Estimating COVID-19 Reproduction Numbers

Kevin Durant

June 2020

Abstract

This is a short description of an analytic, *almost* fully Bayesian solution to the problem of inferring effective reproduction numbers for an epidemic from reported infection counts. Under the assumptions of a negative binomial likelihood function and a simple beta prime distributed predictive prior, the posterior distribution of the daily infection rate follows a beta prime distribution whose parameters can be expressed in closed form. These posteriors can in turn be used to estimate effective reproduction numbers.

1 Introduction

The effective reproduction number of an epidemic, R_t, refers to the expected number of new infections caused by a single infected individual at a given stage of the epidemic's course. Along with other, complementary measures, these numbers can help give an indication of the rate at which the epidemic is spreading.

The method for estimating effective reproduction numbers discussed here is a slightly different take on one described elsewhere by Kevin Systrom [3], which is itself based on work by Bettencourt and Ribeiro [1]. Unlike those two approaches, the solution described here is almost entirely analytic—the only numerical computation required being that of percentiles of a beta prime distribution. All three methods are based on a key point made by the latter authors mentioned above: that under the assumptions of a standard epidemic susceptible-infected (SIR) model, the effective reproduction number of a virus at time t can be estimated from the number of new cases recorded between times $t-1$ and t, and t and $t+1$.

More specifically, let R_t be the effective reproduction number at time t, and assume that this number remains constant over the interval $(t-1, t]$

(most likely a single day)[1]. Likewise, let λ_t be the average number of new infections that occur during this interval—i.e., the current rate of infection. One then has the following approximation [1]:

$$\lambda_t \approx \lambda_{t-1} \exp(\gamma(R_t - 1)),$$

in which γ is the reciprocal of the infectious period of the virus. Equivalently,

$$R_t \approx \frac{1}{\gamma} \log\left(\frac{\lambda_t}{\lambda_{t-1}}\right) + 1. \tag{1}$$

The approach outlined here involves modelling the number of observed infections k_t as a stochastic process, assuming that k_t depends on an underlying rate of infection λ_t via a negative binomial distribution (a Poisson distribution can also be used). One can then infer the infection rates λ_t analytically, and use them to estimate effective reproduction numbers by applying equation (1).

Specifically, one finds that the posterior rate of infection λ_t can be described using a beta prime distribution:

$$\mathrm{P}\left(\frac{\lambda_t}{r} \;\middle|\; k_1, \ldots, k_{t-1}\right) = \mathrm{BP}\left(\frac{\lambda_t}{r} \;\middle|\; \alpha_t, \beta_t\right),$$

in which

$$\begin{aligned} \alpha_t &= \frac{a_1}{c^{t-1}} + \sum_{i=0}^{t-1} \frac{k_{t-i}}{c^i}, \\ \beta_t &= \frac{b_1}{c^{t-1}} + \sum_{i=0}^{t-1} \frac{r}{c^i} = \frac{b_1}{c^{t-1}} + r\frac{1 - \frac{1}{c^t}}{1 - \frac{1}{c}}. \end{aligned} \tag{2}$$

Both r and c are model parameters that can easily be optimised, because their marginal likelihood $\mathrm{P}(k_1, \ldots, k_{t-1} \mid r, c)$ is available in closed form. The constants a_1 and b_1 are parameters of the initial prior on λ_1.

We apply this model to the estimation of reproduction numbers in section 2, by making use of a simplified version of equation (1) in which λ_{t-1} is replaced with a point estimate λ^*_{t-1}. We use, for example, the median value of $\mathrm{P}(\lambda_{t-1} \mid k_1, \ldots, k_{t-2})$ as such an estimate.

[1]Note that our notation differs slightly from that used by Bettencourt and Ribeiro—our R_t and λ_t correspond to their R_{t-1} and $\Delta T(t)$ respectively.

2 The stochastic process

Let $\mathbf{k} = k_1, \ldots, k_{t-1}$ be a sequence of observed infection counts, and $\boldsymbol{\lambda}$ the corresponding sequence of unknown infection rates. The primary assumption is that each k and λ are related via a negative binomial distribution:

$$\mathrm{P}(k_t \mid \lambda_t, \mathbf{k}) = \mathrm{P}(k_t \mid \lambda_t) \sim \mathrm{NB}\left(k_t \;\middle|\; r, \frac{\lambda_t}{r + \lambda_t} = p_t\right), \tag{3}$$

where r is an unknown dispersion parameter and p_t is a reparameterisation of λ_t as a 'success' probability. Parameterised in this way, the negative binomial distribution converges to a Poisson distribution of rate λ_t as $r \to \infty$, and by adjusting r we can control the level of variance inherent to the distribution (smaller values of r result in higher variance).

