National Institute Economic Review (2021), 256: 1, 110–126
doi:10.1017/nie.2021.12

NIER CAMBRIDGE UNIVERSITY PRESS

ARTICLE

TRACKING THE MUTANT: FORECASTING AND NOWCASTING COVID-19 IN THE UK IN 2021

Andrew Harvey[1], Paul Kattuman[2*] and Craig Thamotheram[3]

[1]Faculty of Economics, University of Cambridge, Cambridge, United Kingdom
[2]Cambridge Judge Business School, University of Cambridge, Cambridge, United Kingdom
[3]National Institute of Economic and Social Research, London, United Kingdom
*Corresponding author. Email: p.kattuman@jbs.cam.ac.uk

(Received 10 February 2021; revised 01 March 2021; accepted 03 March 2021)

A new class of time series models is used to track the progress of the COVID-19 epidemic in the UK in early 2021. Models are fitted to England and the regions, as well as to the UK as a whole. The growth rate of the daily number of cases and the instantaneous reproduction number are computed regularly and compared with those produced by SAGE. The results from figures published each day are compared with results based on figures by specimen date, which may be more accurate but are subject to substantial revisions. It is then shown how data from the two different sources can be combined in bivariate models.

Keywords: data revisions; epidemic; Kalman filter; reproduction number (*R*); state-space model.

JEL codes: C22; C32.

1. Introduction

The application of classical time series methods to data on epidemics is relatively undeveloped. Most of the emphasis has been on building models to simulate the path of an epidemic under different assumptions about behaviour and policies, and the forecasting performance has often been unimpressive; see Avery *et al.* (2020) and Ioannidis *et al.* (2020). Here, we show how a new class of time series models can be used to track the progress of an epidemic and forecast key indicators. The methods draw much of their inspiration from econometrics, but take into account the special characteristics of time series for epidemics.

The univariate time series model described in Harvey and Kattuman (2020a)—hereafter HK—fits a trend to the logarithm of the growth rate of the cumulated series of the target variable, which is usually new cases, hospital admissions or deaths. Allowing this trend to be time-varying introduces flexibility which, in the context of an epidemic, enables the effects of changes in policy and population behaviour to be tracked. Such stochastic trend models are a standard econometric tool, and they are easily handled within a state-space framework. Application of the Kalman filter (KF) enables nowcasts and forecasts of variables of interest, such as the growth rate of the daily number of cases and the instantaneous reproduction number, to be made. Estimation of the models is by maximum likelihood and goodness of fit can be assessed by standard statistical test procedures.

This article describes our experience tracking the progress of the COVID-19 epidemic in the UK in early 2021. This period is of considerable interest, because a new variant of the virus appeared in the south-east of England in December 2020 and started to spread throughout the country. The lockdown of 5 January 2021 was partially in response to this new variant. The number of new cases quickly rose to a peak around the beginning of the new year and then started to fall. The ability of models to respond to these movements in a timely fashion is clearly important. Here, we investigate how our models fared by

showing how the response, as captured by both nowcasts and forecasts, adapted to observations available on a daily basis.[1]

We first examined the results for the country as a whole before moving on to monitor the regions. Regional variation is significant, because in October, some areas of the country, such as north-west England, were particularly hard hit, whereas the big rises in December came primarily from the new variant and were mainly in the south-east. There are systematic movements in daily observations according to the day of the week with the figures for the weekend tending to be lower. Our model is able to take account of these movements without using 7-day moving averages (MA7s) which tend to result in a delayed response when there are rapid upward or downward movements.

Multivariate state-space models can combine information in different series. There are two data sources for new COVID cases. One is the figure published each day, whereas the other is by specimen date. The second series is subject to substantial revisions as new data are processed and the series only settles down after about 3 days. However, it may be a better indicator of the spread of the epidemic, and so the question arises as to whether the information it contains can be combined with that in the published data. This is essentially a question of combining different 'vintages', something which is often done with economic data. Sometimes, the observations are made in a different way and at different frequencies, for example, by surveys; see Harvey and Chung (2000) and, more recently, Anesti *et al.* (2021). Our treatment of published and specimen data owes much to this literature, but there are some novel features, primarily concerned with time-varying slopes and the notion of balanced growth. The methods may be generalised to deal with leading indicators as in Harvey (2020).

Section 2 of the paper reviews the model and explains how estimates of the growth rate of daily numbers can be made and how these yield corresponding estimates of instantaneous reproduction number, R_t. Our experience with UK data in January is reported in Section 3, and the multivariate models are described and implemented in Section 4.

2. Forecasting and nowcasting with the dynamic Gompertz model

The observational model uses data on the time series of the cumulated total of confirmed cases or deaths, Y_t, $t = 0, 1, \ldots, T$, and the daily change. HK show how the theory of generalised logistic growth curves suggests models for $\ln y_t$, where $y_t = \Delta Y_t = Y_t - Y_{t-1}$, and the logarithm of the growth rate of the cumulated series, $\ln g_t$, where $g_t = y_t / Y_{t-1}$ or $\Delta \ln Y_t$. For the special case of the Gompertz growth curve, the implication is that $\ln g_t$ follows a downward linear trend. However, additional flexibility is needed to cope with situations where there are recurrent waves. This may be achieved by a stochastic, or time-varying, trend, so that

$$\ln g_t = \delta_t + \varepsilon_t, \quad \varepsilon_t \sim NID\left(0, \sigma_\varepsilon^2\right), \qquad t = 1, 2, \ldots, T, \tag{1}$$

where[2]

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

[1]This methodology forms the basis for the weekly projections of new cases and the *R* number for the UK, its constituent nations and the regions of England, published weekly by NIESR from 18 February 2021 (https://www.niesr.ac.uk/latest-weekly-covid-19-tracker).

[2]HK had a negative sign in front of γ in (1) and (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.

https://doi.org/10.1017/nie.2021.12 Published online by Cambridge University Press

and the normally distributed 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 is an integrated random walk (IRW). HK found the IRW trend to be particularly useful for tracking an epidemic, and it will be adopted in the applications here. The speed with which a trend adapts to a change depends on the signal–noise ratio, which for the IRW is $q = \sigma_\zeta^2/\sigma_\varepsilon^2$; the trend is deterministic when $q = 0$.

Allowing γ_t to change over time means that the progress of the epidemic is no longer tied to the proportion of the population infected as it would be if Y_t followed a deterministic growth curve. Instead, the model adapts to movements brought about by changes in behaviour and policies. If γ_t falls to zero, the growth in Y_t becomes exponential, whereas a positive γ_t means that the growth rate is increasing.

