B_{[1:q_1]}(0) \\ \vdots \\ \left[I(T-1) - I(T-1)^2/N - I(T-1)R(T-1)/N\right] B_{[1:q_1]}(T-1) \end{bmatrix} \in \mathbb{R}^{T \times q_1},$$

$$Q_{12} = \begin{bmatrix} I(0)\tilde{B}_{[1:q_2]}(0) \\ \vdots \\ I(T-1)\tilde{B}_{[1:q_2]}(T-1) \end{bmatrix} \in \mathbb{R}^{T \times q_2},$$

$$Q_{21} = 0 \in \mathbb{R}^{T \times q_1}, \quad Q_{12} = -Q_{21} \in \mathbb{R}^{T \times q_2}$$

Setting the offset $X_{\text{offs}} = [I(0), \ldots, I(T-1), R(0), \ldots, R(T-1)]^T$, a generalized additive model with Poisson response and identity link function can be used, assuming that the observations are independent, which might be false in practice. One can use the IWLS algorithm to get the estimates. Although the estimated parameters successfully recover the true reproduction number, large instabilities occurred, with large confidence bands and potentially negative estimates. This

is largely due to the transmission and removal rates not being constrained to remain positive, which is the case when they are assumed to be of the same form as in equation (4.5). The `glmc` package in `R` [Chaudhuri et al., 2006] provides methods to estimate the parameters of various families of generalized linear models with constraints. However, spline estimates are still very unstable (see Appendix B Figure 20).

To deal with this issue, in the following we propose an alternative approach, that estimates $\beta(t)$ and $\gamma(t)$ successively. We assume that these two quantities are as given in equation (4.5). The discretized equation for the curve of recovered cases $R(t)$ does not depend on the transmission rate, so

$$R(t+1) = R(t) + \exp\left[\sum_{j=1}^{q_2} \gamma_j \tilde{B}_j(t) + \log I(t)\right].$$

As the number of recovered cases is an increasing function of time, we are guaranteed that $R(t+1) - R(t) \geq 0$ for each t, and we can assume that $R(t+1) - R(t)$ follows a Poisson distribution with mean $\mu_R(t) = \exp\{\sum_{j=1}^{q_2} \gamma_j \tilde{B}_j(t) + \log I(t)\}$. This is of GAM form, with Poisson response and log link function, and the parameters can be estimated with the IWLS algorithm. Then, given an estimate $\hat{\gamma}(t) = \sum_{j=1}^{q_2} \hat{\gamma}_j \tilde{B}_j(t)$, we can write, based on the first equation (4.4), the following model for the infected curve

$$I(t+1) - I(t) + I(t) \exp\left\{\sum_{j=1}^{q_2} \hat{\gamma}_j \tilde{B}_j(t)\right\} \sim \text{Pois}\{\mu_I(t)\} \tag{4.9}$$

with $\mu_I(t) = \exp\left[\sum_{j=1}^{q_1} \beta_j B_j(t) + \log\{I(t) - I(t)^2/N - I(t)R(t)/N\}\right]$, and the term inside the logarithm being an offset. To allow for overdispersion, one can turn to a quasi-Poisson or negative binomial model. As mentioned in Section 4.1, the uncertainty brought by the estimation of γ_t can be incorporated into the estimation of β_t by re-adjusting the weights of the observations according to their inverse variances. The result of this method based on the simulated epidemic curve is shown in Figure 13, for splines with $q_1 = 50$ and $q_2 = 50$. The estimates reconstruct R_t with fidelity, though the confidence intervals are large for the first phase of the epidemic.

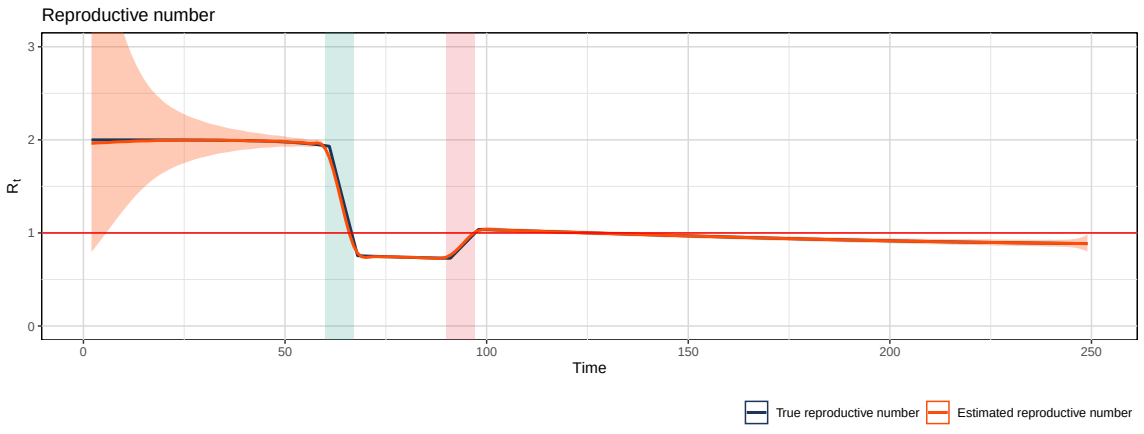


Figure 13: Estimated reproductive number, based on the true curves of infected and recovered individuals. The 95% confidence bands are shown in light red. The estimates for R_t are highly variable during the first days, with confidence bands that include the threshold 1. $\hat{R}_t$ is an adequate estimates of the true reproductive number.

4.2.2 Sensitivity analysis

As for the approach based on smoothing splines, this method requires the true incidence curve for estimating R_t, as well as the recovered cases curve. Since their observed versions do not necessarily reflect the true dynamic of the epidemic, one must use a deconvolution step. However, and this differs from the first approach, the SIR model also requires the true curve of recovered cases, which is different from the observed one. The techniques introduced in Section 3 can be used, but require the specification of a delay interval (which corresponds to the incubation period for the incidence curve). As the delay for the recovery is different from the delay of symptom onset, one must estimate both in order to reach an estimate for R_t. However, this method does not require an estimate of the generation interval. Here, we assess as in Section 4.1.1 the effects of misspecification of the mean and the variance for the delay interval, for both the incubation and recovery periods. The results are very similar to those obtained in the previous sensitivity analysis. When the mean is underestimated for the recovered curve, the resulting estimate is lower, and vice-versa (see Figure 14). Overestimating the mean of the recovered curve can results in wiggly effects, which can occur when the number of recovered cases is greater than the total number of infected cases. This can cause the splines to have wavy behaviour, is shown in Figure 14 with the dark blue curve, with a mean recovery delay taken as 15 days instead of 5.

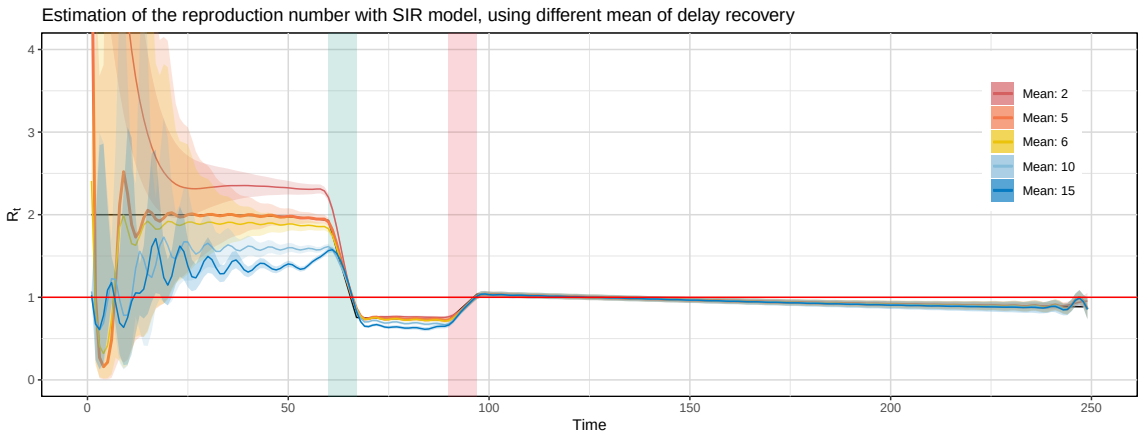


Figure 14: $\hat{R}_t$ for different mean recovery delay. The incubation period is fixed, with mean 5 days. No penalty is used here in order to see the behaviour of each estimated reproductive number during the two transition periods. This results in biases for the last days, as mentioned in Section 3.5. The generation interval is known.