Inference of the rate sequence $\boldsymbol{\lambda}$ is performed iteratively, by repeated application of Bayes' rule:

$$\mathrm{P}(\lambda_t \mid k_t, \mathbf{k}) \propto \mathrm{P}(k_t \mid \lambda_t)\, \mathrm{P}(\lambda_t \mid \mathbf{k}).$$

The first term on the right-hand side—the likelihood function—is simply the negative binomial distribution given above. The second term is a predictive prior on λ_t given only the *past* infection counts $\mathbf{k}$. Technically the inference will be done with respect to $\mathbf{p} = p_1, \ldots, p_{t-1}$, not $\boldsymbol{\lambda}$, but with the right choice of prior the translation between the two is seamless.

The conjugate prior for the negative binomial likelihood function (with known dispersion) is the beta distribution, so our second assumption is that the prior distribution on p_t is of this form. Note that the change-of-variable formula for probability density functions implies that when p_t follows a beta distribution, λ_t / r is distributed according to a beta prime distribution with identical parameters:

$$p_t \sim \mathrm{B}(p_t \mid \alpha, \beta) \Rightarrow \frac{\lambda_t}{r} \sim \mathrm{BP}\left(\frac{\lambda_t}{r} \;\middle|\; \alpha, \beta\right).$$

As mentioned above, this allows us to work with p_t instead of λ_t while deriving posteriors and marginal likelihoods, but still consider λ_t when computing R_t.

The third and final assumption we make involves the way in which the predictive prior $\mathrm{P}(p_t \mid \mathbf{k})$ is derived from the previous posterior $\mathrm{P}(p_{t-1} \mid \mathbf{k})$. In the case of a Gaussian stochastic process, one would derive the prior by assuming additive Gaussian noise on the previous latent variable, resulting in a distribution that has the same mean as the previous posterior, but higher variance. Doing so involves solving an integral of the form

$$\mathrm{P}(p_t \mid \mathbf{k}) = \int \mathrm{P}(p_t \mid p_{t-1})\, \mathrm{P}(p_{t-1} \mid \mathbf{k})\, dp_{t-1},$$

which is tractable in the Gaussian case.

Although the situation is not quite as straightforward here, we can achieve a similar outcome by simply assuming the relationship to the previous posterior directly: specifically, if

$$\mathrm{P}(p_{t-1} \mid \mathbf{k}) \sim \mathrm{B}(p_{t-1} \mid \alpha_{t-1}, \beta_{t-1}), \tag{4}$$

we might assume a predictive prior of the form

$$\mathrm{P}(p_t \mid \mathbf{k}) \sim \mathrm{B}(p_t \mid \alpha_{t-1}/c, \beta_{t-1}/c) = \mathrm{B}(p_t \mid a_t, b_t). \tag{5}$$

This is a straightforward prior that has the same mean as the predictive posterior on p_{t-1}, but a variance that is larger *roughly* by a factor c—since the mean and variance of a beta distribution with parameters α and β are given by

$$\mathrm{E}[X] = \frac{\alpha}{\alpha+\beta}, \quad \mathrm{V}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)},$$

and in general c will be small relative to $\alpha + \beta$.

One could just as easily make use of a predictive prior that introduces additive noise, unlike the multiplicative noise described above. The main reason for choosing multiplicative noise here is that changes in scale made to p_t (and λ_t) result in additive changes to R_t, so in this way one is effectively introducing additive noise into the overarching reproduction number process.

The remaining details of the stochastic process now follow from assumptions (3) and (5). Firstly–and most importantly—the posterior on p_t is given by

$$\begin{aligned}
\mathrm{P}(p_t \mid k_t, \mathbf{k}) &\propto \mathrm{P}(k_t \mid p_t)\,\mathrm{P}(p_t \mid \mathbf{k}) \\
&= \mathrm{NB}(k_t \mid r, p_t)\ \mathrm{B}(p_t \mid a_t, b_t) \\
&\sim p_t^{k_t}(1-p_t)^r \cdot p_t^{a_t-1}(1-p_t)^{b_t-1} \\
&\Rightarrow \mathrm{B}(p_t \mid a_t + k_t, b_t + r).
\end{aligned}$$

That is (extending equation (4)):

$$\mathrm{P}(p_t \mid k_t, \mathbf{k}) \sim \mathrm{B}(p_t \mid a_t + k_t, b_t + r) = \mathrm{B}(p_t \mid \alpha_t, \beta_t). \tag{6}$$

Combining this with equation (5) one can solve for α_t and β_t recursively, leading to equation (2). Note however that these solutions presume that the time series of data points is complete; in the presence of a missing datum k_t one will need to specify α_t and β_t explicitly—for example simply by setting them to a_t and b_t respectively.