Stochastic trend models can be estimated using techniques based on state-space models and the KF; see Durbin and Koopman (2012) and Harvey (1989). The computations for the multivariate model were performed using the STAMP package of Koopman *et al.* (2020), whereas the results reported in Section 3 were obtained with a new program in the R language specifically written for this project. 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 $\left(\delta_{t|t}, \gamma_{t|t}\right)'$ and given by the contemporaneous filter while the predictive filter outputs $\left(\delta_{t+1|t}, \gamma_{t+1|t}\right)'$. The smoother estimates the state at time t based on all T observations in the series and is denoted $\left(\delta_{t|T}, \gamma_{t|T}\right)'$. Estimation of the unknown variance parameters is by maximum likelihood. Tests for normality and residual serial correlation are based on the one-step ahead prediction errors, $v_t = \ln g_t - \delta_{t|t-1},\ t = 3, \ldots, T$.

Additional components, such as day of the week effects, can be added to (1). These may be deterministic or stochastic. Stationary autoregressive or ARMA components may also be included as may explanatory variables, including interventions. However, isolated outliers are most easily handled by treating them as missing observations.

Remark 1. *When the observations on daily cases or deaths are small, a negative binomial distribution for y_t, conditional on past observations including Y_{t-1}, may be appropriate. HK show how the model may be modified to deal with this possibility for a univariate time series. Software can be found in Lit et al. (2020). Estimates of the state based on small numbers are likely to be unreliable, but if the KF is to operate during periods when numbers are small, as they were for COVID-19 cases in the summer of 2020, it may be better to set $v_t = g_t \exp\left(-\delta_{t|t-1}\right) - 1$ rather than to treat the observation as missing.*

2.1. Forecasting and nowcasting the growth rate of daily observations and R

The direction in which an epidemic is moving is best tracked by nowcasts and forecasts of $g_{y,t}$, the growth rate of y_t. Harvey and Kattuman (2020b) construct the nowcast of $g_{y,t}$ from the filtered estimates in the state-space model [(1) and (2)]. Thus, $g_{y,t|t} = g_{t|t} + \gamma_{t|t}$. These estimates can be translated into estimates of the instantaneous reproduction number R_t, in a number of ways, as described in Wallinga and Lipsitch (2007). Harvey and Kattuman (2020b) argue that the most useful for COVID-19 are

$$\widetilde{R}_{t,\tau} = 1 + \tau g_{y,t|t} \quad \text{and} \quad \widetilde{R}^e_{\tau,t} = \exp\left(\tau g_{y,t|t}\right), \tag{3}$$

where $\tau = 4$; τ is the generation interval, that is, the number of days that must elapse before an infected person can transmit the disease. The nowcasts of y_t peak when $g_{y,t|t} = 0$, corresponding to $\widetilde{R}_{t,\tau} = \widetilde{R}^e_{\tau,t} = 1$.

For tracking and forecasting the epidemic, all that is needed are estimates of $g_{y,t}$. The estimates of R_t are a by-product. Despite being dependent on assumptions about the generation interval, estimates of R_t have become the main metric for reporting the state of the epidemic.

Predictions of $g_{y,t}$, and hence of R_t, are given by

$$g_{y,T+\ell|T} = \exp \delta_{T+\ell|T} + \gamma_{T+\ell|T} = \exp\left(\delta_{T|T} + \gamma_{T|T}\ell\right) + \gamma_{T|T}, \qquad \ell = 1, 2, \ldots. \tag{4}$$

https://doi.org/10.1017/nie.2021.12 Published online by Cambridge University Press

If $\gamma_{T|T}$ is zero, the growth of y_t is exponential, and it is helpful to characterise it by the doubling time, $\ln 2/g_{y,T|T} = 0.693 \exp\left(-\delta_{T|T}\right)$.

When $\exp\delta_{T|T} + \gamma_{T|T} > 0$, so that the nowcast $\widetilde{g}_{y,T|T}$ is positive and the estimates of R_T given by (3) are greater than one, there is still a saturation level so long as $\gamma_{T|T}$ is negative; correspondingly, as $\ell \to \infty$, $\widetilde{R}^e_{\tau,T+\ell|T} \to \exp\left(\tau\gamma_{T|T}\right) < 1$. Hence, a negative $\gamma_{T|T}$ signals a flattening of the curve and an upcoming peak in y_t.

Remark 2. *The basic forecasts are made with the estimates of δ_T and γ_T. However, alternative scenarios in which γ_t is assumed to evolve in a certain way, perhaps to reflect changing behaviour and policies, may also be envisaged. If a future scenario arises in terms of a time path for $R_{T+\ell|T}$, it can easily be translated into one for $\gamma_{T+\ell|T}$. The time path for $\gamma_{T+\ell|T}$ leads directly to the forecasting equations of (10), and so no simulations are needed for the predictions of $y_{T+\ell}$.*

2.2. Sampling variability of nowcasts and forecasts

Harvey and Kattuman (2020b) show that the conditional distribution of nowcasts of $g_{y,t}$ can be approximated by the conditional distribution of γ_t, which is normal with mean $\gamma_{t|t}$ and variance $\sigma^2_{\gamma,t|t}$, both of which are produced by the KF.

When $\widetilde{R}_{t,\tau}$ is defined as $1 + \tau g_{y,t}$, its distribution, conditional on current and past observations, can be treated as $N\left(g_{y,t|t}, \tau^2\sigma^2_{\gamma,t|t}\right)$. On the other hand, the conditional distribution of $\widetilde{R}^e_{t,t}$ is lognormal with mean

$$E_t\left(\widetilde{R}^e_{t,t}\right) = \exp\left(\tau\left(g_{t|t} + \gamma_{t|t} + (\tau/2)\sigma^2_{\gamma,t|t}\right)\right) \tag{5}$$

and standard deviation

$$SD_t\left(\widetilde{R}^e_{\tau,t}\right) = E_t\left(\widetilde{R}^e_{\tau,t}\right)\sqrt{\left(\exp\tau^2\sigma^2_{\gamma,t|t} - 1\right)}. \tag{6}$$

Note that $\exp\tau^2\sigma^2_{\gamma,t|t} - 1 \simeq \tau^2\sigma^2_{\gamma,t|t}$, so when $E_t\left(\widetilde{R}_{t,\tau}\right)$ is close to one, $SD_t\left(\widetilde{R}^e_{\tau,t}\right) \simeq SD_t\left(\widetilde{R}_{t,\tau}\right)$. The probability that R_t exceeds one is $\Pr\left(g_{y,t|t} > 0\right)$, and this does not depend on τ or the formula used to estimate R_t from $g_{y,t|t}$.

