INTERFACE

royalsocietypublishing.org/journal/rsif

Research

Cite this article: Harvey A, Kattuman P. 2021 A farewell to *R*: time-series models for tracking and forecasting epidemics. *J. R. Soc. Interface* **18**: 20210179.
https://doi.org/10.1098/rsif.2021.0179

Received: 2 March 2021
Accepted: 2 September 2021

Subject Category:
Life Sciences–Mathematics interface

Subject Areas:
bioinformatics

Keywords:
COVID-19, Gompertz curve, Kalman filter, state-space model, stochastic trend, waves

Author for correspondence:
Paul Kattuman
e-mail: p.kattuman@jbs.cam.ac.uk

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.5619079.

THE ROYAL SOCIETY PUBLISHING

A farewell to *R*: time-series models for tracking and forecasting epidemics

Andrew Harvey[1] and Paul Kattuman[2]

[1]Faculty of Economics, and [2]Cambridge Judge Business School, University of Cambridge, Cambridge, UK

AH, 0000-0003-3082-0440; PK, 0000-0003-1209-5169

The time-dependent reproduction number, R_t, is a key metric used by epidemiologists to assess the current state of an outbreak of an infectious disease. This quantity is usually estimated using time-series observations on new infections combined with assumptions about the distribution of the serial interval of transmissions. Bayesian methods are often used with the new cases data smoothed using a simple, but to some extent arbitrary, moving average. This paper describes a new class of time-series models, estimated by classical statistical methods, for tracking and forecasting the growth rate of new cases and deaths. Very few assumptions are needed and those that are made can be tested. Estimates of R_t, together with their standard deviations, are obtained as a by-product.

1. Introduction

The degree of infectiousness of a disease is given by the basic reproduction number, R_0, defined as the number of infections that are expected to result from a single infectious individual in a completely susceptible population. As an infection spreads, immunity starts to develop and for serious diseases, such as coronavirus disease 2019 (COVID-19), social behaviour may change endogenously, or may be modified, perhaps by the imposition of lockdown and social distancing measures. The progress of an epidemic is then usually tracked by the effective, or instantaneous, reproduction number, R_t, which is the number of people in a population who get infected by an individual at any specific time (e.g. [1–3]). Such tracking is of considerable importance for planning, but it raises the question of whether estimating R_t is to be regarded as an end in itself or as a means to an end, namely tracking and forecasting the number of new cases, hospital admissions and deaths.

Harvey & Kattuman [4]—hereafter HK—developed a class of generalized logistic (GL) time-series models for predicting future values of a variable which, when cumulated, is subject to an unknown saturation level. These models are relevant for many disciplines, but attention in HK was focused on applications for coronavirus.[1] Observations on the cumulative series are transformed to growth rates and the logarithms of these growth rates are modelled with a time trend. Allowing this trend to be time varying introduces flexibility and enables the effects of changes in policy and the environment to be tracked by filters for the level and slope. The filters are functions of current and past observations implied by the model. They can produce nowcasts of the current level of the incidence curve, together with forecasts of its future direction. Estimation is by maximum likelihood (ML) and goodness of fit can be assessed by standard statistical test procedures.

The methods used by epidemiologists to assess the current state of an infectious disease use time-series observations on new infections, together with information on the distribution of the serial interval of transmissions, sometimes called the infection profile (e.g. [5–8]). A brief description of the method in [5] can be found in appendix A. Bayesian methods are often used to combine the information on the serial interval with the observations on new cases, often smoothed by a simple, but to some extent arbitrary, moving average. These formulae

effectively link estimates of R_t to the growth rate in new cases, as do the more general formulae given in [1].

In our approach, estimates of the growth rate of new cases are produced directly by the time-series model from the raw data. The nowcasts and forecasts of R_t, together with the equivalent of Bayesian credible intervals, therefore emerge as a by-product. The underlying assumptions are clear and are subject to diagnostic tests, so estimates of R_t are implicitly validated. In contrast to R_t, which is not observed directly, the accuracy of forecasts of future observations can be assessed *ex post*, providing further testing of the effectiveness of the model.

The HK model is reviewed in §2 and in §3 it is shown how new cases growth rate estimates can be used to nowcast R_t and make short-term predictions. The implicit weights in the model-based filter are compared with the weights in the simple moving average ratio estimators used by the Robert Koch Institute, Germany (RKI). In §4, data from Germany and Florida are used to illustrate how the model is able to assess the importance of spikes in new cases and track second waves. Section 5 concludes by suggesting that tracking an epidemic by methods dependent on R_t may be neither necessary nor desirable: the focus should be on the growth rates of new cases and deaths, together with their predicted time path.

2. The dynamic Gompertz model and its implementation

The model in HK uses data on the time series of the cumulative total, Y_t, of a target series, such as confirmed cases of a disease or deaths. HK show how the theory of GL growth curves suggests observational models of the form

$$\begin{aligned} \ln y_t &= \rho \ln Y_{t-1} + \delta + \gamma t + \varepsilon_t, \\ \rho &\geq 1, \quad \gamma < 0, \quad t = 2, \ldots, T, \end{aligned} \tag{2.1}$$

where $y_t = \Delta Y_t = Y_t - Y_{t-1}$ is the daily change and ε_t is a disturbance term. The model for the Gompertz curve is obtained by setting $\rho = 1$, but subtracting $\ln Y_{t-1}$ from both sides gives a simple time trend regression for the logarithm of the growth rate of the cumulated series, that is, $\ln g_t$ where $g_t = y_t/Y_{t-1}$ or $\Delta \ln Y_t$.

Remark 2.1. The growth curve for the ascending phase of an epidemic proposed in [9] implies an observational equation of the form (2.1) with $\gamma = 0$ and with ρ a deceleration parameter in the range $0 \leq \rho \leq 1$; see also [7]. When $\rho = 1$ the cumulative total grows exponentially. The introduction of a time trend with $\gamma < 0$ gives sub-exponential growth.

Deterministic trends are too inflexible for most practical time-series modelling. A stochastic trend may be introduced into the equation for $\ln g_t$ and this extra flexibility allows ρ to be set to 1. The resulting dynamic Gompertz model is

$$\ln g_t = \delta_t + \varepsilon_t, \quad \varepsilon_t \sim NID(0, \sigma_\varepsilon^2), \quad t = 2, \ldots, T, \tag{2.2}$$

where[2]

$$\left.\begin{aligned} \delta_t &= \delta_{t-1} + \gamma_{t-1} + \eta_t, \quad \eta_t \sim NID(0, \sigma_\eta^2), \\ \text{and} \quad \gamma_t &= \gamma_{t-1} + \zeta_t, \quad \zeta_t \sim NID(0, \sigma_\zeta^2), \end{aligned}\right\} \tag{2.3}$$

and the normally distributed and serially independent irregular, level and slope disturbances, ε_t, η_t and ζ_t respectively, are mutually independent. When σ_ζ^2 is positive but $\sigma_\eta^2 = 0$, the trend, δ_t, is an integrated random walk (IRW). It is this form of the stochastic trend that turns out to be most useful for tracking an epidemic because it is the movements in slope γ_t which are crucial for that purpose. The key parameter is then the signal–noise ratio, $q = \sigma_\zeta^2/\sigma_\varepsilon^2$. A deterministic trend is obtained when q is zero. Other components, such as day of the week effects, may be included in the right-hand side of (2.2).