Secondly, we can derive the marginal likelihood of observation k_t given the previous observations:

$$\begin{aligned}
\mathrm{P}(k_t \mid \mathbf{k}) &= \int \mathrm{P}(k_t, p_t \mid \mathbf{k})\, dp_t \\
&= \int \mathrm{NB}(k_t \mid r, p_t)\ \mathrm{B}(p_t \mid a_t, b_t)\, dp_t \\
&= \binom{k_t + r - 1}{k_t} \frac{1}{\mathrm{B}(a_t, b_t)} \int p_t^{a_t + k_t - 1} (1 - p_t)^{b_t + r - 1}\, dp_t \\
&= \begin{cases} \dfrac{\mathrm{B}(a_t, b_t + r)}{\mathrm{B}(a_t, b_t)} & \text{if } k_t = 0, \\ \dfrac{1}{k_t \mathrm{B}(k_t, r)} \dfrac{\mathrm{B}(a_t + k_t, b_t + r)}{\mathrm{B}(a_t, b_t)} & \text{if } k_t > 0, \end{cases}
\end{aligned} \tag{7}$$

where $\mathrm{B}(x, y)$ denotes the beta function. (The final expression can also be rephrased in terms of α_t and β_t using equation (6).) This allows us to compute the overall marginal likelihood iteratively, since

$$\mathrm{P}(\mathbf{k}) = \prod_{i=1}^{t-1} \mathrm{P}(k_i \mid k_1, \ldots, k_{i-1}).$$

The overall marginal likelihood will in turn allow us to compare the relative likelihoods of values of r and c, which are the model's two tunable parameters.

Equations (5) and (6) allow us to derive, via p_t, a posterior distribution for each λ_t. Ideally one would hope for a distribution on $\lambda_t / \lambda_{t-1}$, since this is the ratio on which R_t depends (equation (1)), however this would require one to either specify $\mathrm{P}(p_t \mid p_{t-1})$—something we explicitly avoiding doing above to keep things tractable—or treat λ_t and λ_{t-1} as independent for the purposes of deriving R_t.

Here we adopt a simpler approach: replace λ_{t-1} in equation (1) with a point estimate λ^*_{t-1}, resulting in the approximation

$$R_t \approx \frac{1}{\gamma} \log\left(\frac{\lambda_t}{\lambda_{t-1}}\right) + 1 \approx \frac{1}{\gamma} \log\left(\frac{\lambda_t}{r} \frac{r}{\lambda^*_{t-1}}\right) + 1, \tag{8}$$

to which our inferred posterior on λ_t / r can directly be applied. This is not unlike a simplification made by the other authors [1, 3], who set $\lambda^*_{t-1} = k_{t-1}$. In what follows, we have set λ^*_{t-1} to the median of the posterior distribution $\mathrm{P}(\lambda_{t-1} \mid k_1, \ldots k_{t-1})$ (which is simply r times the median of the posterior on λ_{t-1}/r).

Before moving on, we note that Bettencourt and Ribeiro also describe a negative binomial process in which r is not constant, but rather $r_t = k_{t-1}$. We

have briefly tested this approach, and in practice it performs similarly to the one we have described above; the main difference being that the precision of the likelihood function (and thus posterior) varies with the observed counts $\mathbf{k}$. The approach outlined in this section can still be applied, and the resulting analytic solution differs only slightly.

An application to South African data

In this section we apply the steps described above to data stemming from the COVID-19 epidemic in South Africa. The data set used here is obtained from South Africa's National Institute for Communicable Diseases, via the University of Pretoria [2]. The data take the form of daily cumulative infection counts per province, from which we derive new infection counts per day.

