# A SPLINE-BASED TIME-VARYING REPRODUCTION NUMBER FOR MODELLING EPIDEMIOLOGICAL OUTBREAKS

Eugen Pircalabelu

LIDAM Discussion Paper ISBA
2021 / 30

UCLouvain

ISBA | LIDAM
Louvain Institute of Data Analysis and Modeling in economics and statistics

# A spline-based time-varying reproduction number for modeling epidemiological outbreaks

**Eugen Pircalabelu**

April 10, 2021

*UCLouvain, Institute of Statistics, Biostatistics and Actuarial Sciences*
*Voie du Roman Pays 20, 1348 Louvain-la-Neuve, Belgium*
eugen.pircalabelu@uclouvain.be

**Abstract**

We develop in this manuscript a method for performing estimation and inference for the reproduction number of an epidemiological outbreak. The estimator is time-dependent and uses spline modeling to adapt to changes in the outbreak. This is accomplished by directly modeling the series of new infections as a function of time and subsequently using the derivative of the function to define a time-varying reproduction number.

*Keywords: reproduction number, time dependency, spline modeling, COVID-19*

*Classification: Physical Sciences - Statistics; Biological Sciences - Applied Biological Sciences*

Significance statement: This research helps epidemiologists, stakeholders and policy-makers alike to asses the spread of a virus in a susceptible population by keeping track of the reproduction number of the virus. Based on the daily number of new infections it can easily and in real-time be assessed how fast the virus spreads. The procedure is flexible enough to allow accounting for interventional measures that attempt to keep the epidemic under control.

## 1 Introduction

As more and more governments try to balance the relaxation of lock-down measures for the population while keeping the COVID-19 pandemic and its effects under control, it has become ever more clear that trustworthy metrics which quickly and accurately inform policy-makers of the state of evolution of the outbreak are of crucial importance. Different countries have used different interventional strategies to keep the spread of the SARS-CoV-2 virus under control and it is natural to assess the efficacy of such measures being enforced. Multiple key indicators are being focused upon, but one metric that is fundamental to the assessment of the spread of the virus in a susceptible population is the *basic reproduction number* ($R_0$) of the virus which is a measure of its transmission capacity. It represents an average number of secondary infections that an infected person can produce in the population, where it is assumed that every person is susceptible (Dietz, 1993).

It is a useful metric to follow during outbreaks as its magnitude serves to determine to which degree interventional strategies are necessary to prevent an epidemic or to maintain it at acceptable levels. Epidemic theory associates outbreaks of a certain disease when the reproduction numbers are estimated to be greater than one, and consequently the spread of the virus is judged to be under control when the reproduction numbers are estimated to be lower than one. In the context of the COVID-19 pandemic, many recent studies give different estimates for this number, very much dependent on the assumptions the model makes, the country where the outbreak is observed and the period when the analysis was performed. See Liu et al. (2020), Tsang et al. (2020) or Salje et al. (2020) among a rapidly growing line of research.

Different approaches exist to estimate the reproduction number, however the large majority estimate one single number for the entire period in which the virus is active, which is an

unrealistic, constant throughout time summary measure of the virus potential in the population. Notable exceptions include Wallinga and Teunis (2004), Wallinga and Lipsitch (2007), Hens et al. (2011) and Cori et al. (2013). More recently, sparked by the developments regarding the spread of SARS-CoV-2, Hong and Li (2020) and Koyama et al. (2021) also proposed time-varying reproduction numbers either based on a time-dependent susceptible-infectious-removed (SIR) model or a state space model.

## 2 Proposed model

Following the work of Wallinga and Lipsitch (2007) and Obadia et al. (2012) we have that the basic reproduction number $R_0$ can be obtained as:

$$R_0 = 1/M(-r) \tag{1}$$

where $r$ is an exponential growth rate of the disease, defined as the per capita change in the number of new cases per unit of time and $M(\cdot)$ is the moment generating function of the generation time (defined as the mean duration between the time of infection of a primary infectee and that of a secondary infectee) distribution. It is assumed here $M(-r)$ exists and that $M(\cdot)$ is known.

To propose a time-varying reproduction number $R(t)$, we consider first $I(t)$ as being the number of members in the population that get infected at time $t$. We assume next, that $I(t)$ follows a model from the exponential family (EF) as:

$$\begin{aligned} I(t) &\sim \text{EF}(\mu(t), \phi) \\ g\{\mu(t)\} &= f(t) \end{aligned} \tag{2}$$

where $\mu(t) = \text{E}\{I(t)\}$ is the expected number of new infections at time $t$, $g(\cdot)$ is a known link function and $\phi$ is a dispersion parameter. Moreover, we consider $f(t)$ to be a smooth, unknown function of time.

As eq. (1) requires a change in the number of new cases per unit of time, it seems natural to augment $R_0$ to the metric $R(t)$ defined as:

$$R(t) = 1/M\{-r(t)\}$$

with $r(t) = \frac{d}{dt} f(t)$. The rationale behind the proposed instantaneous or effective $R(t)$ is first, the fact that $r$ plays in eq. (1) the role of a growth rate across time which is substituted by the derivative of a flexible function that can adapt better to fluctuations across time. Secondly, the Poisson assumption that is used often in practice to model infectious counts, might not be an appropriate one due to the overdispersion phenomenon that is quite prominent for such count data. By allowing for a larger class of models from the exponential family, we provide more flexibility. The idea of introducing a time-varying reproduction number $R(t)$ based on the derivative of a spline model is a novel approach that has not been explored so far.