Stochastic trend models can be estimated using techniques based on state-space models and the Kalman filter (KF) [10,11]. Here the computations were performed using the STAMP package[3] [12]. The KF outputs the estimates of the state vector $(\delta_t, \gamma_t)'$. Estimates of the state at time t conditional on information up to and including time t are denoted $(\delta_{t|t}, \gamma_{t|t})'$ and given by the contemporaneous filter; the predictive filter, which outputs $(\delta_{t+1|t}, \gamma_{t+1|t})'$, estimates the state at time $t+1$ from the same information set. It may sometimes be useful to review past movements by the smoother, denoted $(\delta_{t|T}, \gamma_{t|T})'$, which is the estimate of the state at time t based on all T observations. Estimation of the unknown variance parameters is by ML. Tests for normality and residual serial correlation are based on the standardized innovations, that is, one-step-ahead prediction errors, $v_t = \ln g_t - \delta_{t|t-1}$, $t = 3, \ldots, T$.

Figure 1 shows data for the logarithm of the growth rate of the cumulated series of new cases of COVID-19 in England from early November 2020 until 17 February 2021.[4] The model, which includes a day of the week component and a signal–noise ratio set to $q = 0.005$, was fitted to observations up to and including 4 February 2021. The bold line indicates the smoothed estimates of the trend and, as will be shown in the next section, it is the estimates of the trend and slope at the end of the series that provide the information needed to compute the nowcasts of R_t. The dashed line shows the forecasts from 5 February 2021 onwards. As can be seen these two-week-ahead forecasts are successful in capturing the trend and day of the week movements. Recursions for making forecasts of the actual number of new cases, that is, the y_t's, are given in HK. However, the emphasis in this article is on nowcasting and forecasting of the growth rate of y_t and this only requires estimates of δ_t and γ_t.

Remark 2.2. When y_t is small, it may be better to specify its distribution, conditional on past values, as discrete. The usual choice is the negative binomial. The way in which a dynamic model may be constructed is set out in HK; software can be found in [13].

3. Tracking *R*

A simple and transparent estimator of R_t is

$$\widehat{R}_{k,\tau,t} = \frac{\sum_{j=0}^{k-1} y_{t-j}}{\sum_{j=\tau}^{k+\tau-1} y_{t-j}} = \frac{\sum_{j=0}^{k-1} y_{t-j}}{\sum_{j=0}^{k-1} y_{t-\tau-j}} = \frac{\Delta_k Y_t}{\Delta_k Y_{t-\tau}}, \tag{3.1}$$

where $\Delta_k Y_t = Y_t - Y_{t-k}$, $k = 1, 2, \ldots$. The lag of τ reflects the generation interval, which is the number of days that must elapse before an infected person can transmit the disease.

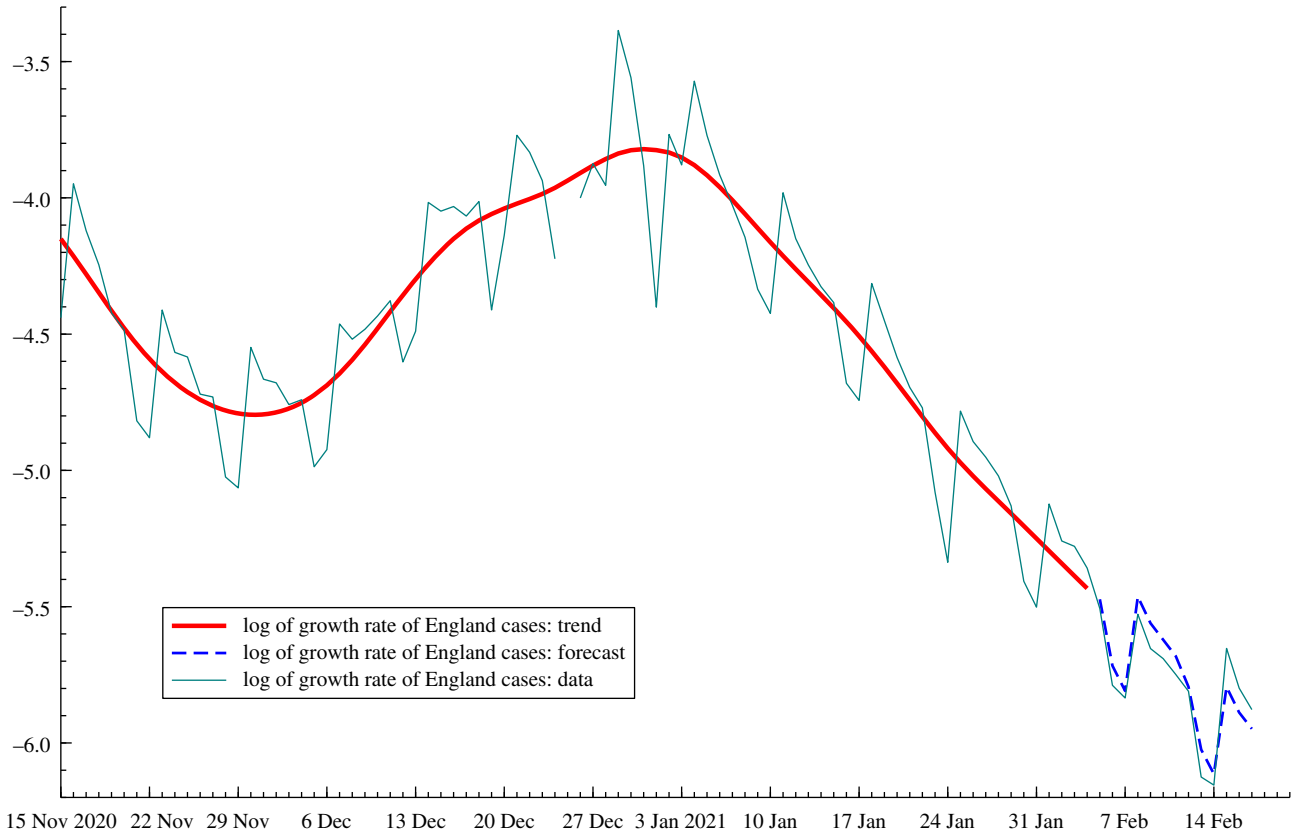

Figure 1. COVID-19 in England from early November 2020 until 17 February 2021. Trend from the fitted model is given by the bold line and the forecasts from 5 February 2021 are shown by dashes. The estimate of the trend and its slope on 4 February 2021 are $\delta_{T|T} = -5.433$ and $\gamma_{T|T} = -0.046$, respectively.