Remark 3. *For the Spanish flu data Chowell et al. (2007), discuss two approaches to estimating R_t based on Susceptible-Exposed-Infectious-Removed models, the more complex one having eight nonlinear differential equations. They also use the Bayesian method of Bettencourt and Ribeiro (2008). Estimates of R_t obtained from the model discussed at the end of this section are not out of line with those reported by Chowell et al. (2007), and they are simpler, more transparent and open to diagnostic checks on the statistical assumptions.*

As with nowcasts, the predictive distribution of $g_{y,T+\ell}$, and hence of $R_{T+\ell}$, can be approximated from the conditional distribution of $\gamma_{T+\ell}$ given observations up to and including time T. This is Gaussian with mean $\gamma_{T|T}$ and variance $\sigma^2_{\gamma,T+\ell|T}$. These estimates are produced by the predictive equations of the KF as in Harvey (1989, eq. 3.5.5, p. 147). For an IRW trend, it can be shown that

$$Var_T\left(g_{y,T+\ell}\right) \simeq Var_T\left(\gamma_{T+\ell}\right) = Var_T(\gamma_T) + \ell\sigma^2_\zeta = \sigma^2_{\gamma,T|T} + \ell q\sigma^2_\varepsilon \tag{7}$$

when the effect of the daily component is not included. The factor by which the variance of an ℓ step ahead forecast of $R_t = 1 + \tau g_{y,t}$ is inflated above that of the variance of the corresponding nowcast is the same as it is for $g_{y,t}$. For example, when $q = 0.005$, $\sigma^2_{\gamma,T|T} = 0.001$ and $\sigma^2_\varepsilon = 0.02$, expression (7) indicates that the SDs of g_{yt} and R_t will increase by 30 per cent for $\ell = 7$ and 55 per cent for $\ell = 14$.

The probability that $R_{T+\ell} > 1$ is $\Pr\left(g_{y,T+\ell} > 0\right) \simeq \Pr\left(z > -g_{y,T+\ell|T}/SD_{y,T+\ell|T}\right)$, where $z \sim N(0,1)$ and $SD_{y,T+\ell|T}$ is the square root of (7). Thus, for new cases in England by date of publication, the nowcast

https://doi.org/10.1017/nie.2021.12 Published online by Cambridge University Press

made on 18th January was −0.048, whereas the 14-day forecast was −0.054. The value of $\sigma^2_{\gamma,T|T}$ for $q = 0.005$, and a daily effect included, was 0.0004, while σ^2_ε was estimated to be 0.014. These values give $\Pr(R_T > 1) \simeq \Pr\left(z > 0.048/\sqrt{0.0004}\right) = \Pr(z > 2.40) = 0.008$ and $\Pr(R_{T+14} > 1) \simeq \Pr(z > 1.30) = 0.10$.

The ability to make predictions offers a way to deal with reporting delay, as described in Abbott *et al.* (2020, pp. 3–4). If the observation for time $t-k$ is not available until time t, the current R_t is better estimated by a k-step ahead forecast. Taking the parameter values of the previous paragraph gives an increase in the SD of 14 per cent for $k = 3$.

2.3. Moving averages

In the UK, the current level of new infections or deaths is usually reported together with the MA7, which is more stable than the daily figure and irons out the daily effects. The moving average figure is often divided by 100,000 so as to give a standardised measure. Estimates of $g_{y,t}$ and R_t can be calculated directly from the moving average. For example,

$$\widehat{R}_{t,k,\tau} = \frac{\sum_{j=0}^{k-1} y_{t-j}}{\sum_{j=\tau}^{k+\tau-1} y_{t-j}} = \frac{\sum_{j=0}^{k-1} y_{t-j}}{\sum_{j=0}^{k-1} y_{t-\tau-j}} = 1 + \tau \widehat{g}_{y,t}, \tag{8}$$

where the sum in the denominator starts at a lag of τ and the sums in the numerator and denominator may overlap. The lag of τ reflects the generation interval, which is number of days that elapse before an infected person can transmit the disease. The Robert Koch Institute (RKI) estimator[3] has $\tau = 4$, and $k = 4$ or 7; setting $k = 7$ has the advantage of smoothing out the daily effect.

Following on from (8), estimates of $g_{y,t}$ can be calculated directly from the moving average. However, because the observations are best captured by a location/scale model in which the level is proportional to scale, estimates formed from the level have poor statistical properties. A better approach would be to take logarithms before averaging. Harvey and Kattuman (2020b, Section 3.3) show that doing so would give a result much closer to that obtained from the model.

A disadvantage of using simple moving averages to track the epidemic is that they give the last seven observations equal weights and so can be slow to respond to upward or downward movements. By contrast, the model deals directly with day of the week effects and so is able to gradually discount past observations. Hence, it can respond more quickly. Figure 1 shows the nowcasts of the underlying trend in new cases produced by the model for Germany (European Centre for Disease Prevention and Control [ECDC] data), together with the MA7. The attraction of the model is clear, and, of course, it also has the advantage of being able to produce forecasts. Lagging the MA7 so it is centred at $t-3$ would shift it more in line with the observations but at the cost of losing the last three observations.

2.4. Forecasting the trend in future observations

The forecasts of the trend in future values of $\ln g_t$ in the dynamic Gompertz model are given by $\delta_{T+\ell|T} = \delta_{T|T} + \gamma_{T|T}\ell$, $\ell = 1, 2, \ldots$, where $\delta_{T|T}$ and $\gamma_{T|T}$ are the KF estimates of δ_T and γ_T at the end of the sample. Forecasts of the trend in the daily observations, y_t, may be obtained from a recursion for the trend in their cumulative total, Y_t, namely

$$\mu_{T+j|T} = \mu_{T+j-1|T}\left(1 + g_{T+j|T}\right) = \mu_{T+j-1|T}\left(1 + \exp\delta_{T+j|T}\right), \qquad j = 1, 2, \ldots, \ell, \tag{9}$$

with $\mu_{T|T} = Y_T$. The trend in the daily figures is then

[3]There is some prior nowcasting to account for reporting delays; the methodology is based on Höhle and an der Heiden (2014).

https://doi.org/10.1017/nie.2021.12 Published online by Cambridge University Press