*Remark* 1. Note that model (2) offers a relatively large palette of possible models with the usual Poisson distributional assumption that is most often used in practice as a special case. If desired, one could simply replace the Poisson distributional assumption in favor of any other, more appropriate distribution such as the negative binomial or a quasi-Poisson model, that directly account for overdispersion. Also, if other time-dependent covariates $X_t$ are available (such as vaccination information, weekend effects or mobility information), one could simply consider an extended additive model of the form $g\{\mu(t)\} = f_1(t) + f_2(X_t)$ without much difficulty.

As $f(t)$ is a considered a smooth function of time we propose to model it in this manuscript with splines and we consider for this application truncated polynomials and radial bases splines:

$$f(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \ldots + \beta_m t^m + \sum_{k=1}^{K} u_k (t - \kappa_k)_+^m$$

$$f(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \ldots + \beta_m t^m + \sum_{k=1}^{K} u_k |t - \kappa_k|^m$$

where $(\kappa_1, \ldots, \kappa_K)^\mathsf{T}$ are a set of specified knots $1 < \kappa_1 < \ldots < \kappa_K < T$, $(a)_+ = \max(a, 0)$, while $(\beta_0, \beta_1, \ldots, \beta_m)^\mathsf{T}$ and $(u_1, \ldots, u_K)^\mathsf{T}$ are unknown coefficients.

The reason for these bases choices is two-fold. First, with such simple polynomial models it is straightforward to obtain $\frac{d}{dt} f(t)$ and secondly, this choice leads to computationally efficient estimation algorithms that exploit connections with a mixed model reformulation as will be detailed next. We note however, that $B$-splines (de Boor, 2001) and $P$-splines as in the formulation of Eilers and Marx (1996) could be interesting alternatives, with slightly more involved derivative expressions due to recursive formulas.

In matrix notation we have that $f(t)$ can be rewritten as $f(t) = [\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}]_t$ where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_m)^\mathsf{T}$, $\boldsymbol{u} = (u_1, \ldots, u_K)^\mathsf{T}$, $\boldsymbol{X} = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ 1 & 2 & \ldots & 2^m \\ \vdots & \vdots & \vdots & \ddots \\ 1 & T & \ldots & T^m \end{bmatrix}$ and in the case of truncated polynomials $\boldsymbol{Z} = \begin{bmatrix} (1-\kappa_1)_+^m & (1-\kappa_2)_+^m & \ldots & (1-\kappa_K)_+^m \\ (2-\kappa_1)_+^m & (2-\kappa_2)_+^m & \ldots & (2-\kappa_K)_+^m \\ \vdots & \vdots & \vdots & \ddots \\ (T-\kappa_1)_+^m & (T-\kappa_2)_+^m & \ldots & (T-\kappa_K)_+^m \end{bmatrix}$. A very similar expression for $\boldsymbol{Z}$ holds for radial bases splines where one uses the terms $|t - \kappa_k|^m$. We assume next that $\boldsymbol{\beta}$ represents a fixed set of parameters and $\boldsymbol{u}$ represents a set of Gaussian random variables. Conditionally on $\boldsymbol{u}$, eq. (2) becomes

$$\begin{aligned} I(t)|\boldsymbol{u} &\sim \text{EF}(\mu(t), \phi) \\ \boldsymbol{u} &\sim N(\boldsymbol{0}, \boldsymbol{G}) \\ g\{\mu(t)\} &= f(t). \end{aligned} \tag{3}$$

For simplicity we assume that $\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{G} = \sigma_u^2 \boldsymbol{I})$, where $\sigma_u$ is an unknown parameter and $\boldsymbol{I}$ is the identity matrix of dimension $K \times K$.

Model (3) is relatively standard in the treatment of splines reparametrized as random effects models and the monograph of Ruppert et al. (2003) provides an excellent exposition on the subject.

## 3 Estimation aspects

In this section we tackle the estimation of $R(t)$ by presenting a step by step procedure that estimates all necessary quantities.

In the exponential family model with normal random effects presented in (3), the likelihood contribution at time $t$ has the form:

$$p(I_t|\boldsymbol{u}) = \exp\left\{ \frac{I_t \theta_t - b(\theta_t)}{\phi} + c(I_t, \phi) \right\}$$

for some functions $b(\cdot)$ and $c(\cdot)$ and where $\theta_t$ is the natural parameter of the exponential family, while the density of the random effects is $p(\boldsymbol{u}) \propto \exp\{-\frac{1}{2}\boldsymbol{u}^\mathsf{T}\boldsymbol{G}^{-1}\boldsymbol{u}\}$. This implies that the likelihood function becomes:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}) &= \int_{\mathbb{R}^K} p(\boldsymbol{I}_t|\boldsymbol{u})p(\boldsymbol{u})d\boldsymbol{u} \\ &\propto \int_{\mathbb{R}^K} \exp\left\{\sum_{t=1}^{T}\frac{I_t\theta_t - b(\theta_t)}{\phi} - \frac{1}{2}\boldsymbol{u}^\mathsf{T}\boldsymbol{G}^{-1}\boldsymbol{u}\right\}d\boldsymbol{u} \end{aligned} \tag{4}$$

where $\boldsymbol{I}_t = (I_1, I_2, \ldots, I_T)^\mathsf{T}$. Due to the presence of a $K$-dimensional integral, direct maximization of (4) is regarded as a highly expensive computational problem. Breslow and Clayton (1993) approximate this integral based on Laplace's approximation and based on their result, in practice $\boldsymbol{\beta}$ is estimated by maximizing the penalized (conditional) log-likelihood defined as:

$$\ell(\boldsymbol{\beta}, \boldsymbol{u}) = \log(p(\boldsymbol{I}_t|\boldsymbol{u})) - \frac{1}{2}\boldsymbol{u}^\mathsf{T}\boldsymbol{G}^{-1}\boldsymbol{u}.$$