5 The COVID-19 epidemic in Switzerland

5.1 Introduction

First identified in December 2019 in Wuhan (China), the COVID-19 disease was declared as a global pandemic by the World Health Organization on March 11, 2020 [WHO, 2020], as nearly 118'000 cases were already reported across 110 countries. On June 25, 2020, the threshold of 10 million infected people was crossed, with more than 504'000 deaths, and as of December 24, 2020, nearly 80 million infected and 1.7 million deaths were reported. The virus is spread between people during close contact, and fever, cough, shortness of breath and fatigue are the most common symptoms. Pneumonia and acute respiratory distress syndrome are the two main complications of the disease, which can lead to the death of the affected individual. Worried that local health care systems would be overwhelmed, countries shut their borders and implemented strict sanitary measures in order to contain the pandemic. The COVID-19 disease has been monitored by state authorities to understand how the virus is spread and how public health measures such as traffic restriction, social distancing and home quarantine can control the outbreak, in particular via the daily information provided by the World Health Organization and Johns Hopkins University's Coronavirus Resource Center [JHU, 2020].

As of December 2020, many potential vaccines are under study and in the testing phase, and their effectiveness will depend on several factors, such as their speed of development, manufacture and delivery. Meanwhile, the use of daily report and death curves to track the spread of the pandemic are still of great help, in particular for estimating the reproductive number of the disease. In the first part of this report, we reviewed and proposed several methods to estimate it, taking into account potential delays for the report of infected individuals. In the following, we will put these tools into practice to estimate the reproductive number of the COVID-19 pandemic in Switzerland.

5.2 Data aggregation and collection

In order to apply the proposed methods, we used data from various data sources that provided the daily number of infected, deceased and recovered individuals for Switzerland and other countries. For Switzerland, the data were obtained through the Federal Office of Public Health [FOPH, 2020], which also provides various information on the current situation of the COVID-19 pandemic, including estimates of the reproductive number. Reports are updated each day, including during weekends, so that the resulting infected and recovered curves that are used do not have too large weekend effects that can be directly linked to the reporting procedure. The two other sources that we used for estimating the reproductive number for other countries are from the Humanitarian Data Exchange website [HDX, 2020] and a GitHub data repository website (`https://github.com/owid/covid-19-data/tree/master/public/data`), which aggregates daily counts from various sources, including from the John Hopkins University and the European Centre for Disease Prevention and Control [ECDC, 2020].

Figure 21 Appendix C shows the number of newly infected, deceased and recovered individuals in Switzerland, from February 25, 2020 to December 24, 2020. A 20 days ahead prediction is added, with the 95% confidence intervals shown in light red. One important feature of the data is the strong weekly pattern, with large drops of the daily counts during the weekends, and above average counts on the Mondays. The maximum new number of infected individuals was multiplied by 8 for the second epidemic wave, reaching more than 10'000 new infected cases on November 1, 2020. Since, these figures have decreased, and now stagnate, raising concerns about a third wave.

As mentioned in Section 3, predicting the next values avoids having estimates for the true

incidence curves with large drops at the end. Although the deconvolution method based on splines approximations can handle this issue, results are better by predicting the next values, and then account for the uncertainty of the prediction directly in the deconvolution process. For the extension, we used thin plate regression splines to extrapolate the future reported cases. More formally, we suppose that $I_t \sim \text{Pois}(\mu_t)$ with

$$\mu_t = \exp\left(\sum_{i=1}^{q} \alpha_i \phi_i(t) + \sum_{j=1}^{q_d} \beta_j B_j(d)\right), \quad t = 1, \ldots, T + 20, \tag{5.1}$$

where $\phi_i(t)$ is the thin plate spline basis, that models the incidence curve as a smooth function of the time. To account for the weekly pattern, we added a second smoothing component, for the days of the week, with basis $B_j(d)$, where d denotes the day of the week. As the data exhibit strong signs of overdispersion, we used a negative binomial model instead of the Poisson. The quantile-quantile and residual plots can be found in Appendix C. One issue that arises when using generalized additive models to represent time series is that the spline basis are constructed based on the data that are used to fit the model, which can cause the extrapolating estimates to have very wiggly behaviour. Following Simpson [2018], one can use penalty terms to control the behaviour of the spline functions outside the range of data points that were used to fit the model.

We used this extrapolation step for all three curves: the infected, death and recovered ones. The death curve does not show a strong weekly pattern, and the curve for new recovered cases has on the other hand high variability. Thus for the latter, a centered-moving average of 7 days was used before the prediction, and the last 10 values were assigned a weight of 0 in the fitting process, as the reported number of recovered individuals was 0 for few consecutive days. Doing so avoids badly behaving extrapolating splines, but as shown in Figure 21c, this comes at the cost of having larger 95% confidence intervals.

5.3 Deconvolution

We performed the deconvolution process as explained in Section 3.4. Following Lauer et al. [2020], we used an incubation period with mean 5.1 days and 95% confidence interval 4.1-5.8 days. The resulting estimated true incidence curve is shown in Figure 15, along with its 95% point-wise and simultaneous-wise confidence bands. To deal with overdispersion, a quasi-Poisson model was used, and for the P-IWLS algorithm, the predicted values that extend the observed incidence curve were assigned a weight corresponding to their inverse variance. The other observed values were assigned a weight of 1. Lastly, we extended the data with ones for the first 20 values, so that the deconvolution does not overestimate the number of true new cases at the beginning, due to the behaviour of spline functions near the extremities of the range of fitted values. The dimension of the spline basis was chosen from a range of potential values so as to minimize the mean square error between smooth observed incidence events and the convolution of true estimated incidence events. Using the smooth observed curve avoids selecting a basis dimension too high, that would overfit the observed incidence curve and thus produce weekly pattern that would emphasize the observed ones. Given a basis dimension q, the constant λ for the penalization term Δ (see Section 3.4 for more details) is chosen with the same method, from a range of potential values. In the end, we kept the value $q = 54$, which corresponds to one parameter every 6 days. A comparison of several deconvolved curves is shown in Figure 23 Appendix C. As expected, the confidence intervals are wider for the last values, and both an increase or a decrease in the number of new cases are plausible. This is due to the uncertainty brought by predicting the number of new cases for the following 20 days.

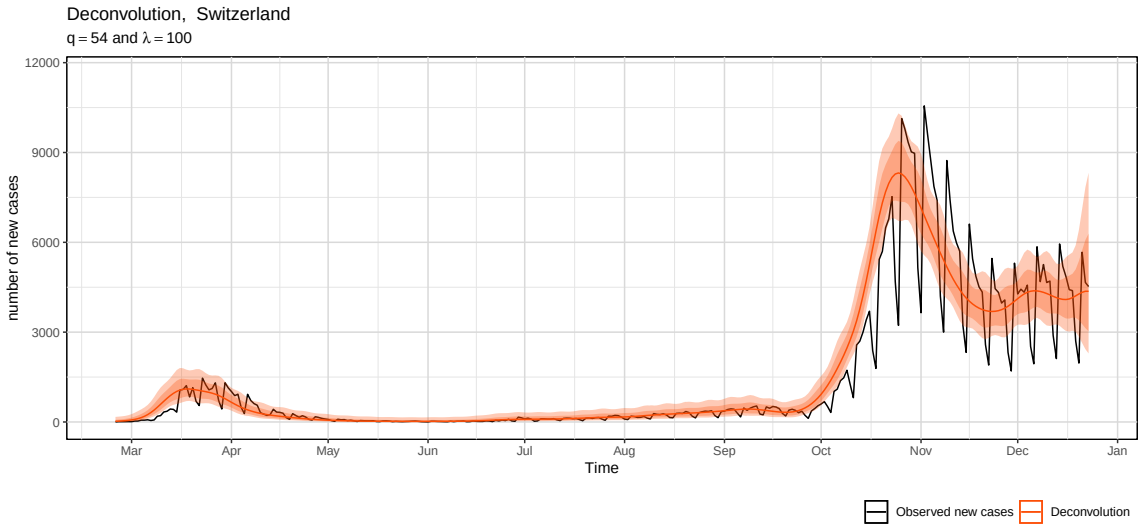


Figure 15: Deconvolution with smoothing splines and automatic degree selection, with quasi-Poisson model. The observed incidence curved in shown in black, and the estimated true one in red. The 95% point-wise confidence bands is in light red, and the 95% simultaneous one in lighter red.