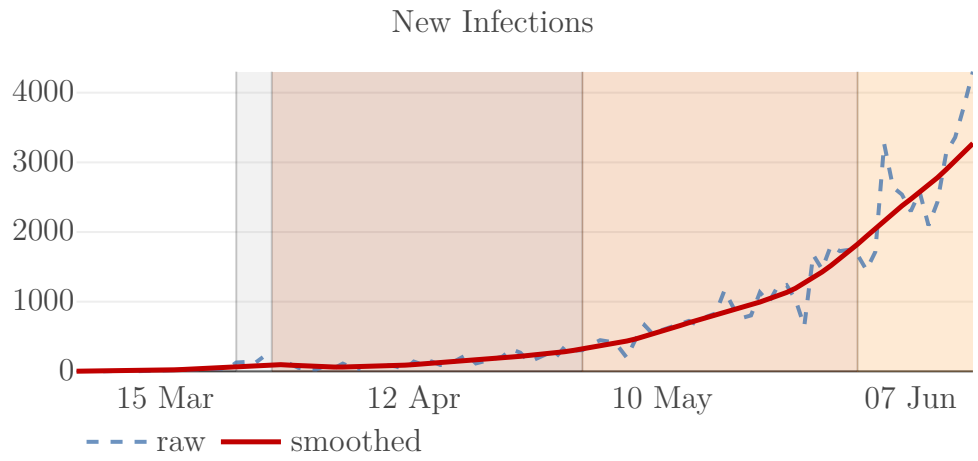

Figure 1: Daily new infection counts for the COVID-19 epidemic in South Africa, smoothed using a Gaussian window with a standard deviation of $\sigma = 3.5$. Shaded areas indicate different levels of the nationwide lockdown, with the first, grey area depicting the period between the announcement of the lockdown and its initiation.

Before inferring daily infection rates, we first need to decide whether or not the raw counts should be smoothed, and if so, to what extent (see figure 1). The obvious argument against doing so is that the variance of R_t depends indirectly on the variance of the k_t, so smoothed infection counts may result in artificially precise posteriors.

On the other hand, one might argue that the underlying assumption of the model is that k_t represents the *true* number of infections on a given day, of which smoothed counts are likely a more appropriate indication. A second,

more technical argument for reducing the variance of the reported infection counts is the fact that the equation $\lambda_t \approx \lambda_{t-1} \exp(\gamma(R_t - 1))$ contains an implicit assumption: that $R_t \geq 0$, and thus $\lambda_t/\lambda_{t-1} \geq \exp(-\gamma)$.

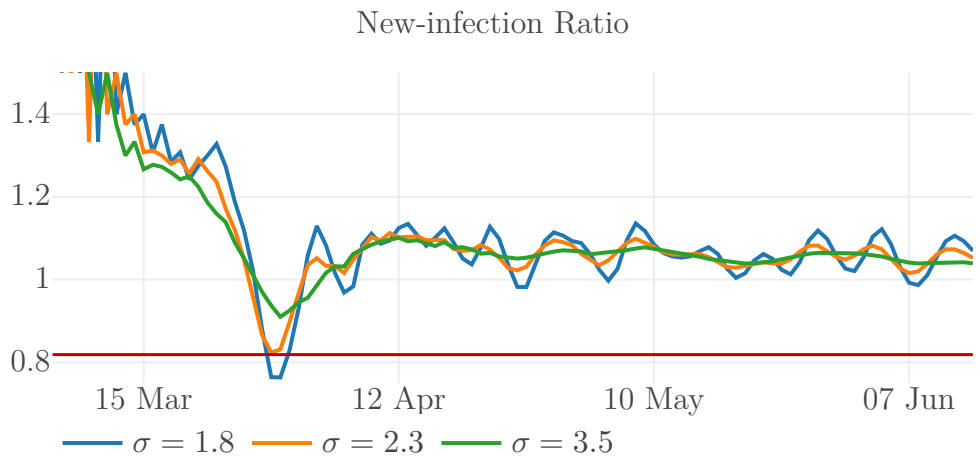

Figure 2: Ratios of consecutive infection counts k_t/k_{t-1} for various levels of Gaussian smoothing. The red line indicates the bound implied by equation (1).

In figure 2 we plot the ratio k_t/k_{t-1} for various levels of Gaussian averaging—roughly corresponding to smoothing windows of 0.5, 1, and 2 weeks (if one uses four standard deviations as a guide). One sees that in this case, windows with scales greater than $\sigma = 7/3$ appear both to respect the implicit bound mentioned above, as well as alleviate much of the periodicity visible in the raw count sequence.