Assume next for ease of exposition that one works under the canonical link function ($g(\cdot) = b^{-1}(\cdot)$) implying that

$$\begin{aligned}\theta_t = \eta_t &= [\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}]_t = f(t) \\ b'(\eta_t) &= \mu_t = \mathrm{E}\{I(t)|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{u}\} \\ b''(\eta_t) &= \mathrm{Var}\{I(t)|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{u}\}.\end{aligned}$$

As for the iteratively reweighted least squares (IRLS) algorithm, we define next the 'working' vector

$$\begin{aligned}\boldsymbol{I}_t^w &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + (\boldsymbol{I}_t - \boldsymbol{\mu}_t)\boldsymbol{W}^{-1} \\ &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \epsilon^w\end{aligned}$$

where $\boldsymbol{\mu}_t = (\mu_1, \mu_2, \ldots, \mu_T)^\mathsf{T}$ and $\boldsymbol{W} = \mathrm{diag}\{b''(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})\}$. Since $\boldsymbol{I}_t^w$ is expressed as a linear mixed model, the penalized quasi-likelihood (PQL) method can be used to obtain estimates as it proceeds by minimizing first the expression

$$(\boldsymbol{I}_t^w - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})^\mathsf{T}\boldsymbol{W}^{-1}(\boldsymbol{I}_t^w - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}) + \boldsymbol{u}^\mathsf{T}\boldsymbol{G}^{-1}\boldsymbol{u},$$

for which the solution is obtained in closed form as

$$\begin{bmatrix}\hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}}\end{bmatrix} = (\boldsymbol{C}^\mathsf{T}\boldsymbol{W}\boldsymbol{C} + \boldsymbol{B})^{-1}\boldsymbol{C}^\mathsf{T}\boldsymbol{W}^{-1}\boldsymbol{I}_t^w$$

with $\boldsymbol{C} = [\boldsymbol{X}|\boldsymbol{Z}]$ and $\boldsymbol{B} = \begin{bmatrix}\boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G}^{-1}\end{bmatrix}$. Using the estimated coefficients one updates next the working vector $\boldsymbol{I}_t^w$ and this two-step process continues until a convergence metric is sufficiently small.

Once estimates for $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{u}})^\mathsf{T}$ are obtained, we can obtain as well the best linear unbiased predictor of $\hat{r}(t) = \frac{d}{dt}f(t)\big|_{(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{u}})^\mathsf{T}}$ and then $R(t)$ is estimated by direct plug-in $\hat{R}(t) = 1/M\{-\hat{r}(t)\}$.

*Remark* 2. Note that in this section it is assumed $\boldsymbol{G}$ is fully specified implying that $\sigma_u^2$ is known but if this is not the case, then one can obtain as well a consistent estimator for it, by using restricted maximum likelihood estimation. More details are available in Searle et al. (2006) or Stroup (2013) among many others.

We have also assumed that $\phi$, the dispersion parameter is known, but if this is not the case then a (conservative) estimator for it could be obtained as:

$$\hat{\phi} = \frac{1}{T-(m+K+1)} \sum_{t=1}^{T} \frac{I_t - \hat{\mu}_t}{b''(\hat{\eta}_t)}.$$

This would be particularly useful if one assumes a quasi-Poisson model that allows for over-dispersion relative to a standard Poisson model for which $\phi$ is known to be $\phi = 1$. Note that for such a quasi-Poisson model the estimates for the coefficients are identical to those from the standard Poisson model, but the elements of the variance-covariance matrix are multiplied by $\hat{\phi}$, generally larger than 1, hence the term 'over'-dispersion. More details are available in Agresti (2015).

# 4 Inferential aspects

It is of interest to provide standard errors to quantify the uncertainty in the estimation of $R(t)$ and as such we proceed in several steps. The first building block is realizing that determining pointwise error bars for the fitted function $\hat{f}(t) = [\boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{u}}]_t$ requires the standard error $\mathrm{SE}(\hat{f}(t))$ defined as:

$$\mathrm{SE}(\hat{f}(t)|\boldsymbol{u}) = \left(\boldsymbol{C}_t^{\mathsf{T}} \mathrm{Cov}((\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{u}})^{\mathsf{T}}|\boldsymbol{u})\boldsymbol{C}_t\right)^{1/2} \text{ with}$$
$$\mathrm{Cov}((\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{u}})^{\mathsf{T}}|\boldsymbol{u}) \approx (\boldsymbol{C}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{C} + \boldsymbol{B})^{-1}\boldsymbol{C}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{C}(\boldsymbol{C}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{C} + \boldsymbol{B})^{-1}$$

and where $\boldsymbol{C}_t$ represents the vector of data corresponding to the $t$-th row of $\boldsymbol{C}$.

Moreover, due to the nature of the spline functions we chose to use in this manuscript, one can always rewrite the predicted value for the first derivative at time $t$ as $\hat{f}'(t) = [\tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}} + \tilde{\boldsymbol{Z}}\hat{\boldsymbol{u}}]_t$ where the columns of the design matrices $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Z}}$ are directly obtainable from the columns of $\boldsymbol{X}$ and $\boldsymbol{Z}$. As such with $\tilde{\boldsymbol{C}} = [\tilde{\boldsymbol{X}}|\tilde{\boldsymbol{Z}}]$ one has direct access to estimated standard errors for the first derivative at each time point $t$:

$$\mathrm{SE}(\hat{f}'(t)|\boldsymbol{u}) = \left(\tilde{\boldsymbol{C}}_t^{\mathsf{T}} \mathrm{Cov}((\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{u}})^{\mathsf{T}}|\boldsymbol{u})\tilde{\boldsymbol{C}}_t\right)^{1/2}.$$