5.4 Estimating R_t

We used the two methods described in Section 4 to produce estimates of R_t, denoted $\hat{R}_t$, based on estimated true incidence events that we obtained with the deconvolution. For the first approach, which directly estimates the reproduction number with smoothing splines, we used generation intervals with mean 5.2 days (95%CI 3.78-6.78) and standard deviation 1.72 (95%CI 0.91-3.93), following the estimates of Tapiwa et al. [2020] from clusters in Singapore and Tianjin. For these estimates, the incubation period distribution was assumed to be known. Estimated R_t and its first derivative are shown in Figure 16. We used the inverse of the variance of estimated true incidence events as weights for fitting the generalized additive model, and multiplied them by the square root of the fitted values in order not to penalize large values with large confidence intervals more than small estimated values with small confidence bands, as we are using a negative binomial model. Finally, we capped the weights for the first values (which were very large in comparison with the others) to avoid having bad behaving confidence bands. For the spline basis, we used a dimension of 50, which lead to an estimated degree of freedom of 37.1, giving roughly one parameter every 8 days.

The estimates of R_t show that the first sanitary measures had an impact on the reproductive number for COVID-19 in Switzerland, which broke down the critical threshold of 1 nearly one week after the implementation of the measures, on March 23, 2020 (with 95% CI March 21 - March 26). During the first weeks of March 2020, values for $\hat{R}_t$ are quite high, and may suffer from a lack of observations (see Section 4.1). $\hat{R}_t$ then stayed under 1 for more than two months, gradually increasing after the restrictive measures have ceased to be applied. It reached a top around mid-June to 1.43, before decreasing again and reaching a plateau between 1 and 1.4 for more than two months, even after masks became mandatory in public transport. September 18, 2020 marked the beginning of the second wave, with the confidence intervals for the first derivative excluding the value 0. The estimates for R_t during the first half of October

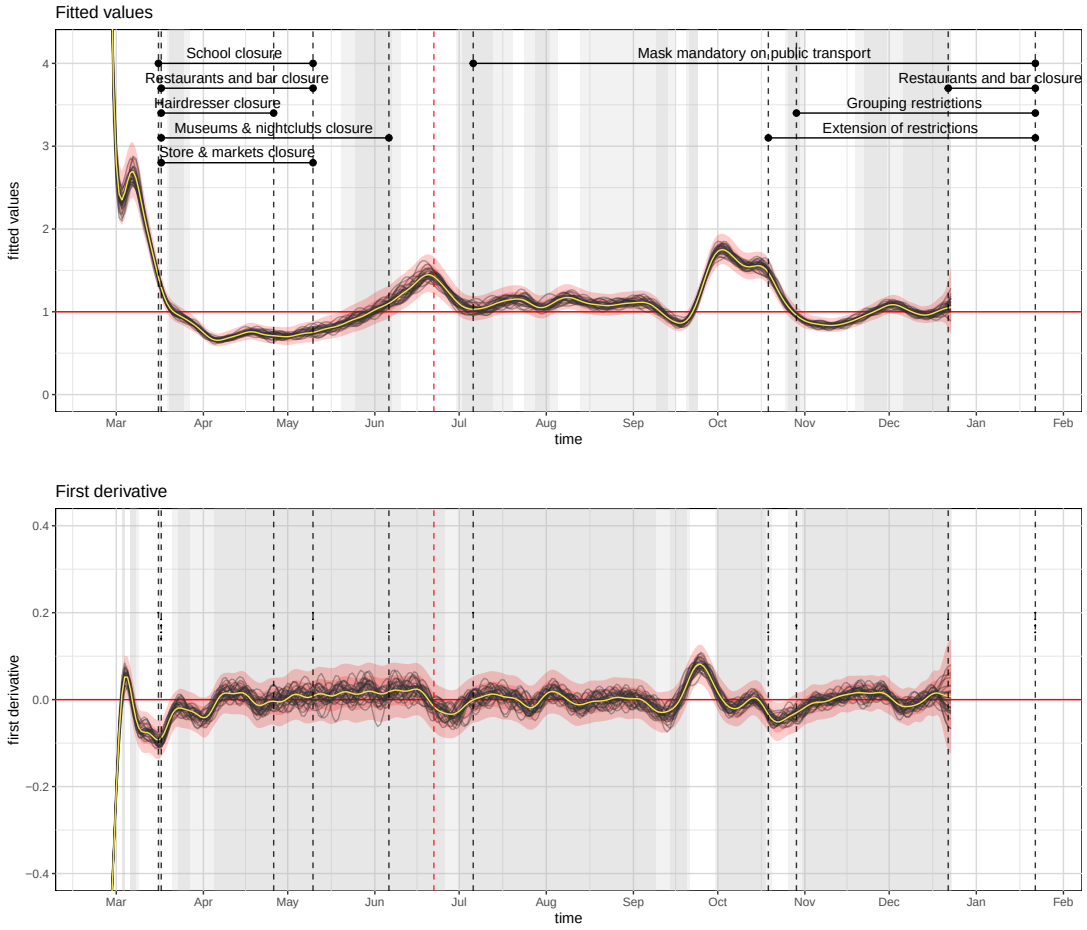


Figure 16: (top) Estimates of R_t and (bottom) its first derivative, for Switzerland data. The estimate is shown by the solid yellow curve, with in light red the point-wise and simultaneous confidence intervals. The solid black curves are splines curves sampled from the estimates of the parameters of the model. The grey regions correspond to period where the critical threshold (either 1 for the fitted values in (top), or 0 for the derivatives in (bottom)) is between the confidence bands. The main restrictive measures are shown with dashed vertical lines, so as to assess their effectiveness.

2020 were above 1.5, and decreased since. As of December 24, 2020, the estimates for the reproductive number are slightly above 1, but we can not exclude the threshold of 1, as the confidence bands are larger at the end due to the infected individuals that are not all reported yet. The stopping of some sanitary measures had a surprising impact on the evolution of the estimated reproductive number, as in late June 2020 (shown by the vertical red dashed line in Figure 16), when restaurants and nightclubs were no longer constrained to close between midnight and 6 a.m. [FOPH, 2020]. Similarly, the increase from mid-September to October was not due to the lifting of a sanitary measure, but correlates with the resumption of classes and teaching in schools. Measures imposed on October 19, 2020 had a clear impact on the estimated reproductive number, accelerating its decrease, with an estimated first derivative being well below the threshold 0.