The length of the moving average, k, determines the degree of smoothing; a value of $k = 7$ has the advantage of removing the day of the week effect but at the cost of a slower response. The rationale for $\widehat{R}_{k,\tau,t}$ comes from [5]. In Germany, the national figure used by RKI is based[5] on setting $\tau = 4$ and $k = 4$ or 7, but with some prior nowcasting of the data as described in [14].

A little algebraic manipulation shows

$$\widehat{R}_{k,\tau,t} = 1 + \tau\widehat{g}_{y,t} \simeq \exp(\tau\widehat{g}_{y,t}), \tag{3.2}$$

where

$$\widehat{g}_{y,t} = \frac{1}{\tau}\frac{\Delta_k Y_t - \Delta_k Y_{t-\tau}}{\Delta_k Y_{t-\tau}}$$

is an implicit estimator of $g_{y,t}$, the growth rate in y_t, and the exponential approximation applies when $\widehat{g}_{y,t}$ is small. In a dynamic Gompertz model, the growth rate of g_t is tracked by the filtered estimates of the slope, that is, $\gamma_{t|t}$, while the growth rate itself is tracked by $g_{t|t} = \exp\delta_{t|t}$. Following the continuous time argument[6] leads to $g_{y,t}$ being estimated as

$$g_{y,t|t} = g_{t|t} + \gamma_{t|t}, \quad t = t', \ldots, T, \tag{3.3}$$

where t' is the time at which the estimates are deemed to be reasonably reliable. The nowcast of R_t suggested by equation (3.2) with $k = \tau$ is

$$\widetilde{R}_{\tau,t} = 1 + \tau g_{y,t|t} \quad \text{or} \quad \widetilde{R}^e_{\tau,t} = \exp(\tau g_{y,t|t}). \tag{3.4}$$

The RKI estimator for COVID-19 implies $\tau = 4$.

A general formula linking R_t to $g_{y,t}$ is given in [1]. When $g_{y,t}$ is estimated by (3.3), the expression is

$$\widetilde{R}^M_t = \frac{1}{M(-g_{y,t|t})}, \tag{3.5}$$

where $M(.)$ is the moment-generating function of the serial interval distribution, defined as the time between the onset of symptoms in a primary case and the onset of symptoms in secondary cases. When the distribution is degenerate, so that all secondary infections occur after exactly τ days, $\widetilde{R}^M_t = \exp(\tau g_{y,t|t})$, which is the same as $\widetilde{R}^e_{\tau,t}$. When the serial interval has a gamma distribution with parameters a and b, implying a mean of ab and a variance of ab^2, $\widetilde{R}^M_t = (1 + bg_{y,t|t})^a$. Keeping the mean constant and letting $b \to 0$ confirms that $\widetilde{R}^M_t = \exp(\tau g_{y,t|t})$, where τ is the mean generation interval. Setting $a = b = 2$, which is consistent with some of the estimates of the mean and variance obtained for COVID-19, yields

$$\widetilde{R}^M_t = (1 + 2g_{y,t|t})^2 = 1 + 4g_{y,t|t} + 4g^2_{y,t|t},$$

so, when $g_{y,t|t}$ is small, $\widetilde{R}^M_t \simeq 1 + 4g_{y,t|t} = \widetilde{R}_{4,t}$. Finally, we note the observation in [1, p. 602] that $\exp(\tau g_{y,t|t})$ is an upper bound for $\widetilde{R}^M_t$ in equation (3.5). Overall it seems that, if a single formula is to be adopted for COVID-19, $\widetilde{R}_{4,t}$ or $\widetilde{R}^e_{4,t}$ is not a bad choice.

3.1. Sampling variability of nowcasts

When q, the signal–noise ratio in the Gaussian IRW model, is treated as known, the distribution of γ_t, conditional on current and past observations, is normal with a mean $\gamma_{t|t}$ and a variance $\sigma^2_{\gamma,t|t}$ that are produced by the KF. The growth rate of the incidence curve, $g_{y,t}$, depends on g_t as well as γ_t but, as argued below, its contribution to the variability of $g_{y,t}$ is dominated by that of γ_t. When the variability in g_t is

ignored, the probability that R_t exceeds 1, that is, $\Pr(\gamma_t > -g_{t|t})$ where $g_{t|t}$ is treated as fixed, can be obtained directly from the conditional distribution of γ_t. This probability does not depend on which equation estimates R_t from $g_{y,t|t}$ and it does not depend on τ.

When $R_{\tau,t}$ is defined as $1 + \tau g_{y,t}$, its distribution, again conditional on current and past observations, is normal with mean $1 + \tau g_{y,t|t}$ and standard deviation (SD) $\tau\sigma_{\gamma,t|t}$. On the other hand, the conditional distribution of $R^e_{\tau,t}$ is lognormal with mean

$$E_t(R^e_{\tau,t}) = \exp(\tau(g_{t|t} + \gamma_{t|t} + \left(\frac{\tau}{2}\right)\sigma^2_{\gamma,t|t})) \tag{3.6}$$

and SD

$$\mathrm{SD}_t(R^e_{\tau,t}) = E_t(R^e_{\tau,t})\sqrt{(\exp\tau^2\sigma^2_{\gamma,t|t} - 1)}. \tag{3.7}$$

Note that $\exp\tau^2\sigma^2_{\gamma,t|t} - 1 \simeq \tau^2\sigma^2_{\gamma,t|t}$, so, when $E_t(R_{\tau,t})$ is close to 1, $\mathrm{SD}_t(R^e_{\tau,t})$ will be very close to the $\mathrm{SD}_t(R_{\tau,t})$.

Why is the variability in $g_{t|t}$ ignored? From equation (2.2), $g_t = \exp\delta_t$ and, because δ_t is normal, g_t is lognormal with mean $\mu_{g,t|t} = \exp(\delta_{t|t} + 0.5\sigma^2_{\delta,t|t})$ and variance $\mathrm{Var}(g_t) = \mu^2_{g,t|t}$ $(\exp\sigma^2_{\delta,t|t} - 1)$, where $\sigma^2_{\delta,t|t}$ is the variance of δ_t. However, $\sigma^2_{\delta,t|t}$ is typically small so $\mu_{g,t|t} \simeq \exp\delta_{t|t} = g_{t|t}$ and $\mathrm{Var}(g_t) \simeq \mu^2_{g,t|t}\sigma^2_{\delta,t|t} \simeq g^2_{t|t}\sigma^2_{\delta,t|t}$. Now

$$\mathrm{Var}(g_{y,t}) = \mathrm{Var}(g_t) + \mathrm{Var}(\gamma_t) + 2\mathrm{Cov}(g_t, \gamma_t),$$

but, although $\sigma^2_{\delta,t|t}$ is usually larger than $\sigma^2_{\gamma,t|t}$, the former is multiplied by $g^2_{t|t}$ to get $\mathrm{Var}(g_t)$, whereas $\mathrm{Var}(\gamma_t) = \sigma^2_{\gamma,t|t}$; note that $\sigma^2_{\delta,t|t}$ itself does not depend on the value of $g_{t|t}$. Although $g_{t|t}$ can be high near the beginning of an epidemic, it tends to fall quite rapidly and once the epidemic is underway it rarely exceeds 0.05. The example of Florida, where the second wave increases $g_{t|t}$, shows that, even in this case, $\mathrm{Var}(g_t)$ remains negligible compared with $\mathrm{Var}(\gamma_t)$.