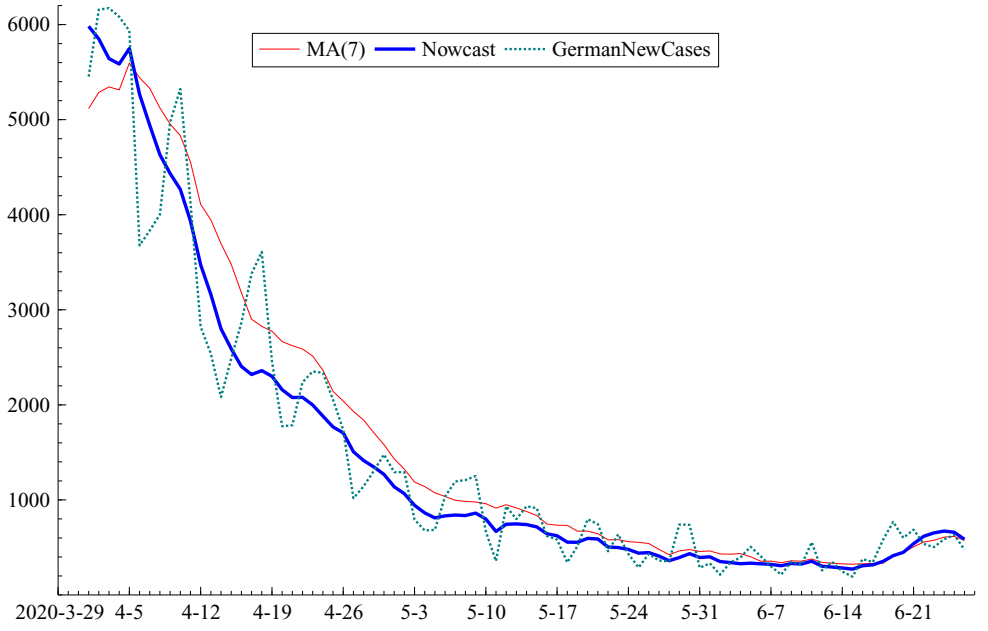

Figure 1. (Colour online) German new cases from 29th March to 26th June (data sourced from ECDC) showing nowcasts from model and 7-day moving averages

$$\mu_{y,T+\ell|T} = g_{T+\ell|T}\mu_{T+\ell-1|T}\,, \quad \ell = 1,2,\ldots.$$

Combining the above equations gives

$$\begin{aligned}\mu_{y,T+\ell|T} &= Y_T \exp\delta_{T+\ell|T}\prod_{j=1}^{\ell-1}\left(1+\exp\delta_{T+j|T}\right), \quad \ell = 2,3,\ldots,\\ \mu_{y,T+1|T} &= Y_T \exp\delta_{T+1|T}.\end{aligned} \tag{10}$$

Daily effects can be added to δ_t. In this case, forecasts of the observations themselves, that is, $\widehat{y}_{T+\ell|T}$ and $\widehat{Y}_{T+\ell|T}$, are given by adding the filtered value of the daily component to the trend component, $\delta_{T+\ell|T}$.

The conditional distribution of future values of the trend, $\delta_{T+\ell}$, in $\ln g_{T+\ell}$ is Gaussian. The conditional distribution of $\exp\delta_{T+\ell}$ is therefore lognormal, but, for more than one-step ahead, the distribution of the corresponding trend in the observations is not lognormal because of the presence of the unknown cumulative total in our equation for the underlying trend which is $\mu_{y,T+\ell} = g_{T+\ell}Y_{T+\ell-1}$, $\ell = 2,3,\ldots$. However, since Y_t changes relatively slowly, it may be possible to ignore its effect by treating it as fixed.

An alternative to working with the distribution of the trend of the observations is to convert a prediction interval for $\ln g_{T+\ell}$ into one for $\mu_{y,T+\ell}$ by replacing $\delta_{T+j|T}$ in (9) by $\delta_{T+j|T} \pm z.\sigma_{\delta,T+j|T}$, where $\sigma^2_{\delta,T+j|T}$ is the conditional variance of δ_{T+j} and z is a constant such as one or two. Again, with Y_t changing slowly, there may be a case for simply constructing a prediction interval from (10) by replacing $\delta_{T+\ell|T}$ by $\delta_{T+\ell|T} \pm z.\sigma_{\delta,T+j|T}$ for $\ell = 1,2,3,\ldots$. If a prediction interval for the observations themselves is wanted, the standard deviation $\sigma_{\delta,T+j|T}$ may be replaced by $\sqrt{\sigma^2_{\delta,T+j|T}+\sigma^2_{\varepsilon}}$ in the preceding formulae. Allowance may also need to be made for a daily component.

https://doi.org/10.1017/nie.2021.12 Published online by Cambridge University Press

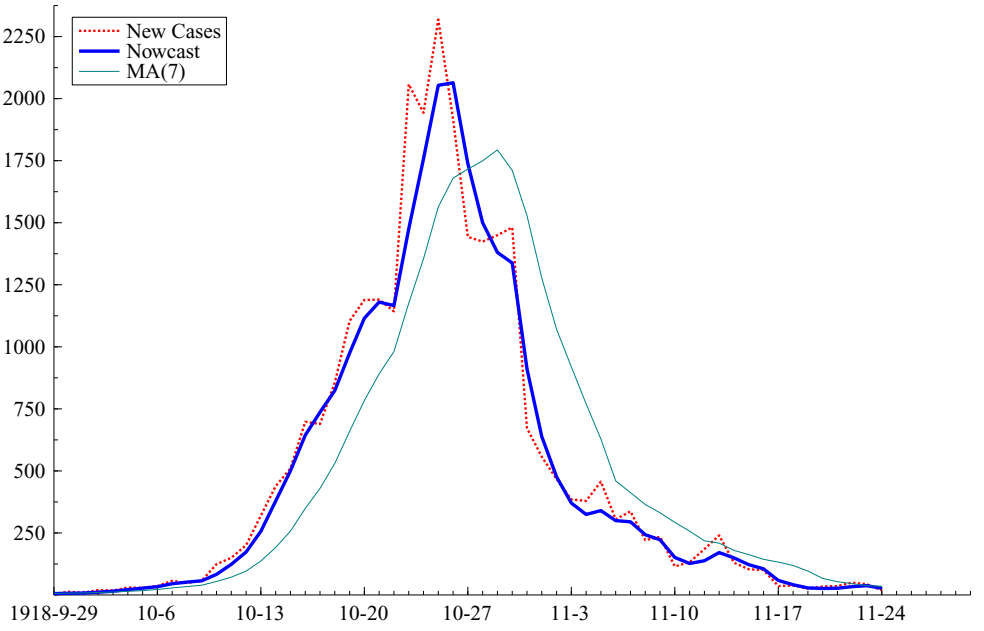

Figure 2. (Colour online) Nowcasts and 7-day moving averages for San Francisco flu from 29 September to 24 October 1918

2.5. Nowcasts of the trend in daily observations

The nowcast for the trend in y_t is

$$\mu_{y,t|t} = Y_{t-1} \exp \delta_{t|t}, \qquad t = t', \ldots, T.$$

Using the current rather than the lagged cumulative total, that is, $Y_t \exp \delta_{t|t}$, makes virtually no difference once Y_t has become relatively large. Since δ_t is conditionally Gaussian, $\exp \delta_t$ is lognormal, and a credible interval may be produced if required.