The approach that uses the SIR model does not need to have estimates for the generation interval, but requires the curves for deceased and recovered individuals, which need to be deconvolved as well as the incidence curve. Thus, the delay interval need to be estimated as well, and may

not correspond to the incubation period. In particular, the date at which an infected person recovered might be hard to assess, and under-estimating the mean for the delay distribution of recovered cases can cause the reproductive number to have high variability during periods of transition of the epidemic (see Section 4.1). For these reasons, we used a delay distribution with mean 16 days and variance 8 days, which is nearly three times the incubation period. These values were found using grid search over the space of parameters for the delay distribution (the mean and the variance), and computing the mean square error with the estimates of R_t from the first approach. These estimates are shown in Figure 17, along with the main sanitary measures that came into force in Switzerland in 2020. Estimates are qualitatively similar, and we can distinguish the main transition periods of the reproductive number. In particular, estimates of R_t went below 1 during the first wave at the end of March 2020, and stayed below this threshold for more than two months. However, estimates with the SIR model reached the value 0.5, much below the minimum values obtained with the approach based on the generation interval. This difference might be caused by several factors. The first is the deconvolution of the curve of recovered (and deceased) cases, which showed large variability. The second might come from the distribution of the recovery delay, that can change over the year depending on the saturation of health systems that prioritize the detection of new cases. As of December 24, 2020, the estimates of R_t from the SIR model are above 1, well over the estimates of the method based on the generation interval. However, the confidence bands are very large, nearly including all the values between 1 and 2. These bad estimates for the last values for the SIR model are due the small number of reported recovered cases since mid-December (see Figure 21c Appendix C), as a result of the end of the year. As such, the estimated recovered cases are declining more rapidly than the new infected ones, and the estimates of R_t are above 1. This is one main drawback of the SIR model, that is, the need to have access to a daily report of the recovered cases, which in practice can be harder to obtain than the curve of newly infected individuals. The advantage is the possibility to estimate the transmission and removal rates of the epidemic, as shown in Figure 24 Appendix C.

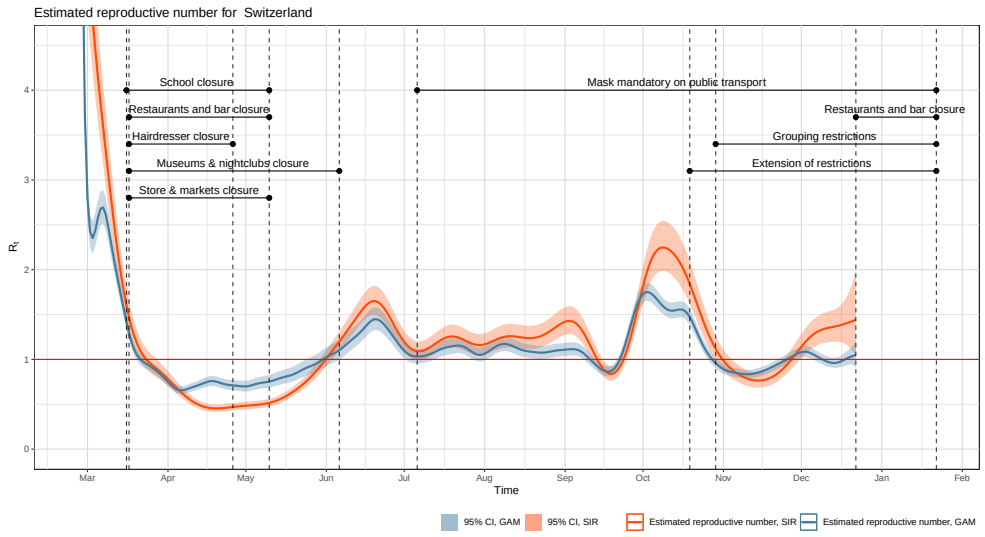


Figure 17: Estimated R_t for Switzerland with the two approaches, namely the one based on the generation interval (in blue) and the one based on the SIR model (in red). Their 95% confidence intervals are shown in their respective lighter color, but the SIR estimates do not account for the uncertainty of the deconvolution step, which explains why the confidence intervals are smaller for the last values.

6 Conclusion

Through this report, we reviewed and extended several approaches to estimate the reproductive number of an epidemic, and apply them for the COVID-19 epidemic in Switzerland. The reproductive number, R_t, which is defined as the average number of secondary cases infected by an infected individual at day t, can give important insight as to how a disease outbreak is evolving. There exists several methods to estimate R_t, among which that of Cori et al. [2013], which has proved to be very effective. As this method and the ones we designed rely on true infection events I_t, our first step was to develop tools to recover the true number of newly infected individuals from the observed one, given a distribution for the incubation period.

Roll back deconvolution is the simplest one, but tends to smooth the observed curve and struggles when the incidence curve has large peaks of new cases. A deconvolution based on the optimization of a Poisson response can be used, and gives very satisfying results given a sufficiently smooth observed incidence curve. However, the confidence bands are hardly obtainable, and this method does not handle well noisy data that have weekend effects. These issues are caused by an over-parameterization of the model, as one parameter is included for each day, making this method not scalable to epidemics that spread over time. The final deconvolution method solves this problem by assuming that the incidence curve is a smooth function of time. We used a generalized additive model to estimate I_t, with negative binomial response to deal with overdispersion. This method handle noisy data well, all the more so when a penalization term is included. The number of spline functions and the penalization term can be chosen so as to minimize the square difference between the theoretical (that is, computed from the estimates of I_t) and observed incidence curve. The main problem that arises when deconvolving the data is that a large bias is produced at the end of the reported case curve. This is because the last values rely on few observations, and are sampled from the tail of the distribution of the incubation period. It is for this reason that real-time monitoring of R_t is in practice hard. All three deconvolution methods suffer from this problem, and as such a prediction step can be added in order to extend the observed data. However, only deconvolution with a GAM can take into account the uncertainty of the prediction for estimating I_t in a efficient way, by weighting the observations in the P-IWLS algorithm.

We explored two methods for estimating R_t. One is based on the generation interval presented in Section 4.1, the other one on the epidemiological SIR model. For the first, we used spline functions to model the reproduction number as a smooth function of time, using a generalized additive model with Poisson response and log link function to constrain R_t to take positive values. We reviewed the effects of incorrectly specifying the generation interval, which causes large bias when R_t is not close to the critical threshold 1, as during the early phase of an epidemic. One can also add in the model lockdown variables, so as to assess the effectiveness of sanitary measures taken to contain the epidemic. In practice, this turns out to be hardly usable, as the effects can be only assessed when the basis dimension is sufficiently low, which would in turn make the estimates poorer. As such, we preferred to estimate the effect of intervention measures using estimates of the derivative of R_t, using a finite difference method.

The approach based on SIR model does not require the generation interval, but the daily report of infected, recovered and deceased cases. Here as well, the incidence and recovery curves need to be deconvolved in order to obtain an unbiased estimate of R_t. By interpreting the infection and recovery updates as generalized additive models, and assuming that the transmission and removal rates can be modelled by spline functions, we were able to estimate the reproductive number R_t. Missspecifying the distribution of the period of recovery can have a large impact on estimates of R_t, and can introduce instabilities in the estimates when the mean of delay recovery is too large. In this method as well, the uncertainties from the deconvolution can be relayed to

the estimator of R_t, so that recent observations and those that have large confidence bands are smaller weights in the inference.

Finally, we applied the proposed tools to estimate true incidence events and the reproductive number of the COVID-19 epidemic in Switzerland. Characterized by a strong weekly pattern, we used a deconvolution step with smoothing splines in order to recover the true infection curve, using an incubation period with mean 5.1 days. We estimated R_t using the approach based on a generation interval and the SIR model, and explored the effectiveness of the sanitary measures that have been put in place to contain the epidemic. Even if the surge in the number of cases in late September is not linked to the ease of some health measures, its decline in late October was heavily helped by the introduction of various restrictions. The method based on the SIR model, although having a strong epidemiological interpretation, suffers from a lack of reported recovered cases during the end of year 2020, which has the effect of increasing the estimates of R_t and their confidence intervals during this period. Together, the deconvolution and R_t estimation methods presented in this report can be used as a real-time monitoring of the evolution of the COVID-19 disease in Switzerland and other countries.