Using further a large sample argument justifying the distributional approximation:

$$\frac{\hat{f}'(t) - \mathrm{E}(f'(t)|\boldsymbol{u})}{\mathrm{SE}(\hat{f}'(t)|\boldsymbol{u})} \approx N(0,1)$$

and letting $h(a) = 1/M(-a)$ we have using the Delta method:

$$\frac{h\big(\hat{f}'(t)\big) - h\big(\mathrm{E}(f'(t)|\boldsymbol{u})\big)}{\mathrm{SE}(\hat{f}'(t)|\boldsymbol{u})|h'(a)|} \approx N\big(0,1\big).$$

Another possibility would be to first construct directly lower and upper bounds for $\mathrm{E}(f'(t)|\boldsymbol{u})$ as $\hat{f}'(t) \pm z_{1-\frac{\alpha}{2}}\mathrm{SE}(\hat{f}(t)|\boldsymbol{u})$ and then construct the approximate confidence interval for $R(t)$ as:

$$\left[\frac{1}{M\big(\hat{f}'(t) - z_{1-\frac{\alpha}{2}}\mathrm{SE}(\hat{f}(t)|\boldsymbol{u})\big)} \;;\; \frac{1}{M\big(\hat{f}'(t) + z_{1-\frac{\alpha}{2}}\mathrm{SE}(\hat{f}(t)|\boldsymbol{u})\big)}\right],$$

where $z_{1-\frac{\alpha}{2}}$ is the quantile of the standard normal distribution at a fixed significance level $\alpha$.

## 5 Simulation study

To test the performance of the proposed spline-based method we have conducted a simulation study where we have first created epidemics of length $T = 150$ days for which we used four different smooth $R(t)$ trajectories depicted in Figure 1 top panels. For each scenario, we simulated 1000 epidemics, starting with 100 index cases. For visualization purposes we have plotted the evolution of an epidemic from each scenario in the bottom panels of Figure 1. Scenario I corresponds to a two-wave epidemic where the second wave is more severe, Scenario II depicts a slowly increasing exponential outbreak, while Scenario III mimics an interventional mechanism that reduces the infections in the second part of the time horizon. The last scenario depicts an epidemic that is fully under control throughout the entire period.

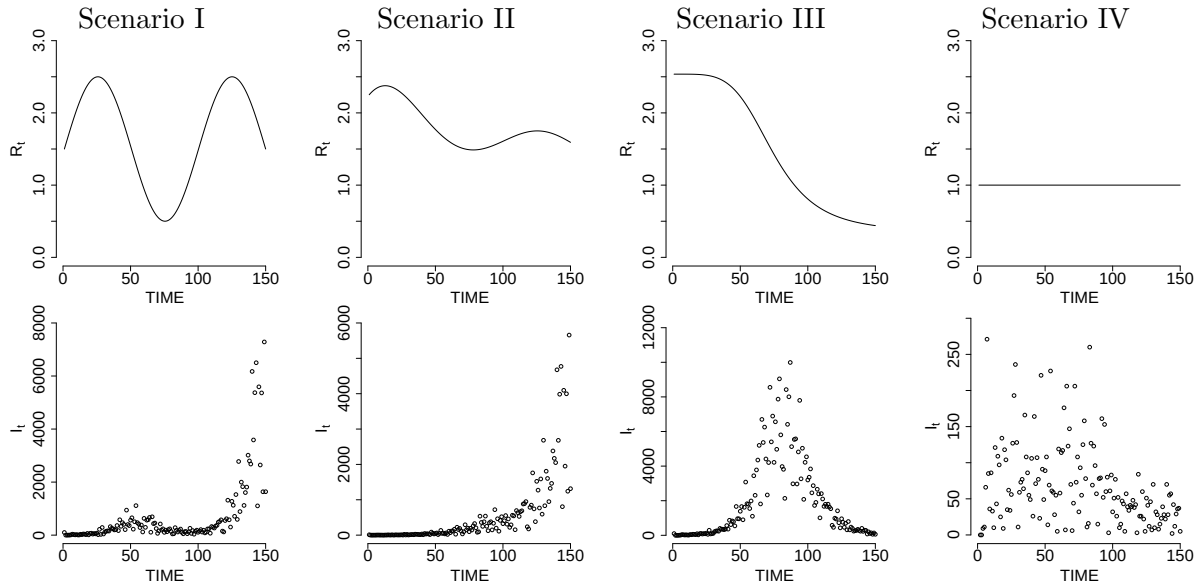


Figure 1: Simulated data from four different scenarios. Top panels correspond to different $R(t)$ functions while the bottom panels represent counts of infections $I(t)$ simulated from models corresponding to each $R(t)$.

We have simulated next data similar to what is proposed in Cori et al. (2013) where we have used a SARS-like serial interval distribution, with mean 8.4 days and standard deviation 3.8 days. For each day the number of new incident cases $I(t)$ was drawn from a negative binomial distribution with mean $\mu(t) = R(t)\sum_{s=1}^{t} I(t-s)w(s)$ and variance $\sigma^2(t) = \mu(t) + \mu(t)^2/\theta$, where $w(s)$ is the discrete serial interval distribution, $I(t-s)$ are the infections $s$ lags in time and $\theta$ is a parameter that allows for overdispersion. Three values for $\theta$ were considered as $\theta \in \{2, 10, 100\}$. The considered competitors are (i) the cubic spline-based $R(t)$ with truncated polynomial and radial bases for which the Poisson, negative binomial and quasi-Poisson distributional assumptions are made, (ii) the estimator proposed in Cori et al. (2013) and (iii) the estimator proposed in Wallinga and Teunis (2004). For these last two competitors we have used the publicly available implementation offered in the `EpiEstim` package from `R`.