3.2. Predictions of *R*

Predictions of R_t in the dynamic Gompertz model can be made from predictions of $g_{y,t}$, that is,

$$\begin{aligned} g_{y,T+\ell|T} &= \exp\delta_{T+\ell|T} + \gamma_{T+\ell|T} \\ &= \exp(\delta_{T|T} + \gamma_{T|T}\ell) + \gamma_{T|T}, \\ \ell &= 1, 2, \ldots, \end{aligned} \tag{3.8}$$

where ℓ is the number of steps ahead. When $g_{y,T|T}$ is positive, so any estimate of R_T given by (3.5) is greater than 1, there is still a saturation level for Y_t so long as $\gamma_{T|T}$ is negative; for example as $T \to \infty$, $\widetilde{R}^e_{\tau,T+\ell|T} \to \exp(\tau\gamma_{T|T})$. When $\gamma_{T|T}$ is zero, the growth of y_t is exponential and in this case it is helpful to characterize it by the doubling time, $\ln 2/g_{y,T|T} = 0.693\ \exp(-\delta_{T|T})$. When $\gamma_{T|T}$ is positive, as can happen at the start of a new wave, predictions of $g_{y,t}$ should not be made from (3.8). However, it may still be useful to quote the doubling time based on $g_{y,T|T}$.

If, as in the previous sub-section, it can be assumed that g_t is relatively small, the predictive distribution of $g_{y,T+\ell}$, and hence of $R_{T+\ell}$, is available because the conditional distribution of $\gamma_{T+\ell}$ given observations up to and including time T is Gaussian with mean $\gamma_{T+\ell|T} = \gamma_{T|T}$ and variance $\sigma^2_{T+\ell|T}$, as produced by the predictive equations of the KF.

Remark 3.1. The ability to make predictions offers insight into how to deal with reporting delay, as described in [15, pp. 3–4]. If the observation at time t actually relates to an event ℓ days earlier, the current R_t is better estimated by an ℓ-step-ahead forecast. When $\gamma_{T|T}$ is negative, this forecast will be less than the nowcast.

3.3. Weights

The filtered estimates of g_t and γ_t in the dynamic Gompertz model, equation (2.2), are obtained by discounting past observations, with the rate of discounting depending on the signal–noise ratio, q. Weights implied by the KF and smoother for estimated states in a linear model can be obtained as output from the STAMP package, using a method described in [16]. The forcing variable in the filter is $\ln g_t$ and the weights assigned to it in the contemporaneous filter are the weights for $\gamma_{t|t}$ plus the weights for $g_{t|t}$. When Y_t is much larger than y_t, as will be the case when an epidemic has been underway for some time, $g_{t|t}$ will be relatively small and attention can be focused on $\gamma_{t|t}$. Then, if the weights for the slope, $\gamma_{t|t}$, are denoted w_j, $j = 0, 1, 2, \ldots$,

$$g_{y,t|t} \simeq \gamma_{t|t} = \sum_{j=0}^{t-2} w_j(\ln y_{t-j} - \ln Y_{t-j-1}) \simeq \sum_{j=0}^{t-2} w_j \ln y_{t-j}, \tag{3.9}$$

where the last approximation follows because $\ln Y_{t-j-1}$ is assumed to be changing very slowly and $\sum_{j=0}^{t-2} w_j = 0$. When multiplied by τ, the weights in equation (3.9) feed directly into the estimators of R_t implied by equation (3.4). In particular

$$\widehat{R}^e_{\tau,t} = \prod_{j=0}^{t-2} y_{t-j}^{\tau w_j} = \frac{\prod_{w_j \geq 0} y_{t-j}^{\tau w_j}}{\prod_{w_j < 0} y_{t-j}^{\tau w_j}}, \tag{3.10}$$

which is similar in form to (3.1) but with summations replaced by products.

Figure 2 shows the weights for the slope produced when q is 0.001 and 0.01. ML estimates of q are typically between these values. Setting $q = 0.005$, which has the first four weights positive and the next 17 negative, is a reasonable default. A higher value gives a faster response, which may be appropriate when there is a sharp change in the environment, perhaps because of a change in policy. However, it comes at the cost of making nowcasts and forecasts less stable.

3.4. The early phase of an epidemic

Any modelling is very difficult at the start of an epidemic because of the lack of data; see, for example, the remarks in appendix 3 of [5]. In the dynamic trend model initializing the KF in the absence of prior information is generally done with a non-informative (diffuse) prior on the level and slope, as in the STAMP package; alternatively they can be estimated as unknown parameters. However, in the early part of an epidemic the growth is exponential or very close to it; see, for example, the analysis of the 1918 outbreak of Spanish flu in [17]. Thus, we could set $\gamma_0 = 0$. The filter for δ_t, the level of $\ln g_t$, can have a prior distribution informed by knowledge about the basic reproduction number, R_0. A rough estimate of $\ln g_0$ is then given from $\widehat{g}_0 = (1/\widehat{\tau})\ln\widehat{R}_0$ or $\widehat{g}_0 = (\widehat{R}_0 - 1)/\widehat{\tau}$. Choosing a suitable variance for δ_0 is more problematic.

4. Waves and spikes

After an epidemic has peaked, daily cases start to fall and the concern shifts to the possibility of a second wave and the