A Deconvolution

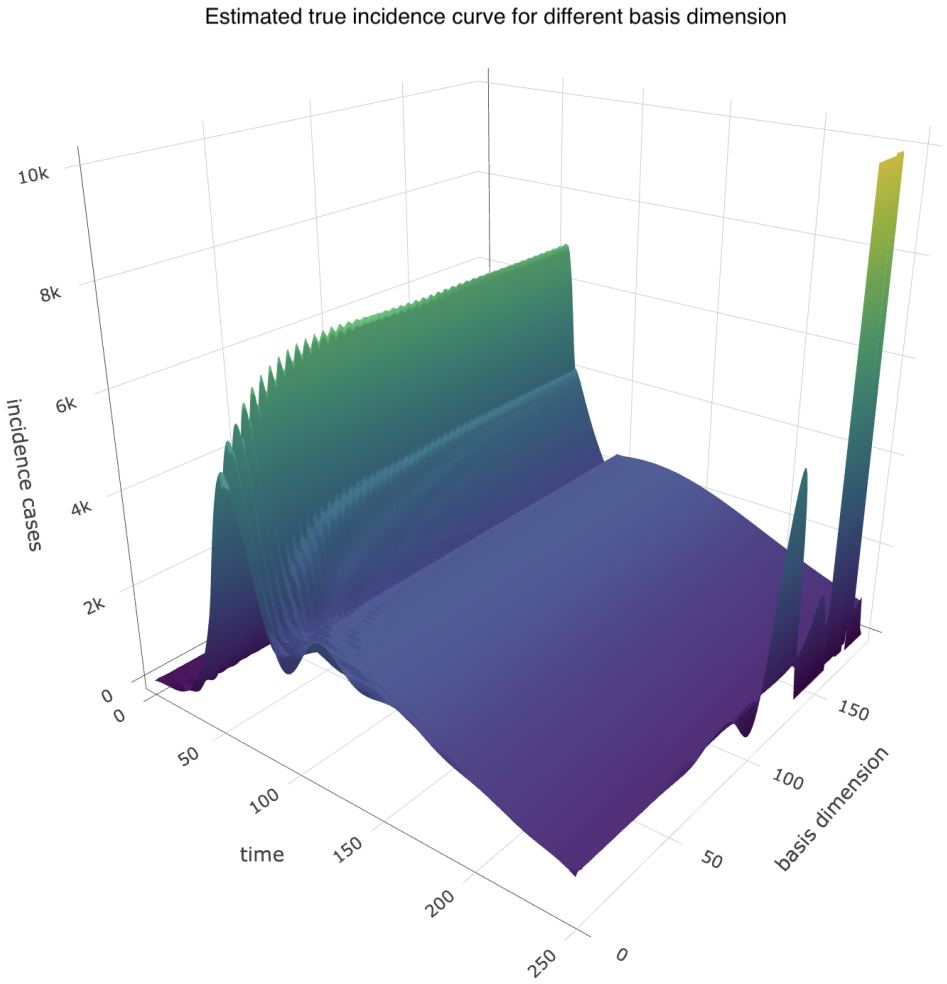
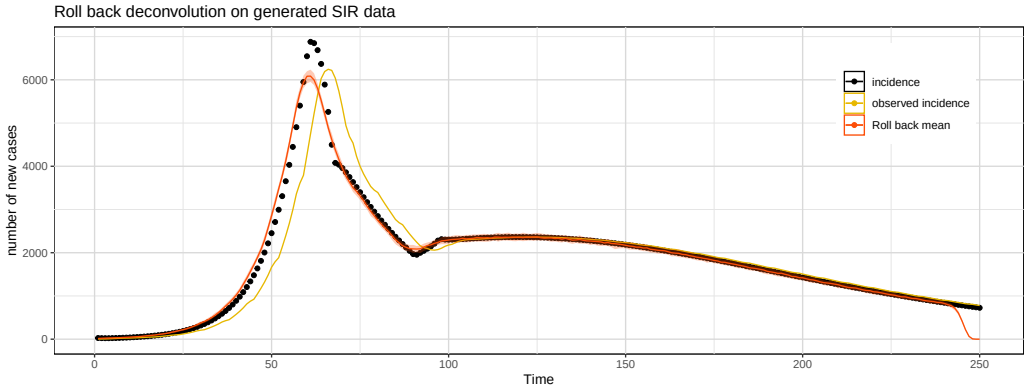
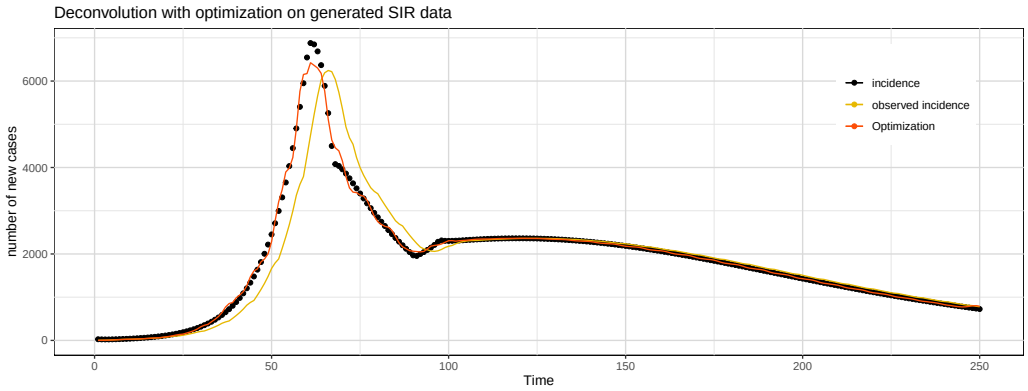


Figure 18: Illustration of the effect of the dimension of the spline basis on the estimates of the incidence curve. The higher the basis dimension is, the better is estimated of true incidence events. However, this creates explosion near the end of the range of fitted values, due to the spline functions. This can be overcome by adding an extra prediction step after the last observed values.



(a)



(b)

Figure 19: Results of the deconvolution methods for incidence curves with a strong weekly pattern: (a) Roll back deconvolution, (b) deconvolution with optimization of a Poisson response (Section 3.3). Here, the observations are first smoothed using a 7-days moving average. These smooth observations are represented by the solid yellow curve.