For the remainder of this example we will counts that have been averaged using a Gaussian window with standard deviation $\sigma = 3.5$. This value will not necessarily be suitable for data sets from other regions or sources, however.

The rest of the application is straightforward: we apply equation (7) to select values for r and c that maximise the marginal likelihood (for the countrywide South African data this yields $r = 1592$ and $c \approx 3.56$), and then use equation 6 to compute the parameters of the posteriors on $\lambda_1, \lambda_2, \ldots$. These posteriors, shown in figure 3, track the smoothed k_t relatively closely, as one might expect.

Finally, equation (8) allows us to plot the estimated evolution of R_t over time, simply by mapping percentiles of λ_t/r to those of $R - t$. This plot is shown in figure 4 for the entire country, and figure 5 on a provincial level.

The only comment we will make on these figures is that the effect of the hard lockdown (implemented on 28 March) on the estimated reproduction

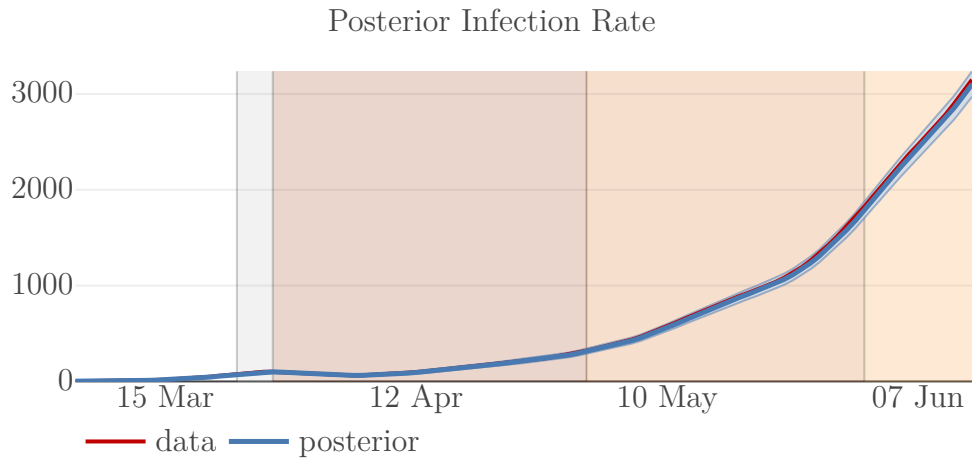

Figure 3: Daily posterior infection rates λ_t, plotted as a median and 5–95th percentile interval.

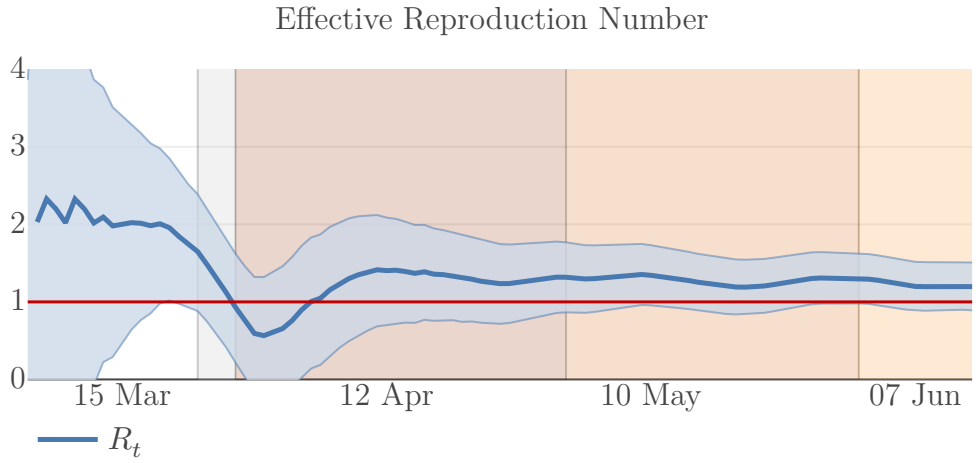

Figure 4: Estimated reproduction numbers R_t for the COVID-19 epidemic in South Africa, plotted as a median and 5–95th percentile interval.