All competitors are evaluated with respect to the accuracy of estimating $R(t)$ which is measured by the total sum of squared errors defined as:

$$\text{SSE} = \sum_{t=1}^{T} \left(\hat{R}(t) - R(t)\right)^2$$

where $R(t)$ is the true value of the reproduction number and $\hat{R}(t)$ is the estimated value of the reproduction number.

Table 1 shows the obtained results as median values over the 1000 different simulated epidemics and upon inspection we conclude that: (i) when there is a high degree of overdispersion (ie. $\theta$ is small) the spline-based $R(t)$ produced better SSE performance than the classical competitors, across all scenarios and regardless of the spline basis function, pointing to the conclusion that for such a setting smoothing helps identifying the underlying trend more accurately; (ii) for low to medium overdispersion the results are closer, but also in this case splines are slightly better, especially for Scenario IV. For visualization purposes, in Figure 2 the estimated $R(t)$ values for the first 100 epidemics using truncated polynomial and radial bases are plotted. The figure suggests also that as expected the obtained trajectories are much smoother for the spline-based method than for the competitors and also the fact that the competitors are much more variable relative to the spline-based estimator.

Overall, estimating $R(t)$ using flexible, spline-based models can provide substantial gains in the accuracy of estimating the transmission capability of virus.

| | Scenario I | | | Scenario II | | |
|---|---|---|---|---|---|---|
| | $\theta = 2$ | $\theta = 10$ | $\theta = 100$ | $\theta = 2$ | $\theta = 10$ | $\theta = 100$ |
| Truncated polynomial (Poisson) | 33.6 | 10.6 | 8.3 | 17.8 | 6.9 | 2.5 |
| Truncated polynomial (quasi-Poisson) | 33.6 | 10.6 | 8.3 | 17.8 | 6.9 | 2.5 |
| Truncated polynomial (Negative binomial) | 17.5 | 10.7 | 8.6 | 11.6 | 5.7 | 3.3 |
| Radial basis (Poisson) | 25.8 | 13.7 | 8.4 | 19.1 | 14.3 | 5.5 |
| Radial basis (quasi-Poisson) | 25.8 | 13.7 | 8.4 | 19.1 | 14.3 | 5.5 |
| Radial basis (Negative binomial) | 18.1 | 10.5 | 8.2 | 10.9 | 5.3 | 3.0 |
| Cori et al. | 31.7 | 10.5 | 5.8 | 28.1 | 7.5 | 2.8 |
| Wallinga & Teunis | 36.2 | 23.3 | 20.5 | 20.5 | 8.0 | 5.2 |
| | Scenario III | | | Scenario IV | | |
| | $\theta = 2$ | $\theta = 10$ | $\theta = 100$ | $\theta = 2$ | $\theta = 10$ | $\theta = 100$ |
| Truncated polynomial (Poisson) | 13.3 | 5.1 | 3.6 | 3.2 | 1.7 | 1.5 |
| Truncated polynomial (quasi-Poisson) | 13.3 | 5.1 | 3.6 | 3.2 | 1.7 | 1.5 |
| Truncated polynomial (Negative binomial) | 9.9 | 5.8 | 4.4 | 2.6 | 1.3 | 1.3 |
| Radial basis (Poisson) | 9.1 | 4.3 | 3.4 | 1.7 | 0.7 | 0.4 |
| Radial basis (quasi-Poisson) | 9.1 | 4.3 | 3.4 | 1.7 | 0.7 | 0.4 |
| Radial basis (Negative binomial) | 10.6 | 6.0 | 4.4 | 1.6 | 0.6 | 0.4 |
| Cori et al. | 24.5 | 7.4 | 3.5 | 11.4 | 3.7 | 1.9 |
| Wallinga & Teunis | 17.8 | 10.4 | 8.8 | 12.9 | 2.6 | 1.2 |

Table 1: Median SSE over 1000 different simulated epidemics from four different scenarios using three different values for $\theta$. $R(t)$ is estimated based on truncated polynomial or radial bases with Poisson, negative binomial and quasi-Poisson assumptions. The performance of the estimators of Cori et al. (2013) and Wallinga and Teunis (2004) is also presented.