Example *Daily cases of influenza*[4] *in San Francisco during the worldwide outbreak of Spanish flu in 1918 show exponential growth in the upward phase; see Chowell et al. (2007). Consequently, a plot of the logarithm of the growth rate (LDL) shows very little downward movement at first. Fitting the Gaussian dynamic Gompertz to the whole series gives* $q = 0.05$. *The slope in LDL adapts, so it is close to zero in early October and then falls so as to capture the downward phase. Figure 2 contrasts the nowcasts with an MA7, which lags behind the observations throughout.*

3. COVID-19 in the UK and regions

Our empirical focus is on trends in new cases in the UK, its nations and English regions. We concentrate on early 2021, when the new strain of COVID-19 was the leading cause of increase in infection rates, initially in the south-east of England.

The daily counts of COVID-19 cases are based on the results of laboratory-based or swab tests for the presence of SARS-CoV-2 virus in specimens taken from people, as well as results of antibody serology tests. The new cases' data are available by the date the specimen was collected (the specimen date series) and by the date the testing process was completed and the case was first included in the published totals

[4]The data are supplementary material to the article by Chowell (2007; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2358966).

https://doi.org/10.1017/nie.2021.12 Published online by Cambridge University Press

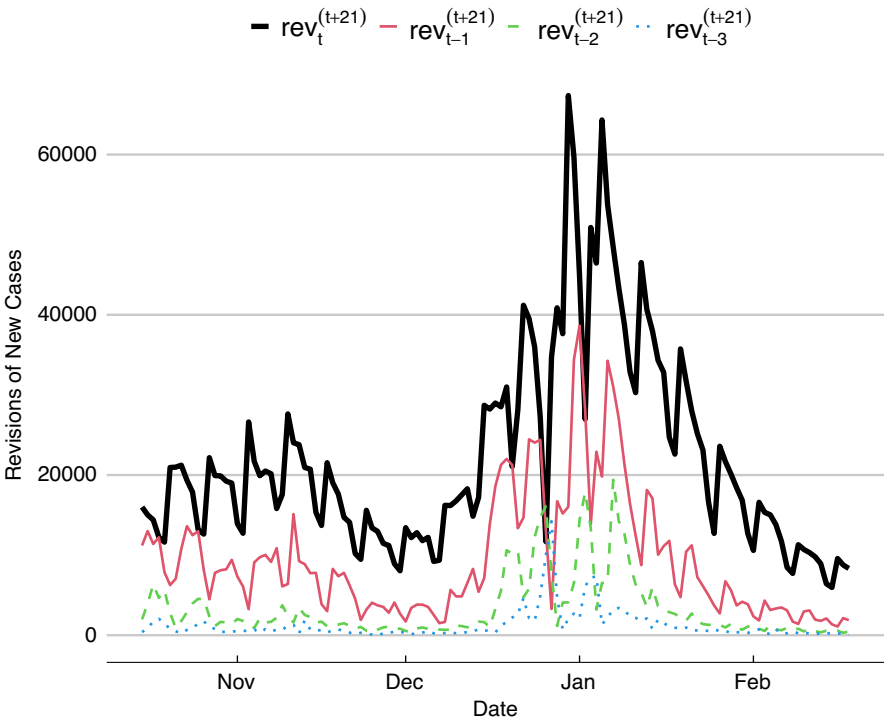

Figure 3. (Colour online) Differences between new cases in the specimen date series at t and the same data revised 3 weeks in the future

(published date series). A key issue is how information in the specimen date series can be combined with that in the published date series. There are pronounced daily effects in the specimen date series, which, however, can be accommodated in a model with seasonal effects. More importantly, the specimen date series is subject to substantial reporting delays and revisions. Figure 3 illustrates the extent of revisions over time of the specimen date series, with reference to the same data as revised 3 weeks in the future by when revisions are almost complete and very small. To be precise, let $y_t^{(v)}$ denote the v-day ahead update (or vintage) for y_t, the specimen cases recorded on date t. Thus, the current vintage of the data is given by the series: $y_t^{(t)}, y_{t-1}^{(t)}, \ldots, y_1^{(t)}$. The date r revision to the current vintage of specimen cases for date i is defined as: $rev_i^{(r)} = y_i^{(r)} - y_i^{(t)}$. Figure 3 presents the revisions 3 weeks ahead for the last four entries of the current vintage, $rev_t^{(t+21)}, \ldots, rev_{t-3}^{(t+21)}$. It is evident that except in the neighbourhood of Christmas day and New Year's day when data quality was very poor, revisions were substantially complete within 3 days.

A technical issue led to a large number of infections that occurred between 25 September and 2 October 2020 going unrecorded and then being assigned to 3rd and 4th October, thereby creating an artificial spike. Rather than attempting to reallocate observations, we start our analysis with data published on 5 October 2020. Cases by specimen date were not affected by the above issue. On 27 November 2020, another technical issue led to the total number of people who tested positive being revised down.

We fit models to the logarithm of the growth rate of new cases as measured by the specimen date up to and including time $t-3$ and report nowcasts and forecasts of $g_{y,t+h}$ and $\widetilde{R}_{t+h}^{e}$ for $h = -3, 0, 7$ and 14. For models fitted to new cases measured by published date, we report nowcasts and forecasts for $h = 0, 7$ and 14. Note that the published data for time t are actually released at $t+1$.

The forecasts we generate make no assumptions about the effects of measures imposed to control the spread of the epidemic. Thus, the forecasts made at the start of the year overshoot the eventual numbers. As the restrictive measures begin to bite, the forecasts made by the model adapt.

https://doi.org/10.1017/nie.2021.12 Published online by Cambridge University Press

Table 1. England: $g_{y,t+h}$ and $\widetilde{R}^{e}_{t+h}$ based on publication date series

	h		
t	**0**	**7**	**14**
g_y			
2020-12-28	5.25	5.76	6.40
2021-01-04	5.47	6.04	6.76
2021-01-11	−1.31	−1.66	−1.95
2021-01-18	−4.78	−5.17	−5.43
2021-01-25	−7.02	−7.31	−7.48
2021-02-01	−4.99	−5.16	−5.27
R			
2020-12-28	1.23	1.26	1.29
2021-01-04	1.24	1.27	1.31
2021-01-11	0.95	0.94	0.92
2021-01-18	0.83	0.81	0.80
2021-01-25	0.76	0.75	0.74
2021-02-01	0.82	0.81	0.81

3.1. Nowcasts and forecasts in January 2021

Tables 1 and 2 present, at weekly intervals starting on 28 December 2020, nowcasts and forecasts of the growth rate of new cases, $g_{y,t}$,$g_{y,t+7}$ and $g_{y,t+14}$, and of the reproduction numbers, R_t,R_{t+7} and R_{t+14}, for England. Table 1 uses the publication date series, whereas table 2 is for the specimen date series. The projections from both series are based on trends without daily effects, and show broadly similar patterns of accelerating growth rates before Christmas that increased through the New Year to the first observations in 2021. The lock down of 5 January 2020 brought both sets of growth rates and reproduction numbers down to the same broad range within a week. The growth rates estimated from both the publication date series and the specimen date series have continued to be negative since then.