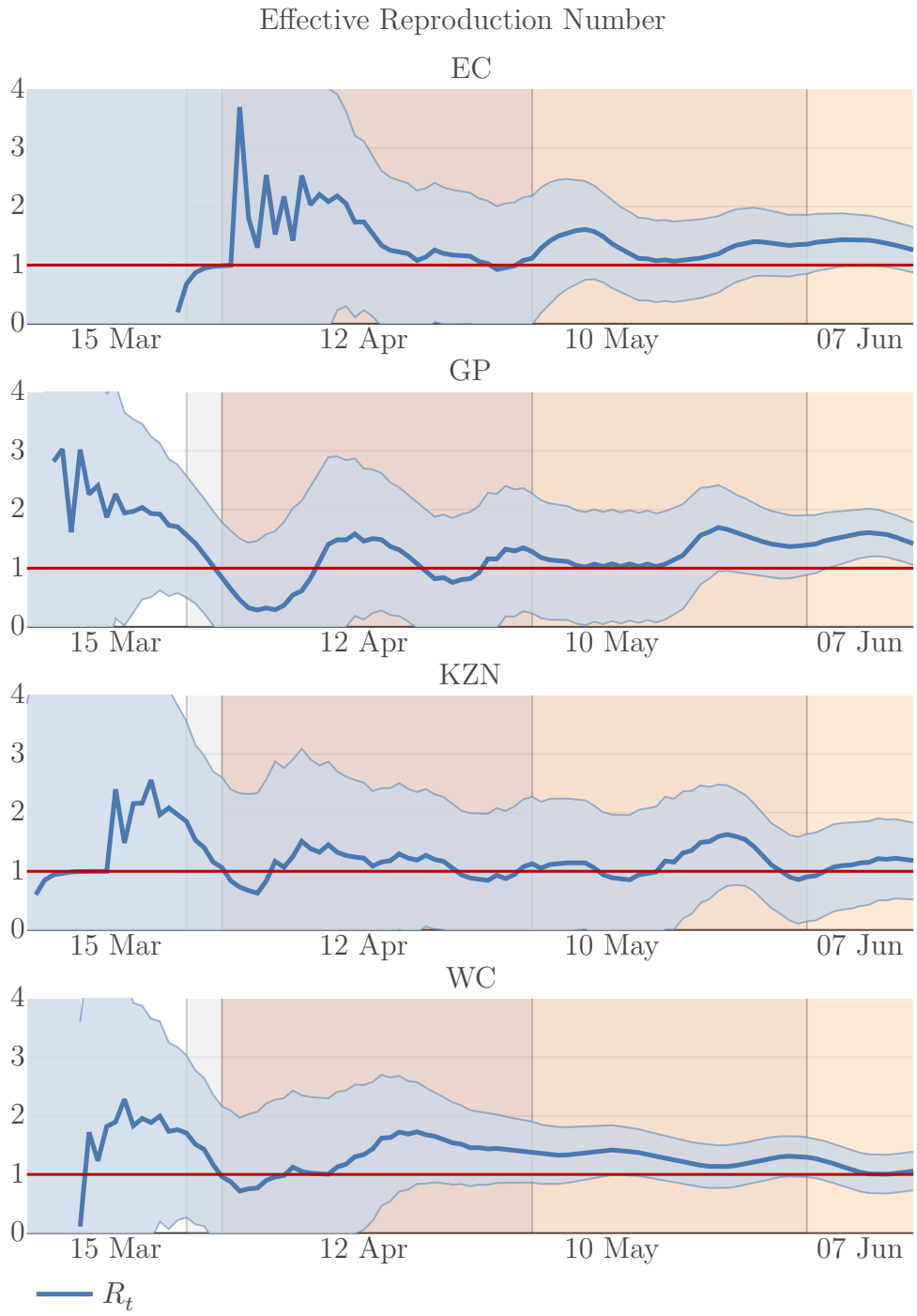

Figure 5: Reproduction numbers for the COVID-19 epidemic in South Africa, estimated at a provincial level.

numbers is clearly visible, as is—to a lesser extent—the slow relaxation of lockdown restrictions and gradual reopening of economic activity that follows.

References

[1] L. M. A. Bettencourt and Ru. M. Ribeiro. "Real Time Bayesian Estimation of the Epidemic Potential of Emerging Infectious Diseases". In: *PLOS ONE* 3.5 (May 2008), pp. 1–9.

[2] University of Pretoria DSFSI. *Coronavirus Data Repository for South Africa*. June 15, 2020. URL: `https://github.com/dsfsi/covid19za`.

[3] K. Systrom. *The Metric We Need to Manage COVID-19*. Apr. 12, 2020. URL: `http://systrom.com/blog/the-metric-we-need-to-manage-covid-19`.