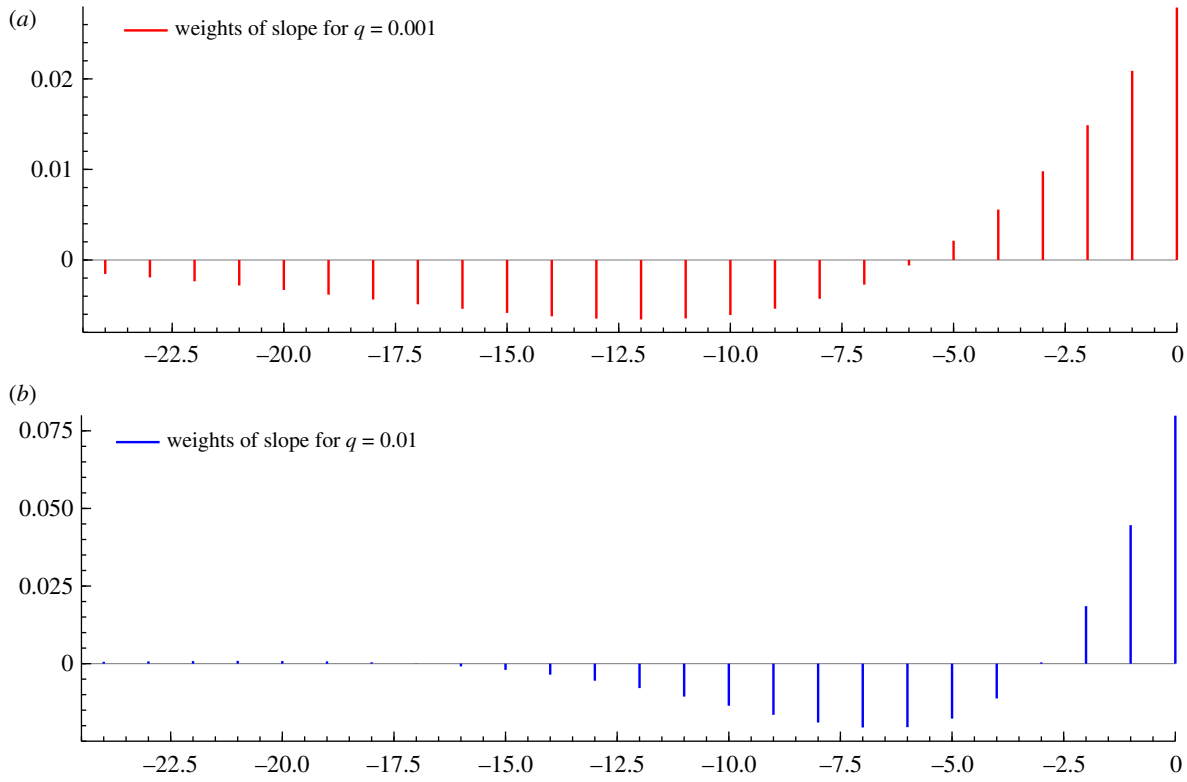

Figure 2. Weights assigned to $\ln g_t$ by the filter for the slope with (*a*) $q = 0.001$ and (*b*) $q = 0.01$.

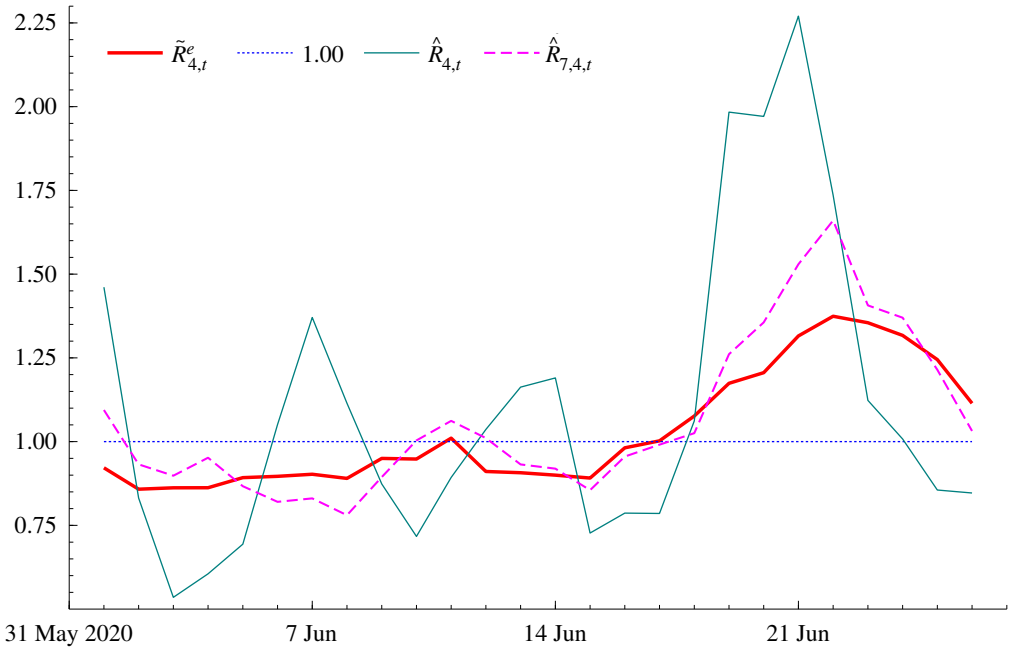

Figure 3. R_t for German new cases in June 2020.

need to deal with outbreaks indicated by spikes in the data so that they do not morph into waves. The monitoring of waves and spikes raises different issues, primarily because a wave applies to a whole nation or a relatively large geographical unit, whereas a spike is localized.

4.1. Spikes

When national numbers are low, a localized outbreak can also result in a jump in the national estimate of R_t. However, such a jump does not indicate that there has been a sudden change in the way the infection spreads and so has few implications for overall policy. Figures for new cases in Germany show a sharp increase towards the end of June 2020, caused by an outbreak at a meat-processing factory in the Gütersloh area in Westphalia. Estimates produced by RKI at the time showed a big increase in R_t, accompanied by what seems to us to be a rather narrow credible interval. Figure 3 compares the model-based reproduction number estimate, $\tilde{R}^e_{4,t}$, with the 4-day and overlapping 7-day moving average estimates, $\hat{R}_{4,t}$ and $\hat{R}_{7,4,t}$. The $\hat{R}_{4,t}$ estimates are very erratic and seriously affected by the failure to take account of the daily pattern. Estimates for Sundays and Mondays are typically lower. The peak in $\hat{R}_{4,t}$ has observations for Wednesday to Saturday in the

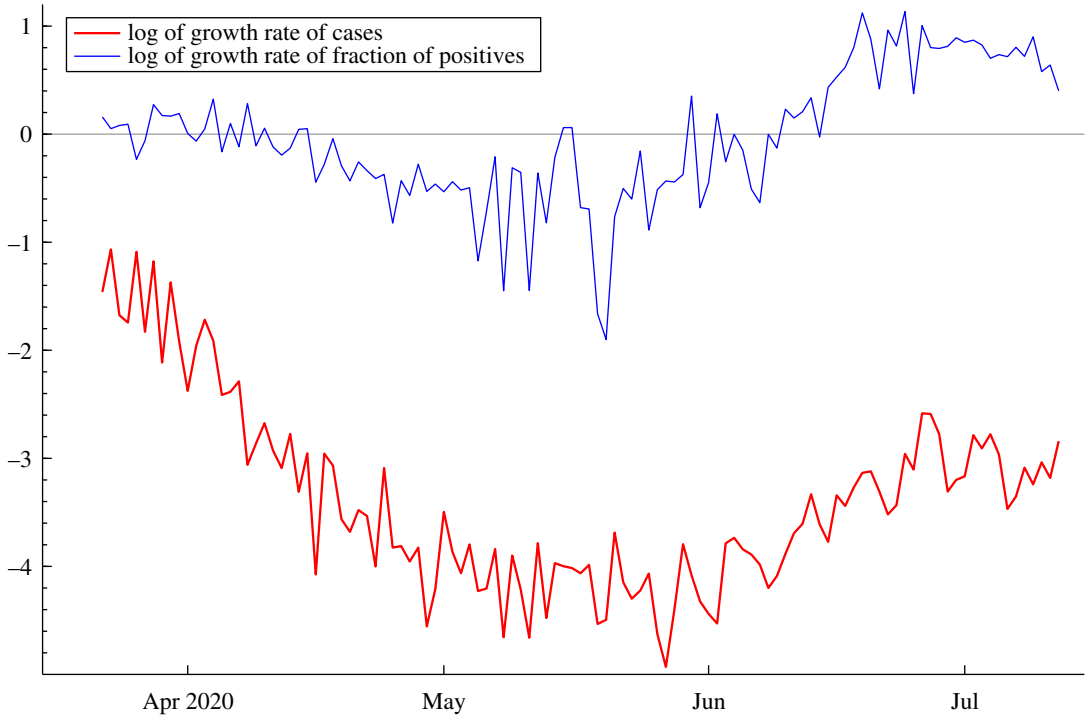

Figure 4. Logarithm of the growth rate of the total number of confirmed cases in Florida, together with the logarithm of the growth rate of the fraction of positives out of the total tested.