Figure 4 gives the forecasts of new cases based on publication date series for England, including the daily effect. Figure 5 gives the forecasts based on the specimen date series. Vertical dashed lines denote the end of the estimation sample. These figures demonstrate that once past the imposition of the January lockdown, the model adapted quickly to the change in the series and in a relatively stable environment provided accurate forecasts.

3.2. Forecast accuracy

We assess forecast accuracy using mean absolute percentage error (MAPE) over the 14-day period from the date on which the data are released. For the publication date series, we evaluate forecasts against subsequent realisations of the same series, whereas for specimen date series, we evaluate forecasts against the first vintage with a release date that allows the first major revisions to vanish from the evaluation sample. This also maintains a fixed number of days for each evaluation date relative to the forecast origin for revisions to enter the evaluation sample. Thus, evaluation data for the specimen data series with vintage dated t require evaluation data of vintage (v) from $y^{(v)}_{t-2}$ to $y^{(v)}_{t+14}$, because we truncate the estimation sample at $y^{(t)}_{t-3}$. We choose a vintage of $v = t + 17$ to allow for the discarding of the heavily

https://doi.org/10.1017/nie.2021.12 Published online by Cambridge University Press

Table 2. England: $g_{y,t+h}$ and $\widetilde{R}^{e}_{t+h}$ based on specimen date series

	h			
t	**−3**	**0**	**7**	**14**
g_y				
2020-12-28	2.02	2.03	2.06	2.10
2021-01-04	4.24	4.37	4.73	5.13
2021-01-11	−2.39	−2.60	−3.01	−3.31
2021-01-18	−3.96	−4.14	−4.46	−4.68
2021-01-25	−4.35	−4.47	−4.70	−4.85
2021-02-01	−4.62	−4.70	−4.85	−4.95
R				
2020-12-28	1.08	1.08	1.09	1.09
2021-01-04	1.18	1.19	1.21	1.23
2021-01-11	0.91	0.90	0.89	0.88
2021-01-18	0.85	0.85	0.84	0.83
2021-01-25	0.84	0.84	0.83	0.82
2021-02-01	0.83	0.83	0.82	0.82

revised data at $t+15$ to $t+17$. Where this is not possible due to lack of data, we set v equal to the last date upon which data are available (which in figure 5 is 23 February 2021). For publication date series ending at time t, we require evaluation data from $y^{(j)}_{t+1}$ to $y^{(j)}_{t+14}$. As the publication data are just the first release of the specimen data, this results in the evaluation sample $y^{(t+1)}_{t+1},\ldots,y^{(t+14)}_{t+14}$.

Table 3 reports the MAPE for the forecasts of new cases based on the publication date series. Recall that data on new cases at t become available at $t+1$ in the publication date series. Table 4 reports the corresponding forecasts for the specimen date series. Accuracy is comparable for nowcasts and forecasts generated from the two series. Once the shocks to data quality over Christmas and the New Year are past and the initial effect of the January lockdown has worked through, both the 7- and 14-day ahead forecasts become more accurate.

3.3. Comparison with *R* published by DHSS and SAGE

The benchmarks for our results are the estimates of the growth rate and R values published jointly by the Department of Health and Social Care and the Scientific Advisory Group for Emergencies, based on contributions by different modelling groups using a variety of data sources. Estimates can vary between different models and are presented as ranges. For example, on 5 February 2021, the published range estimates for England were $[0.7, 0.9]$ for R_t and $[-5\%, -2\%]$ for the growth rate. Note that due to time delays, estimates reflect transmission of the disease over the past few weeks.

Figure 6 presents the model-based estimates of R for England using the publication and specimen date series, and for comparison, the empirical estimate of R based on the RKI estimator (Section 2.3), as well as the range estimates of R published by SAGE, is obtained from https://www.gov.uk/guidance/the-r-number-in-the-uk. The model-based estimates of R are quicker to reveal the effect of the January lockdown on infection transmission than the SAGE estimates.

https://doi.org/10.1017/nie.2021.12 Published online by Cambridge University Press