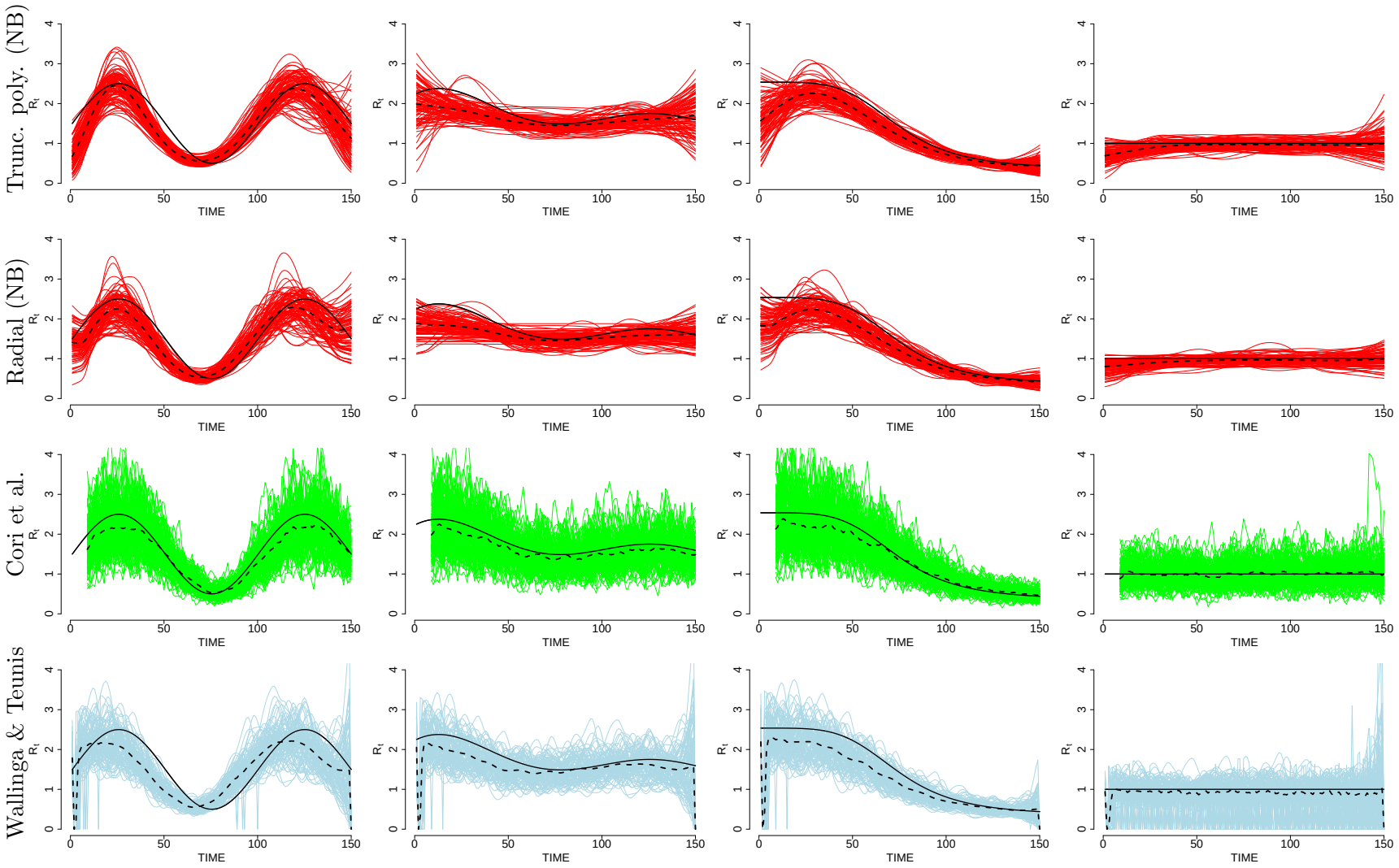


Figure 2: Spline-based estimated $R(t)$ using truncated polynomial or radial bases and a negative binomial (NB) distributional assumption. The estimates for 100 different simulated epidemics are shown in each panel. The solid black line depicts the true $R(t)$ function, the solid colored lines depict the 100 different estimates, while the dotted lines depict the empirical averages over the 100 estimates at each time point. The estimates based on the methods of Cori et al. (2013) and Wallinga and Teunis (2004) are also presented.

# 6 Real examples

We have estimated next the time dependent reproduction number $R(t)$ for six countries (US, Brazil, Italy, UK, Belgium and France) using the proposed spline-based method (with truncated polynomial bases and a negative binomial distributional assumption) and the estimator of Cori et al. (2013). We have used the data on daily infections that are available on the official public health platforms for each respective country, and to eliminate weekend effects and delays in reporting we have taken 7-day moving averages for each time series as the final input data. For stability purposes, the reproduction numbers are estimated from the moment in time when at least 10 new infected cases are confirmed in the population. Not all series start at the same moment in time, due to country reporting issues, however this does not pose any major problems.

Figure 3 presents the obtained results. In general, the two estimators agree quite well with respect to the estimated trends (albeit a slight lag between the two) starting from around beginning of April, 2020 for all six countries. In the first part of 2020, there is a larger degree of differences between the two estimators either with respect to the magnitude of the initial estimates or with respect to the estimated initial evolution, as is the case for France. In general, the estimator of Cori et al. (2013) suggests much higher initial estimates than what the spline-based model proposes and due to the fact that it needs a window of time (set to 7 days for this application) it is unable to capture the trend from the start, whereas our estimator is equipped to estimate reproduction numbers in the beginning part of the epidemic (although with larger uncertainty). As such the spline model suggests that for example, the US, the UK and Belgium are in the beginning part in March, 2020 on an upwards path for about 2 weeks, after which the reproduction number starts decreasing. Given that explicit measures were taken in that period (on March, 13th national emergency is declared in the US together with travel bans [1], on March, 18th Belgium goes into lock-down [2], and on March, 23rd the UK issues 'stay at home' orders [3]) such a trajectory seems plausible.

France presents a very interesting phenomenon, as the official, consolidated series starts around mid May, 2020. The proposed spline-based estimator suggests that the epidemiological situation was relatively under control at that moment, whereas the competitor suggests the reproduction number is higher than 3. Given that on May, 11th the Security council in France started with relaxing lock-down measures [4] and as it was reported that all epidemiological indicators were since two weeks before on a downward slope [5], we are more inclined to trust the spline estimator as a more accurate description of the reality in France at that moment in time.