numerator. Although $\widehat{R}_{7,4,t}$ irons out some of the daily movement, the estimate of R_t is still affected. The model-based $\widetilde{R}^e_{4,t}$ evolves more smoothly. After June the data give no indication of a sustained increase in new cases so the jump in estimates of R_t, particularly $\widehat{R}_{4,t}$, can safely be classed as a spike.

The model was estimated using data from 25 March to 26 June 2020, using cases data sourced from the European Centre for Disease Prevention and Control (ECDC) website.[7] Estimates obtained using RKI's nowcast data are not very different.[8] The fit was good with very little evidence of residual serial correlation; the $Q(15)$ statistic is 9.58. A Gaussian distribution seems a good approximation because the Bowman–Shenton test statistic, which is asymptotically distributed as χ^2_2 under the null hypothesis, is only 0.77. The estimate of q was 0.0026.

The SD of the conditional distribution of γ_t is 0.0276. Thus the SD of $R_t = 1 + 4\gamma_t$ is 0.110. For $R^e_{4,t}$ setting $E_t(R^e_{4,t}) = 1$ gives $\mathrm{SD}_t(R^e_{4,t}) = 0.111$, so the probability that it lies in the interval[9] [0.895, 1.117] is 0.68. It makes little difference whether R_t is taken to be normal or lognormal. As regards the contribution of g_t to the variability $g_{y,t}$, the 26 June value of $g_{T|T}$ was only 0.0030 and $\mathrm{SD}(g_T)$ was less than 1% of the SD of γ_T.

4.2. Waves

The US state of Florida, the third most populous in the USA with a population of around 20 million, provides an example of a second wave. A graph of daily new cases[10] from early March until 19 July 2020 shows a peak in early April followed by a steady decline. This is similar to the pattern for Germany and reflects the fact that Florida, like Germany, was in lockdown during April 2020. After April restrictions in Florida were eased there was a levelling out in May 2020, followed by a sharp rise in June.

Figure 4 shows the logarithm of the growth rate of the number of confirmed cases, deaths and fraction of positives, starting 22 March 2020. (Before 22 March the data are very erratic.) After May there was an increase in testing. However, the growth rate in tests is roughly constant from the end of May onwards and this shows up in figure 4, where the logarithm of the growth in the proportion of positives follows a similar path to that of the logarithm of the growth in total cases. This suggests that confirmed cases are still a good indicator of the path of new infections.

Fitting the dynamic Gompertz model, with a daily component, to data on confirmed cases from 22 March to 12 July 2020 gave residuals with very little residual serial correlation as the $Q(16)$ statistic was only 8.42. The Bowman–Shenton test statistic was only 0.11 so a Gaussian distribution cannot be rejected. Graphical confirmation for the good fit is provided by figure 5.

The signal–noise estimate, q, was 0.0014. Figure 6 shows the filtered estimates of g_t and γ_t. At the beginning of June, $\gamma_{t|t}$ becomes positive and its sharp rise is accompanied by an attendant rise in $g_{t|t}$. The increase in $\gamma_{t|t}$ continues until the end of June, when it changes direction and $g_{t|t}$ peaks. The implied time series of nowcasts of R_t follows directly from their sum, $g_{y,t|t}$. Thus $\widetilde{R}^e_{4,t}$ reaches 1.5 by the end of June 2020 and then falls in July so that ultimately $\widetilde{R}^e_{4,t} \simeq 1.1$.

The filtered estimates of g_t and γ_t for confirmed cases in Florida are very different from those for Germany in that $g_{t|t}$ no longer becomes negligible with time. Indeed for most of June it is of a similar order of magnitude to $\gamma_{t|t}$. Nevertheless its contribution to the variability of $g_{y,t}$ is still negligible. The SD of the conditional distribution of γ_t is 0.0275 while that of δ_t is 0.1296, translating into a SD of 0.0057 for g_t. If the covariance term is ignored, the SD of $g_{y,t}$ is 0.0281, which is only a little above the SD of γ_t.

Re-estimating the model with another week of data has $\gamma_{T|T}$ down to −0.031 but the SD is little changed at 0.027. The estimate of g_T is 0.034, giving $\widetilde{R}^e_{4,T} = 1.01$. Finally estimating using data up to 12 August leaves $\gamma_{T|T}$ virtually unchanged but, because $g_{T|T}$ is down to 0.011, $\widetilde{R}^e_{4,T}$ is below 1, with a value of 0.91.