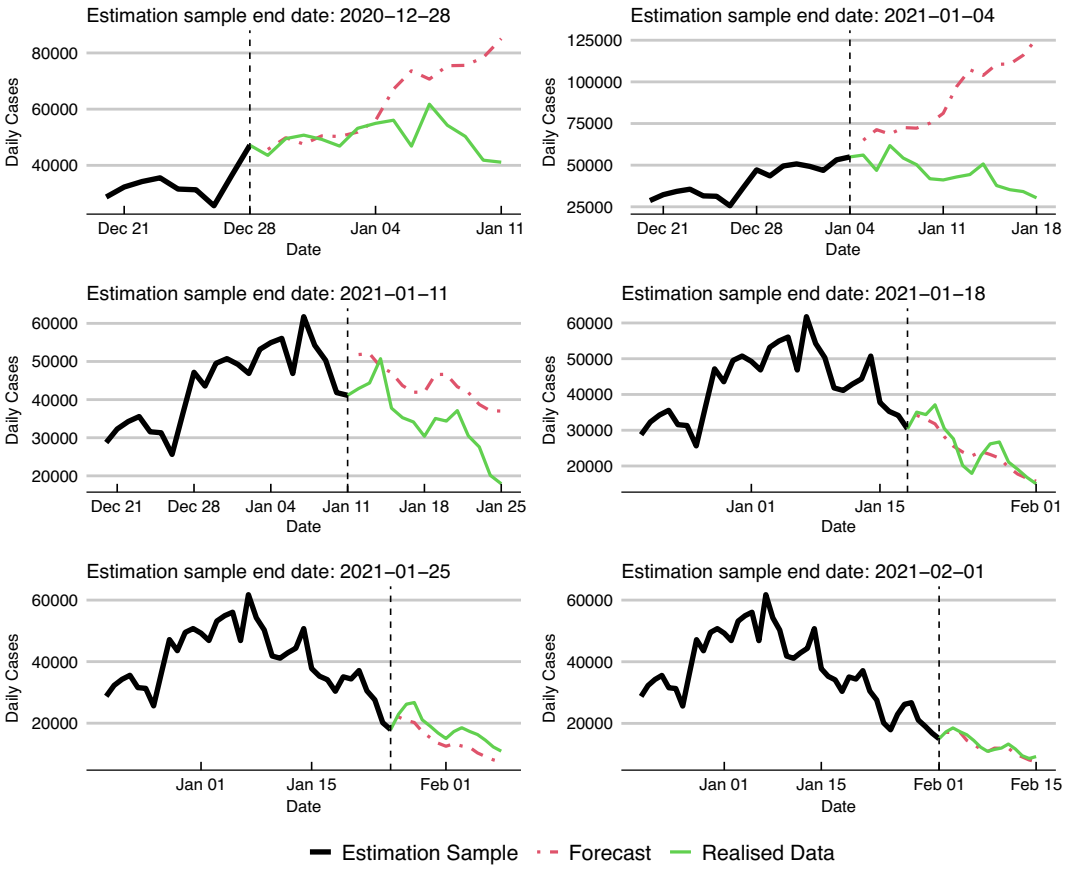

Figure 4. (Colour online) England: forecasts of new cases based on publication date series

Analysis and results corresponding to the above for the UK, other nations and the regions of England are presented in an online supplement to this paper.

4. Combining observations by publication date and specimen date

Methods of dealing with preliminary observations and observations at different vintages have long been employed in econometrics. Our treatment of published and specimen data owes much to this literature, but there are some novel features, primarily concerned with time-varying slopes and the notion of balanced growth. The techniques may be generalised to deal with situations where growth may not be balanced. Similar techniques may be employed when one series is a leading indicator of the other.

The bivariate model has observations on the first variable (published series), which is effectively a leading indicator, available at time t, whereas the second (specimen series) is only observed after k periods. Thus, at time t, the observations on $\ln g_{2t}$ are missing for $t-k+1,\ldots,t$. The model is

$$\begin{aligned} \ln g_{1t} &= \delta_t + \psi_{1t} + \varepsilon_{1t}, \qquad t = 1,\ldots,T, \\ \ln g_{2t} &= \delta_t + \overline{\delta} + \psi_{2t} + \varepsilon_{2t}, \\ \psi_{jt} &= \phi_j \psi_{j,t-1} + \eta_{jt}, \qquad \eta_{jt} \sim NID\left(0, \sigma^2_{\eta j}\right). \end{aligned} \tag{11}$$

https://doi.org/10.1017/nie.2021.12 Published online by Cambridge University Press

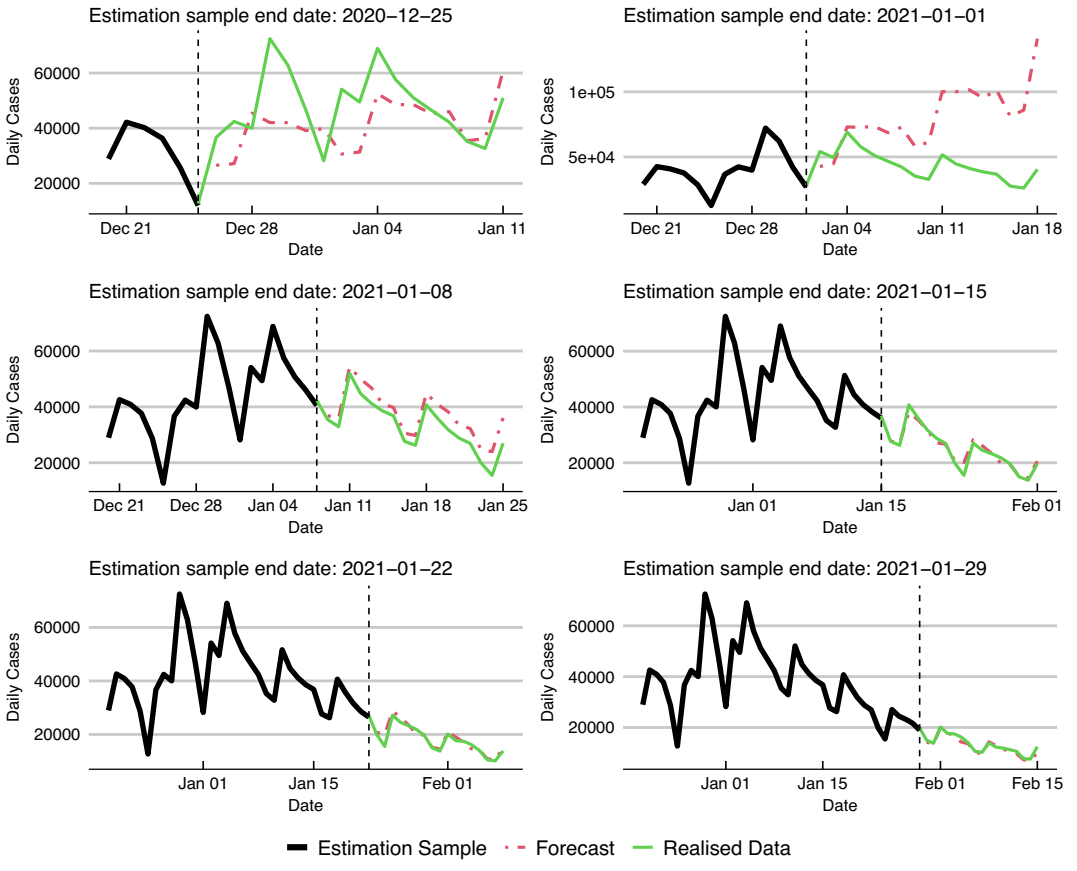

Figure 5. (Colour online) England: forecasts of new cases based on specimen date series

Table 3. England: accuracy (mean absolute percentage error) of forecasts of publication date series of new cases

	h		
t	**1**	**7**	**14**
2020-12-28	4.8	3.7	28.6
2021-01-04	15.9	47.7	119.1
2021-01-11	20.8	21.7	36.1
2021-01-18	2.5	11.7	9.7
2021-01-25	3.9	17.5	24.8
2021-02-01	1.0	5.1	6.5

where $\overline{\delta}$ is a constant term; δ_t will contain a constant that can be (arbitrarily) assigned to series one. As in the univariate models of the last section, the trend, δ_t, is an IRW that contains the information needed to estimate the underlying movements in the growth rate of the target series, $g_{2,y,t}$. All disturbances, including ε_{1t} and ε_{2t}, are Gaussian and assumed to be mutually as well as serially independent. Provided $|\phi_j| < 1, j = 1,2,\ldots$, the series are co-integrated of order (2,2), that is $CI(2,2)$, with balanced growth. The difference $\ln g_{1t} - \ln g_{2t}$ is a stationary ARMA(2,2) process, but setting $\phi_1 = \phi_2$ gives an AR(1) plus noise.