## 6.1 Accounting for interventions

One advantage of the spline-based $R(t)$ measure is the fact that, as discussed briefly in Section 2, one can easily consider models of the form $g\{\mu(t)\} = f_1(t) + f_2(X_t)$. As such, we propose next a semi-parametric extension, where the $R(t)$ estimates are adjusted for interventional effects. As in Fokianos and Fried (2010) and Liboschik et al. (2016) we consider interventional covariates of the form:

$$X_t = \delta^{(t-\tau)}\mathbf{1}_{\{t\geq\tau\}}$$

[1] https://www.nytimes.com/2020/03/13/us/politics/trump-coronavirus-news-conference.html
[2] https://www.belgium.be/en/news/2020/coronavirus_reinforced_measures
[3] https://www.gov.uk/government/speeches/pm-statement-at-coronavirus-press-conference-23-march-2021
[4] https://www.gouvernement.fr/info-coronavirus/les-actions-du-gouvernement
[5] https://www.santepubliquefrance.fr/maladies-et-traumatismes/maladies-et-infections-respiratoires/infection-a-coronavirus/documents/bulletin-national/covid-19-point-epidemiologique-du-14-mai-2020

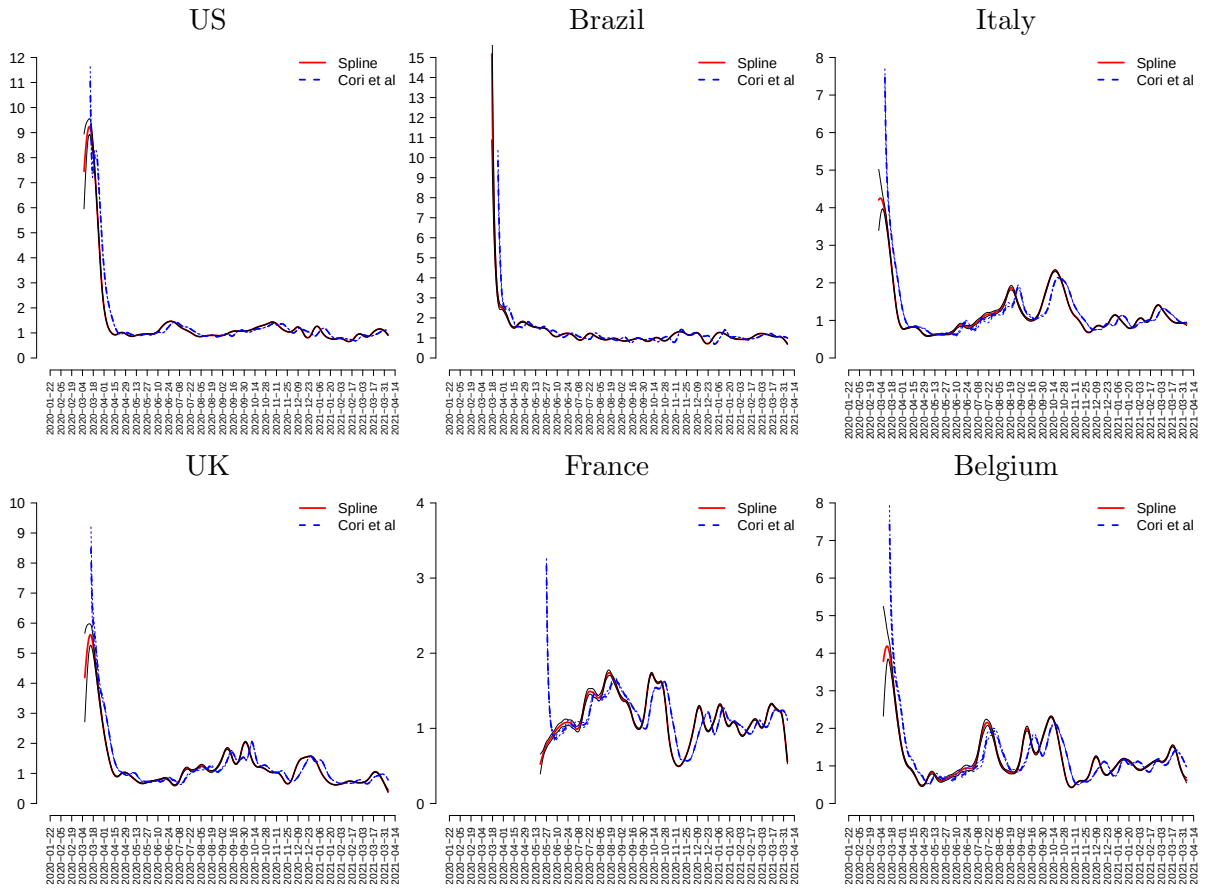


Figure 3: Spline-based estimated $R(t)$ (and 95% pointwise confidence interval) using truncated polynomial bases and a negative binomial assumption for six countries. The estimates based on the method of Cori et al. (2013) are also presented.

where $\tau$ represents the intervention time point, $\mathbf{1}_{\{t \geq \tau\}}$ is the indicator function taking the value 1 if $t \geq \tau$ and 0 otherwise. Moreover, $\delta \in [0, 1]$ specifies the intervention type with $\delta = 0$ denoting an intervention that has an effect only at the time of its occurrence, $\delta = 1$ denoting a persistent effect of the intervention after its occurrence, while $\delta \in (0, 1)$ denotes an exponentially decaying effect.

To illustrate the effect of accounting for interventions, we chose the case of Belgium (due to familiarity with the epidemiological situation) where we have introduced six exponentially decaying interventional effects corresponding to two national lock-downs, one regional restrictive measure for Antwerp and the reopening of the academic year for secondary and higher educational level. Figure 4 shows the estimated reproduction numbers across time and it suggests that (i) accounting for the first lock-down in March, 2020 had the largest impact in modifying the estimates in the beginning of the epidemic. As well, the reopening of schools on September, 1st (coinciding with the end of the vacation period and returning to work for a large proportion of the active population), showed also an effect on impacting the estimates, suggesting a slight shift of reproduction numbers for the month of September, 2020 than what the model without interventional effects would suggest. As expected, these interventional effects dissipate some time after their introduction as the estimated trends in the later part of the epidemic (ie. after November, 2020) are very close together.