B Estimating R_t

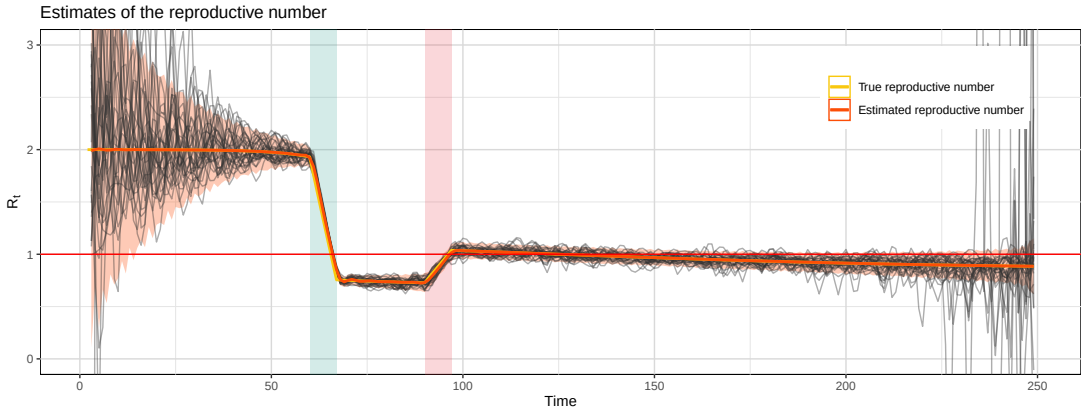
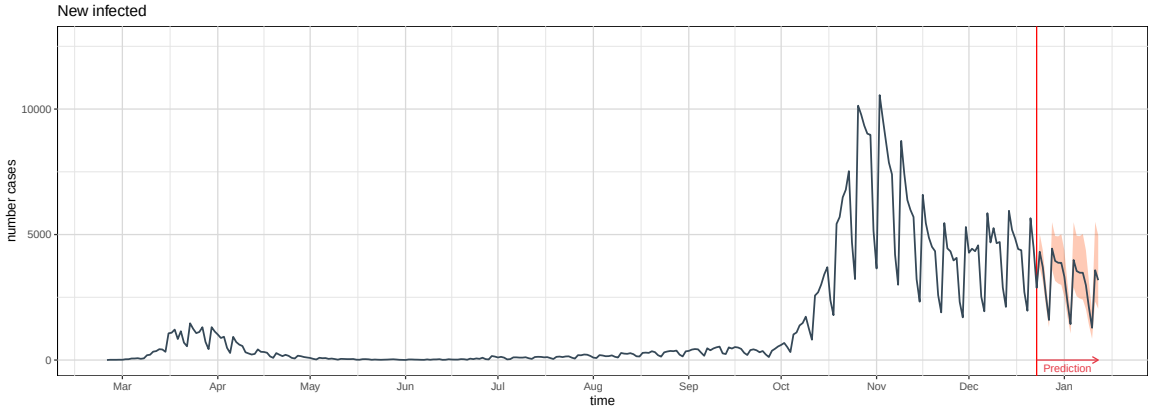
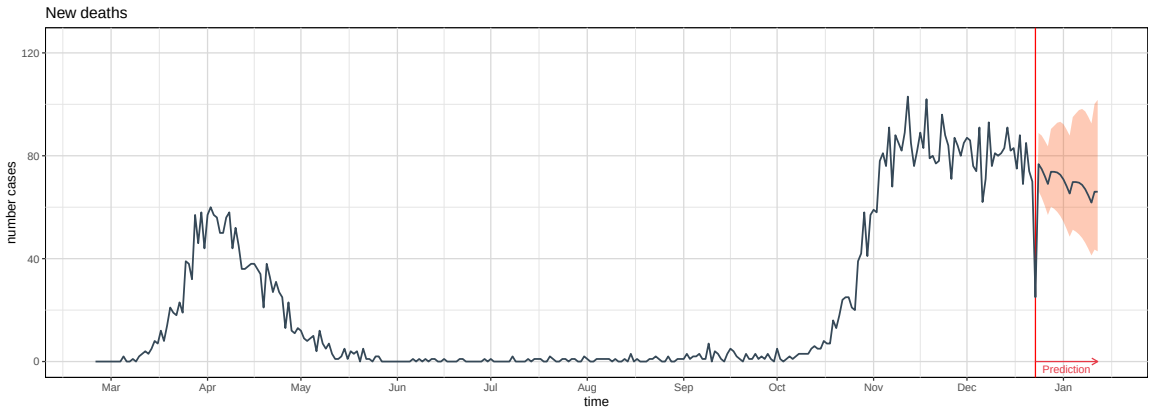


Figure 20: Estimation of R_t based on the SIR approach, with non restricted parameters for the transmission and removal rates β_t and γ_t. The confidence intervals are very wide both for the initial and final phase of the epidemic. The black lines are simulated reproductive number curves based on the mean estimate and covariance matrix for the $\{\beta\}_j$ and $\{\gamma\}_j$.

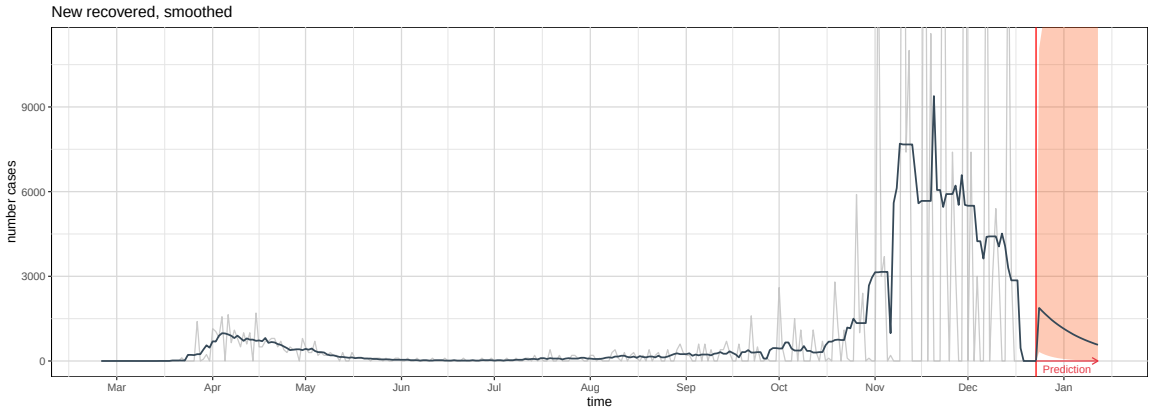
C Switzerland



(a)



(b)



(c)

Figure 21: Reported curves for the new infected (a), deaths (b) and recovered (c) cases, with a 20 day prediction with 95% confidence intervals in light red. For the recovered cases curve, the original data are shown in light gray, and the smoothed ones (using a 7-day moving average) are used to compute the next 20 values. No penalty for extrapolating were used, since it would lead (for this particular prediction time to visually unlikely results).