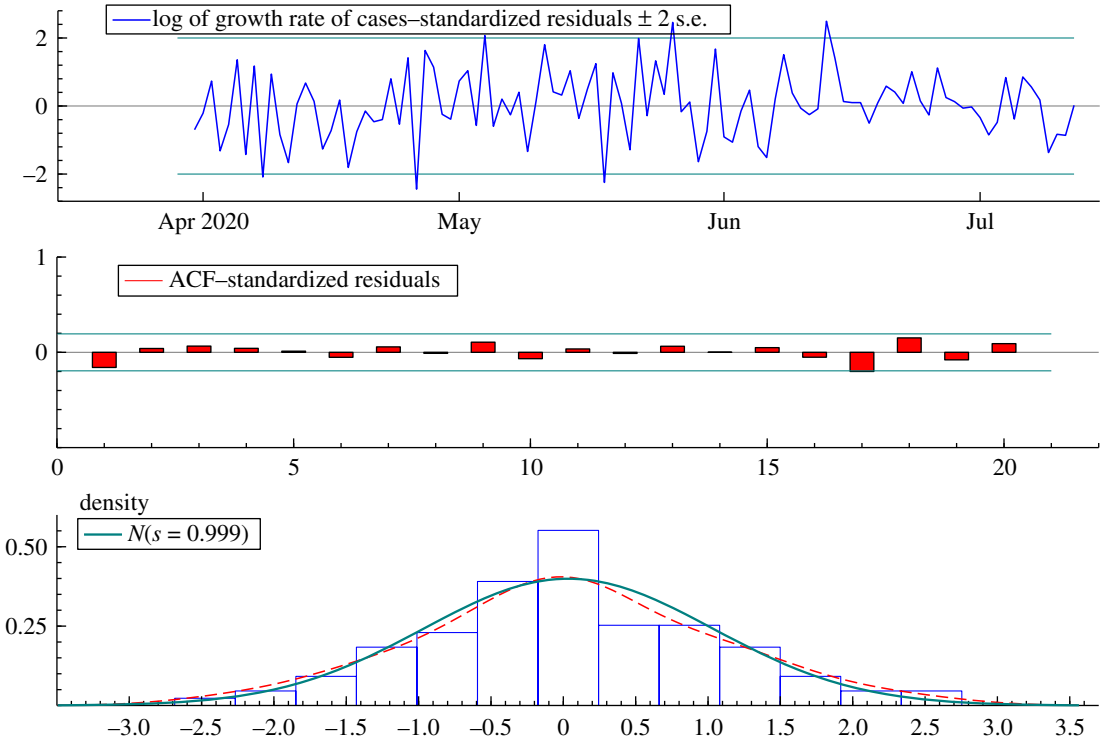

Figure 5. Residuals from fitting the model to the logarithm of the growth rate of Florida cases.

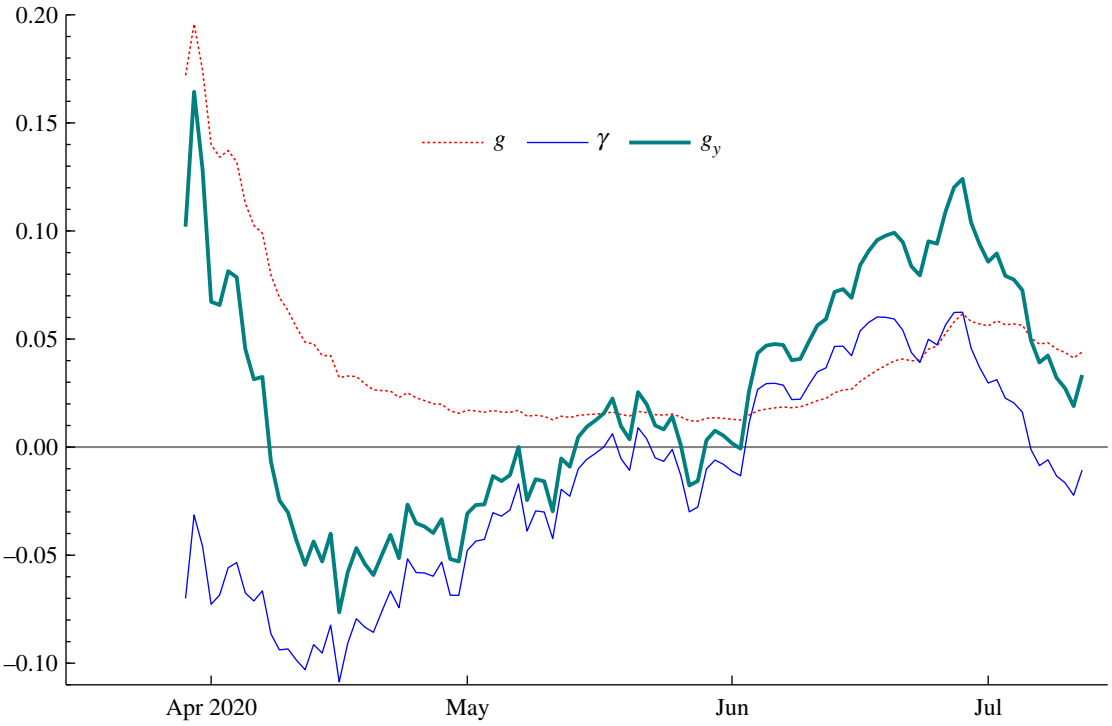

Figure 6. Filtered estimates of the growth rate, g_t, and slope, γ_t, for confirmed cases in Florida.

5. Conclusion

New time-series models are able to track the progress of an epidemic by providing nowcasts and forecasts of the daily number of new cases and deaths. Estimates and forecasts of the instantaneous reproduction number R_t can be computed as a by-product, using a formula that links it to the estimated growth rate of new cases, based on assumptions made about the serial interval distribution. The availability of the full conditional distribution allows the variability of the estimates to be assessed.

Current methods for tracking R_t do not pay due attention to the time-series properties of the data, whereas the method described in this paper is based on time-series techniques that have been shown to be effective in a range of disciplines. The dynamic response depends on a signal–noise ratio that can be estimated from the data rather than being inferred from knowledge about the serial interval of infections. An important element in time-series methodology is diagnostic checking and the fit of the model. We show how diagnostic methods can be applied in the context of epidemics and in doing so we raise questions about some of the assumptions,

explicit or implicit, that are currently made in the estimation of R_t. The ability of the model to track spikes and waves is illustrated with COVID-19 data from Germany and Florida.

We stress again that computing R_t is a by-product of our approach. Information on R_0 could be used at the start of an epidemic, but with a dynamic time-series model its impact soon wears off. After that, calculations involving R_t play no part in nowcasting and forecasting daily cases and deaths.

Data accessibility. Data employed in this study are publicly available. The data for reproducing figure 1 in the paper were downloaded from https://coronavirus.data.gov.uk/. The data are provided as electronic supplementary material.

Authors' contributions. Both authors contributed equally to the study and have approved the final manuscript. Both authors gave final approval for publication and agree to be held accountable for the work performed herein.

Competing interests. We declare we have no competing interests.

Funding. No funding has been received for this article.

Acknowledgements. We would like to thank Michael Ashby, Jouni Helske, Michael Höhle, Mark Salmon, Stefan Scholtes, David Spiegelhalter, Craig Thamotheram, Qingyuan Zhao and three anonymous referees for helpful comments and suggestions. Jonas Knecht supplied valuable research assistance.

Endnotes

[1]The models have been used as the basis for weekly trackers providing forecasts for regions and nations in the UK since the early part of 2021 (see https://www.niesr.ac.uk/latest-covid-19-tracker-0); and for states in India since May 2021 (see https://www.jbs.cam.ac.uk/covid-india). Past issues of both trackers are available on the websites.

[2]HK had a negative sign in front of γ in (2.1) and (2.2) because in a growth curve the growth rate is always falling so it is more convenient to let γ be positive. This ceases to be the case once there are second waves.

[3]A fully documented *R* script to carry out the estimation will be available shortly. An *R* script with the essential code is presented in the electronic supplementary material.

[4]Sourced from: https://coronavirus.data.gov.uk/.

[5]Erläuterung der Schätzung der zeitlich variierenden Reproduktionszahl R. Robert Koch Institute, 15 May 2020.

[6]The continuous-time incidence curve is $\mu'(t) = g(t)\mu(t)$, where $\mu(t)$ is the growth curve and $g(t)$ is its growth rate. Taking logarithms and differentiating provides the rationale for our formula for $g_{y,t\,|\,t}$; see the discussion surrounding eqn 6.1 in HK.