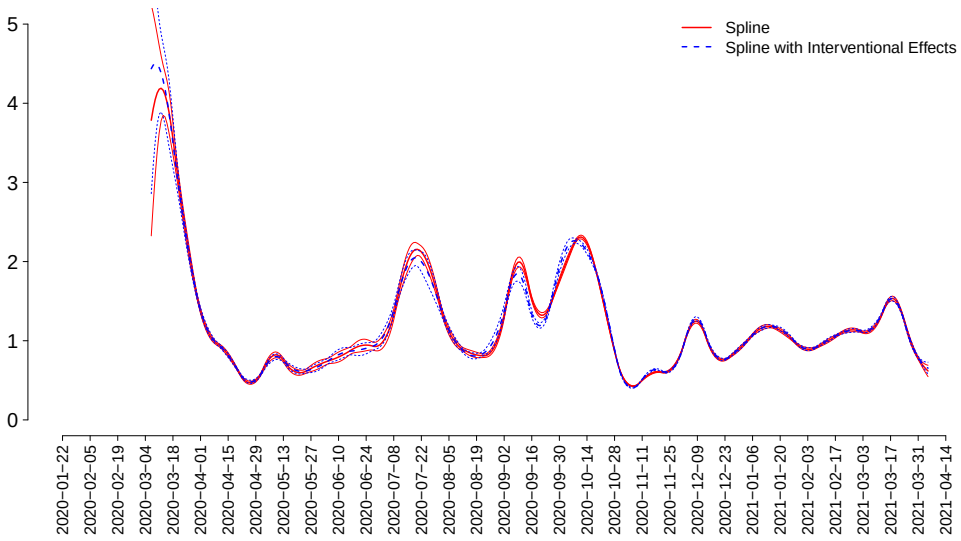


Figure 4: Spline-based estimated $R(t)$ with 95% pointwise confidence interval for Belgium. The estimator accounting for the interventions (dashed line) is presented alongside the estimator that does not account for interventions (full line).

# 7 Discussion

We proposed in this manuscript a time-dependent version of the reproduction number $R(t)$ that is based on a spline approach. The model starts from the representation of the basic reproduction number of Wallinga and Lipsitch (2007) and proposes to obtain a growth rate parameter by inspecting the derivative of a smooth function of time. As such, the model is an extension that allows for an estimation of an instantaneous reproduction number of a virus at any moment in time during the evolution of the epidemic.

On simulated data the proposed plug-in estimator shows good performance relative to other classical estimators that are used in the literature, capable of adapting to sharp fluctuations in the evolution of the epidemic. As well, accounting for overdispersion in the data can easily be done by using a quasi-Poisson or a negative binomial model as our basic formulation is general enough to allow for it. On real data related to the SARS-CoV-2 pandemic, the proposed estimator seemed also to provide more realistic evolutions for the first part of the pandemic, while being close in estimated trends to the classical estimators, for the second part of the pandemic.

# References


Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Cori, A., Ferguson, N. M., Fraser, C., and Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology*, 178(9):1505–1512.

de Boor, C. (2001). *A Practical Guide to Splines*. Springer.

Dietz, K. (1993). The estimation of the basic reproduction number for infectious diseases. *Statistical Methods in Medical Research*, 2(1):23–41.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121 (with comments and a rejoinder by the authors).

Fokianos, K. and Fried, R. (2010). Interventions in INGARCH processes. *Journal of Time Series Analysis*, 31(3):210–225.

Hens, N., Van Ranst, M., Aerts, M., Robesyn, E., Van Damme, P., and Beutels, P. (2011). Estimating the effective reproduction number for pandemic influenza from notification data made publicly available in real time: A multi-country analysis for influenza A/H1N1v 2009. *Vaccine*, 29(5):896 – 904.

Hong, H. and Li, Y. (2020). Estimation of timevarying reproduction numbers underlying epidemiological processes: A new statistical tool for the COVID-19 pandemic. *PLoS ONE*, 15(7):e0236464.

Koyama, S., Horie, T., and Shinomoto, S. (2021). Estimating the time-varying reproduction number of COVID-19 with a state-space method. *PLoS Computational Biology*, 17(1):e1008679.

Liboschik, T., Kerschke, P., Fokianos, K., and Fried, R. (2016). Modelling interventions in INGARCH processes. *International Journal of Computer Mathematics*, 93(4):640–657.

Liu, Y., Gayle, A. A., Wilder-Smith, A., and Rocklöv, J. (2020). The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of Travel Medicine*, 27(2).

Obadia, T., Haneef, R., and Boëlle, P. (2012). The R0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks. *BMC Medical Informatics and Decision Making*, 12.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric regression*. Cambridge University Press.

Salje, H., Tran Kiem, C., Lefrancq, N., Courtejoie, N., Bosetti, P., Paireau, J., Andronico, A., and Hoze, N. Richet, J. D. C.-L. (2020). Estimating the burden of SARS-CoV-2 in France. Technical report: ffpasteur-02548181f.

Searle, S., Casella, G., and McCulloch, C. (2006). *Variance Components*. John Wiley & Sons.

Stroup, W. (2013). *Generalized Linear Mixed Models. Modern Concepts, Methods and Applications*. Chapman & Hall.

Tsang, T. K., Wu, P., Lin, Y., Lau, E. H. Y., Leung, G. M., and Cowling, B. J. (2020). Effect of changing case definitions for COVID-19 on the epidemic curve and transmission parameters in mainland China: a modelling study. *The Lancet Public Health*, 5(5):e289 – e296.

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

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(6):509–516.