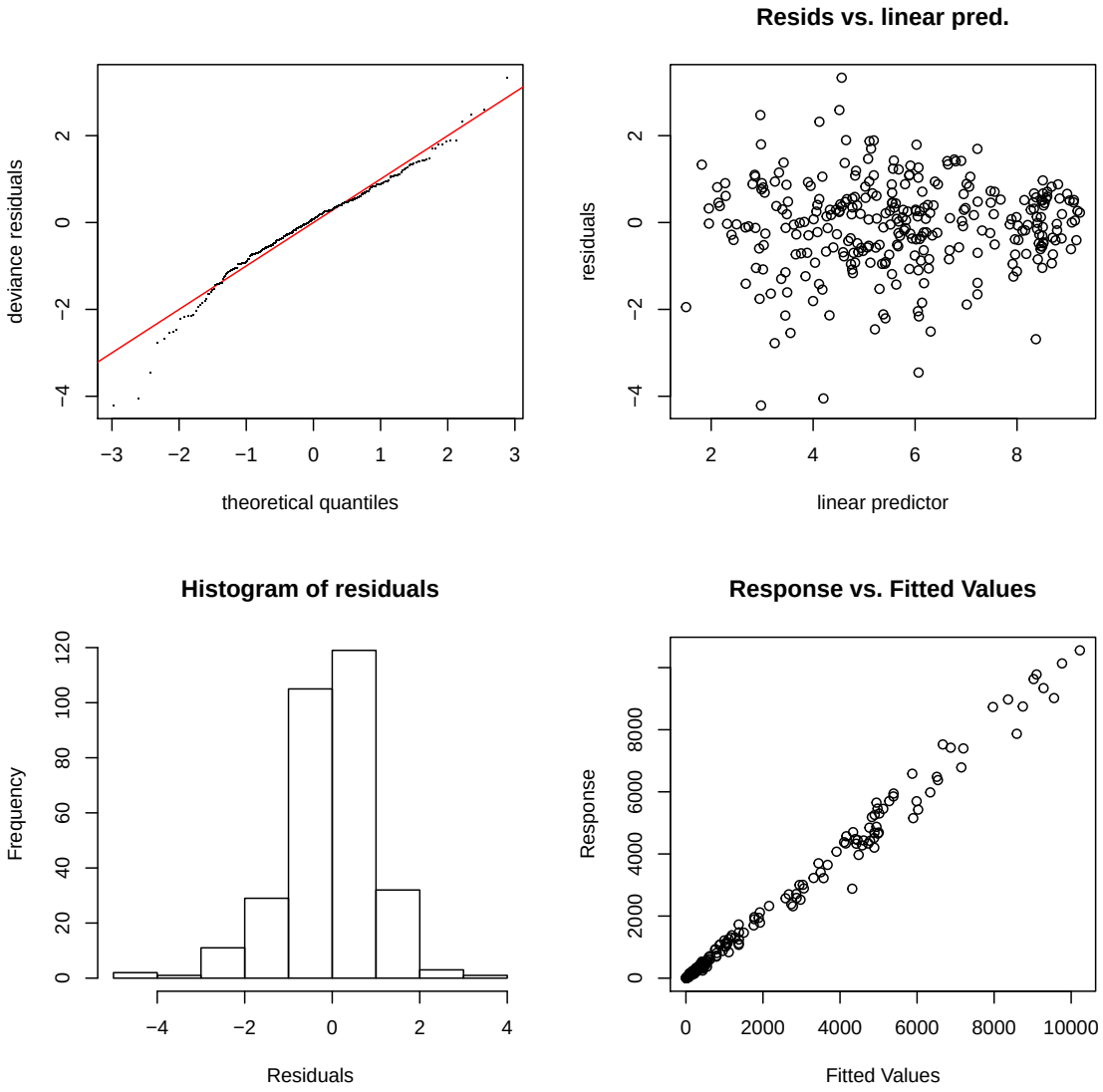
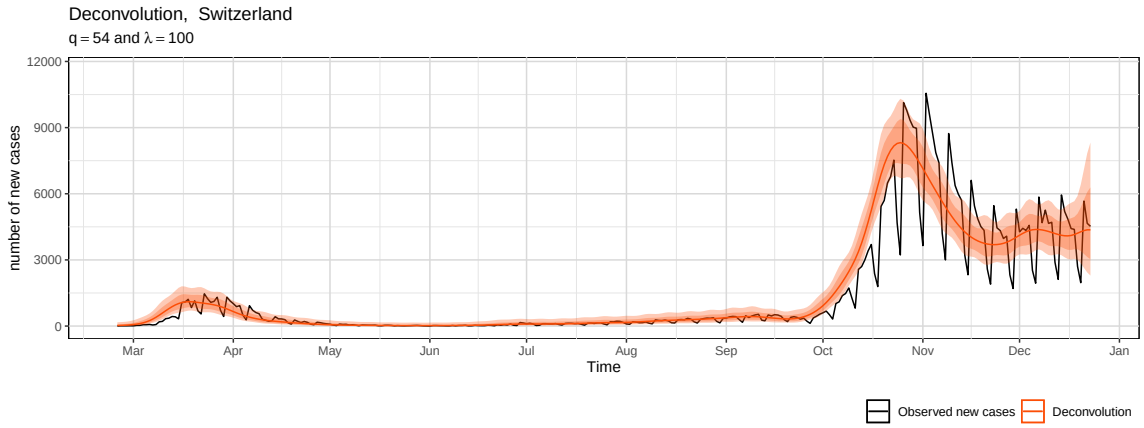
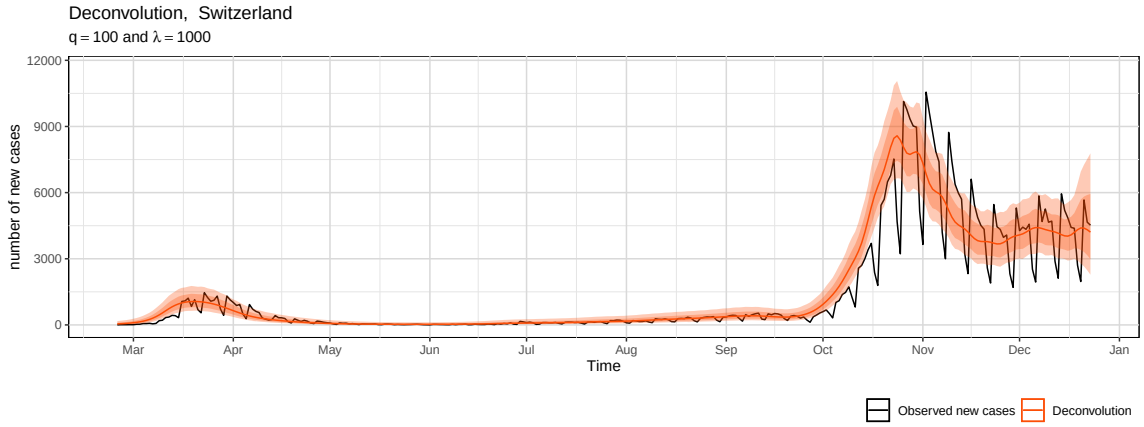


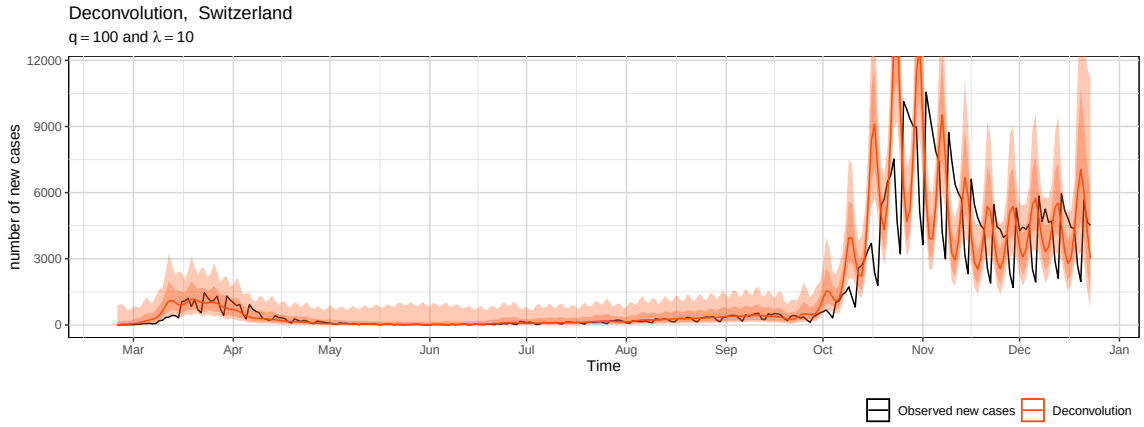
Figure 22: Diagnostic plots for the extrapolation process, from the Negative Binomial model with mean as defined in (5.1). The residuals do not show particular pattern against the linear predictor. As for the quantile-quantile plot, there is very slight signs of under-dispersion, but this is the best model we were able to reach for the extrapolation.



(a)



(b)



(c)

Figure 23: Estimated true incidence curves, using different dimension for the spline basis q and the penalization term λ. In (a), both q and λ were selected so as to minimize the mean square error between the convolved estimated true incidence and a 7-day moving average (MA) of the observed cases. For (b), no MA was used for selecting q, and in (c), no MA was used for both q and λ, which resulted in overfitting the observed curve, so that weekly pattern are now emphasized in the deconvolution. Using such estimated true incidence events can lead to very poor estimates for the reproductive number R_t. This is why we preferred to choose the curve (a) for estimating R_t.