[7]Sourced from: https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide.

[8]Sourced from: https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Projekte_RKI/Nowcasting.html.

[9]Since R_t, like γ_t, is a random variable this is not, strictly speaking, a confidence interval. In a fully Bayesian framework, it would be called a credible interval.

[10]Data on Florida are sourced from: https://covidtracking.com/data.

Appendix A. Cori *et al.* [5] method for estimating R_t

Following Cori *et al.* [5], Thompson *et al.* [6] generate an estimate of the current level of new cases, y_t, that combines the estimate of R_t with an estimate, $\Lambda_{t|t-1}$, of the previous level based on the sum of new cases in the previous time period weighted by the infectivity function, or infectious profile after infection, f_j, $j = 0, 1, 2, \ldots$. This estimate can be written $\Lambda_{t|t-1} = \sum_{j=1}^{t} f_j y_{t-j}$, where $\sum_{j=\tau}^{t} f_j = 1$ with f_j describing the serial distribution. This distribution is based on prior knowledge, such as data collected from household studies in the early phase of an infection. The estimate of R_t is obtained by Bayesian methods. Cori *et al.* [5, appendix 1] assume a Poisson distribution for y_t and a (conjugate) gamma prior for R_{t-1} with parameters a and b. The posterior mean of R_t—its nowcast—is then

$$\widehat{R}^*_{k,t} = \frac{a + \sum_{j=0}^{k-1} y_{t-j}}{b^{-1} + \sum_{j=1}^{t-1} \Lambda_{t-j|t-j-1}} = \frac{a + \sum_{j=0}^{k-1} y_{t-j}}{b^{-1} + \sum_{j=1}^{t-1} \sum_{i=1}^{j} f_j y_{t-i}} = \frac{a + \sum_{j=0}^{k-1} y_{t-j}}{b^{-1} + \sum_{j=1}^{t} w_j y_{t-j}}, \tag{A 1}$$

while the associated nowcast for the mean of new cases is

$$\widehat{\mu}_{t|t} = \widehat{R}^*_{k,t|t} \Lambda_{t|t-1} = \frac{a + \sum_{j=0}^{k-1} y_{t-j}}{b^{-1} + \sum_{j=1}^{t} w_j y_{t-j}} \sum_{j=1}^{t} f_j y_{t-j}.$$

It is proposed in [6] that $a = 1$ and $b = 5$ at the outset so that both the mean, ab, and SD, ab^2, are set to 5. With no prior information $a = 1/b = 0$.

The numerator in $\widehat{R}^*_{k,t}$ provides an estimate of the level at time t based on the last k observations. The value of k reflects a trade-off between response and stability. The choice of equal weights seems to be arbitrary. The weights in the second term reflect the structure implied by the sometimes imperfect knowledge of the distribution of the serial interval.

Cori *et al.* [5, appendix 2] suggest letting the summation in the denominator of equation (A 1) start at $j = \tau$, where τ is the generation interval. Approximating the weights by a simple moving average of length k and assuming no prior information then gives equation (3.1), which is the formula for the instantaneous reproduction number used by RKI. The value of k may be set equal to the length of the serial interval.

In our time-series approach, the lag structure depends solely on the observations, y_t, and the properties of the fitted model. The weights in figure 2 are comparable with those used to construct $\widehat{R}^*_{k,t}$ in equation (A 1), except that they are multiplicative in the implied estimator of R^e_t. The negative weights in figure 2 correspond to the weights in the denominator of $\widehat{R}^*_{k,t}$.

References

1. Wallinga J, Lipsitch M. 2007 How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc. R. Soc. B* **274**, 599–604. (doi:10.1098/rspb.2006.3754)
2. Gostic K *et al.* 2020 Practical considerations for measuring the effective reproductive number. *medRxiv*. (doi:10.1371/journal.pcbi.1008409)
3. Birrell P, Blake J, van Leeuwen E, Gent N, De Angelis D. 2020 Real-time nowcasting and forecasting of COVID-19 dynamics in England: the first wave. *Phil. Trans. R. Soc. B* **376**, 20200279. (doi:10.1098/rstb.2020.0279)
4. Harvey A, Kattuman P. 2020 Time series models based on growth curves with applications to forecasting coronavirus. *Harvard Data Science Review*. Special issue 1—COVID-19. See https://hdsr.mitpress.mit.edu/pub/ozgjx0yn. (doi:10.1162/99608f92.828f40de)

5. Cori A, Ferguson N, Fraser C, Cauchemez S. 2013 A new framework and software to estimate time-varying reproduction numbers during epidemics. *Am. J. Epidemiol.* **178**, 1505–1512. (doi:10.1093/aje/kwt133)
6. Thompson R *et al.* 2019 Improved inference of time-varying reproduction numbers during infectious disease outbreaks. *Epidemics* **29**, 100356. (doi:10.1016/j.epidem.2019.100356)
7. Chowell G, Sattenspiel L, Bansal S, Viboud C. 2016 Mathematical models to characterize early epidemic growth. *Phys. Life Rev.* **18**, 66–97. (doi:10.1016/j.plrev.2016.07.005)
8. Bettencourt LM, Ribeiro RM. 2008 Real time Bayesian estimation of the epidemic potential of emerging infectious diseases. *PLoS ONE* **3**, e2185. (doi:10.1371/journal.pone.0002185)
9. Viboud C, Simonsen L, Chowell G. 2016 A generalized growth model to characterize the early ascending phase of infectious disease outbreaks. *Epidemics* **15**, 27–37. (doi:10.1016/j.epidem.2016.01.002)
10. Durbin J, Koopman SJ. 2012 *Time series analysis by state space methods*. Oxford, UK: Oxford University Press.
11. Harvey A. 1989 *Forecasting, structural time series models and the Kalman filter*. Cambridge, UK: Cambridge University Press.
12. Koopman S, Lit R, Harvey A. 2021 STAMP 9.0: structural time series analyser, modeller and predictor. Timberlake Consultants.
13. Lit R, Koopman S, Harvey A. 2020 Time series lab—score edition. See https://timeserieslab.com.
14. Höhle M. 2014 Bayesian nowcasting during the STEC O104:H4 outbreak in Germany, 2011. *Biometrics* **70**, 993–1002. (doi:10.1111/biom.12194)
15. Abbott S *et al.* 2020 Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. *Wellcome Open Res.* **5**, 112. (doi:10.12688/wellcomeopenres.16006.2)
16. Koopman S, Harvey A. 2003 Computing observation weights for signal extraction and filtering. *J. Econ. Dyn. Control* **27**, 1317–1333. (doi:10.1016/S0165-1889(02)00061-1)
17. Chowell G, Nishiura H, Bettencourt LMA. 2007 Comparative estimation of the reproduction number for pandemic influenza from daily case notification data. *J. R. Soc. Interface* **4**, 155–166. (doi:10.1098/rsif.2006.0161)