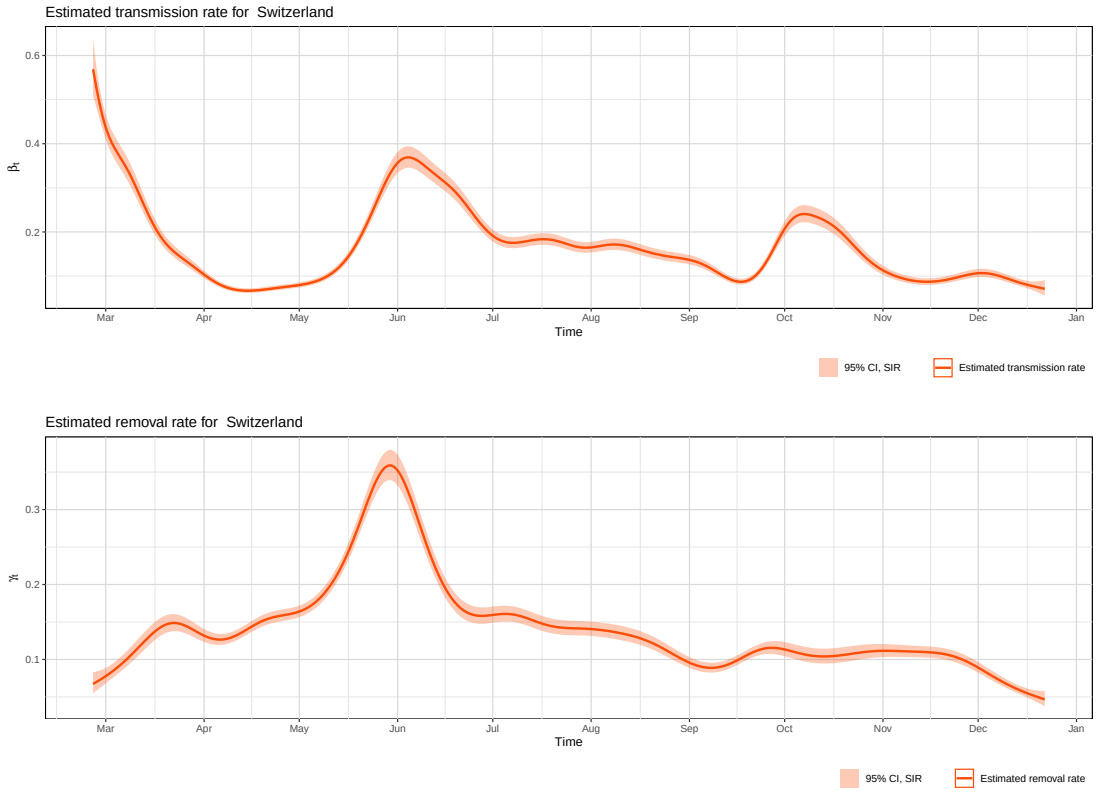


Figure 24: Estimates of the (top) transmission and (bottom) removal rates β_t and γ_t using the SIR model. The confidence bands are shown in light red.

References

Abbott, S., Hellewell, J., Munday, J., Thompson, R. and Funk, S. (2020a). Epinow: Estimate realtime case counts and time-varying epidemiological parameters. https://github.com/epiforecasts/EpiNow.

Abbott, S., Hellewell, J., Thompson, R., Sherratt, K., Gibbs, H., Bosse, N., Munday, J., Meakin, S., Doughty, E., Chun, J. Y., Chan, Y.-W., Finger, F., Campbell, P., Endo, A., Pearson, C., Gimma, A., Russell, T., Flasche, S., Kucharski, A. and Funk, S. (2020b). Estimating the time-varying reproduction number of sars-cov-2 using national and subnational case counts. *Wellcome Open Research* 5: 112, doi:10.12688/wellcomeopenres.16006.1.

Bettencourt, L. and Ribeiro, R. (2008). Real time bayesian estimation of the epidemic potential of emerging infectious diseases. *PloS one* 3: e2185, doi:10.1371/journal.pone.0002185.

Chaudhuri, S., Handcock, M. S. and Rendall, M. S. (2006). glmc: An R package for generalized linear models subject to constraints.

Cori, A., Ferguson, N., Fraser, C. and Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology* 178: 1505–1512, doi:10.1093/aje/kwt133.

Davison, A. C. (2003). *Statistical Models*. Cambridge University Press, doi:10.1017/CBO9780511815850.

Davison, A. C. (spring 2020). *Modern Regression Methods, MATH-408 course*. EPFL.

ECDC (2020). European Centre for Disease Prevention and Control, Data on hospital and ICU admission rates and current occupancy for COVID-19. https://www.ecdc.europa.eu/en/publications-data, retrieved December 24, 2020.

Fahrmeir, L. and Lang, S. (2001). Bayesian inference for generalized additive mixed models based on markov random field priors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 50: 201–220, doi:https://doi.org/10.1111/1467-9876.00229.

FOPH (2020). Federal Office of Public Health. https://www.covid19.admin.ch/en/overview, retrieved December 24, 2020.

Fraser, C. (2007). Estimating individual and household reproduction numbers in an emerging epidemic. *PLOS One* 2: e758, doi:10.1371/journal.pone.0000758.

Goldstein, E., Dushoff, J., Ma, J., Plotkin, J. B., Earn, D. J. D. and Lipsitch, M. (2009). Reconstructing influenza incidence by deconvolution of daily mortality time series. *Proceedings of the National Academy of Sciences* 106: 21825–21829, doi:10.1073/pnas.0902958106.

Gostic, K. M., McGough, L., Baskerville, E., Abbott, S., Joshi, K., Tedijanto, C., Kahn, R., Niehus, R., Hay, J. A., De Salazar, P. M., Hellewell, J., Meakin, S., Munday, J., Bosse, N., Sherratt, K., Thompson, R. M., White, L. F., Huisman, J., Scire, J., Bonhoeffer, S., Stadler, T., Wallinga, J., Funk, S., Lipsitch, M. and Cobey, S. (2020). Practical considerations for measuring the effective reproductive number, r_t. *medRxiv* doi:10.1101/2020.06.18.20134858.

HDX (2020). Humanitarian Data Exchange, Novel Coronavirus (COVID-19) Cases Data. https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases, retrieved December 24, 2020.

Hong, H. G. and Li, Y. (2020). Estimation of time-varying reproduction numbers underlying

epidemiological processes: A new statistical tool for the COVID-19 pandemic. *PLOS One* 15: 1–15, doi:10.1371/journal.pone.0236464.

JHU (2020). Johns Hopkins University's Coronavirus Resource Center. https://coronavirus.jhu.edu/, retrieved December 24, 2020.

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G. and Lessler, J. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine* 172: 577–582, doi:10.7326/M20-0504, pMID: 32150748.

Lucy, L. B. (1974). An iterative technique for the rectification of observed distributions. *Astron. J.* 79: 745–754, doi:10.1086/111605.

Pan, A., Liu, L., Wang, C., Guo, H., Hao, X., Wang, Q., Huang, J., He, N., Yu, H., Lin, X., Wei, S. and Wu, T. (2020). Association of public health interventions with the epidemiology of the COVID-19 outbreak in wuhan, china. *The Journal of the American Medical Association* 323, doi:10.1001/jama.2020.6130.

Perasso, A. (2018). An introduction to the basic reproduction number in mathematical epidemiology. *ESAIM: Proceedings and Surveys* 62: 123–138, doi:10.1051/proc/201862123.

Richardson, W. H. (1972). Bayesian-based iterative method of image restoration. *J. Opt. Soc. Am.* 62: 55–59, doi:10.1364/JOSA.62.000055.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71: 319–392, doi:https://doi.org/10.1111/j.1467-9868.2008.00700.x.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press, doi:10.1017/CBO9780511755453.

Simpson, G. L. (2018). Modelling palaeoecological time series using generalised additive models. *Frontiers in Ecology and Evolution* 6: 149, doi:10.3389/fevo.2018.00149.

Tapiwa, G., Cécile, K., Dongxuan, C., Andrea, T., Christel, F., Jacco, W. and Niel, H. (2020). Estimating the generation interval for COVID-19 based on symptom onset data. *medRxiv* doi:10.1101/2020.03.05.20031815.

Velthuis, A., De Jong, M., Stockhofe, N., Vermeulen, T. and Kamp, E. (2002). Transmission of actinobacillus pleuropneumoniae in pigs is characterized by variation in infectivity. *Epidemiology and Infection* 129: 203–214, doi:10.1017/S0950268802007252.

Velthuis, A. G. J., Bouma, A., Katsma, W. E. A., Nodelijk, G. and De Jong, M. C. M. (2007). Design and analysis of small-scale transmission experiments with animals. *Epidemiology and Infection* 135: 202–217, doi:10.1017/S095026880600673X.

Wallinga, J. and Teunis, P. (2004). Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology* 160: 509–16, doi:10.1093/aje/kwh255.

WHO (2020). World health organization coronavirus update. https://www.who.int/emergencies/diseases/novel-coronavirus-2019, retrieved May 5, 2020.

Wood, S. N. (2008). Fast stable direct fitting and smoothness selection for generalized additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70: 495–518, doi:https://doi.org/10.1111/j.1467-9868.2007.00646.x.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73: 3–36, doi:https://doi.org/10.1111/j.1467-9868.2010.00749.x.

Contents

List of Figures