https://doi.org/10.1017/nie.2021.12 Published online by Cambridge University Press

Table 4. England: accuracy (mean absolute percentage error) of forecasts of specimen date series of new cases

	h			
t	−2	0	7	14
2020-12-28	27.7	25.8	31.4	22.2
2021-01-04	21.3	12.2	47.0	102.1
2021-01-11	3.9	5.9	9.2	16.1
2021-01-18	0.5	3.4	5.7	5.1
2021-01-25	4.2	13.1	6.6	6.0
2021-02-01	5.0	2.8	4.3	6.0

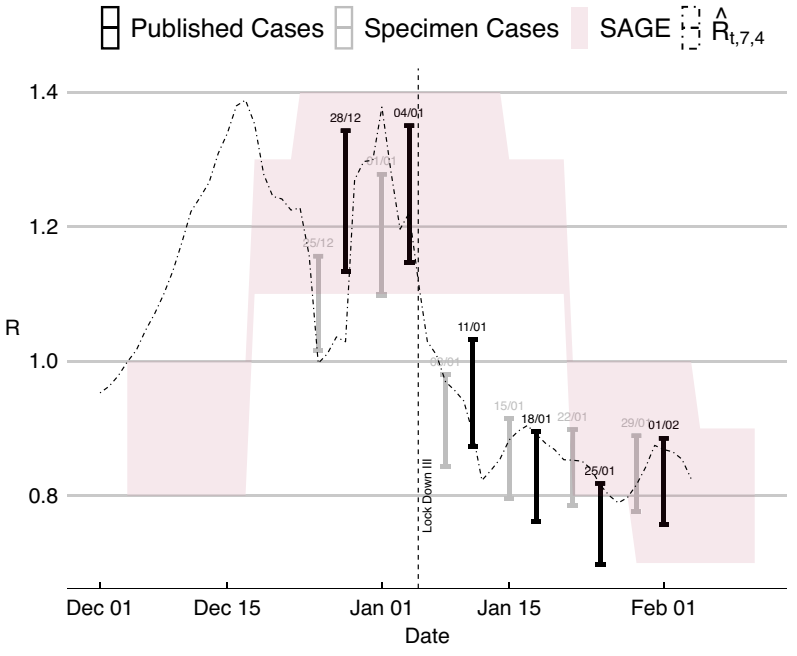

Figure 6. (Colour online) Estimates of *R* based on publication and specimen date series vis-a-vis *R* ranges published by SAGE

The KF provides the (filtered) state estimates needed to compute the nowcasts for $g_{2,y,T}$, R_T and $y_{2T.}$ As new observations become available, these nowcasts are updated by the KF. Smoothed estimates of variables from $t = T - k + 1$ to $t = T - 1$ can be computed if needed. Forecasts of the state beyond time T are made by the predictive KF, and corresponding forecasts of R_t and y_{2t} can be formed. Daily effects are included in the applications and are handled in (11) by adding a 'seasonal' component.

A modified version of the model confines the AR(1) component to the first variable, so that

$$\begin{aligned} \ln g_{1t} &= \delta_t + \psi_t + \varepsilon_{1t}, \qquad t = 1,\ldots,T, \\ \ln g_{2t} &= \delta_t + \overline{\delta} + \varepsilon_{2t}, \\ \psi_t &= \phi\psi_{t-1} + \eta_{1t}, \qquad \eta_t \sim NID\left(0,\sigma_\eta^2\right). \end{aligned} \tag{12}$$

https://doi.org/10.1017/nie.2021.12 Published online by Cambridge University Press

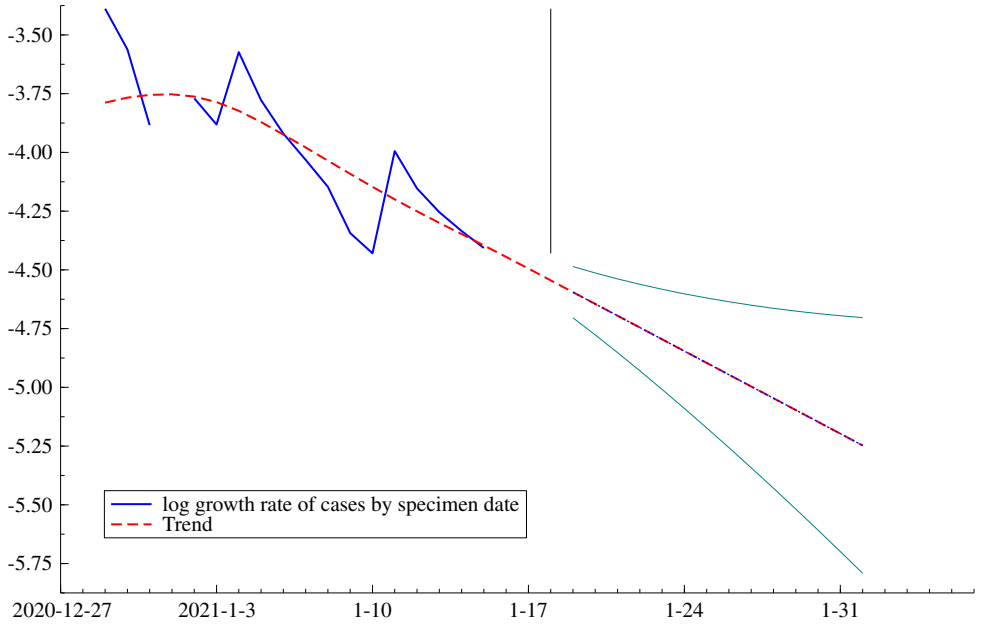

Figure 7. (Colour online) Forecasts of trend in specimen cases made with observations up to 15 January 2021 but using published data up to 18 January 2021

An advantage of this simplification is that the signal–noise ratio in the target can be compared with that of a univariate model and, if desired, set to a preassigned value.

Nowcasts of the trend in observations are obtained from recursions similar to those in Section 2.4, except that filtered estimates of δ_t are replaced by smoothed ones. Thus,

$$\mu_{T-k+\ell|T} = \mu_{T-k+\ell-1|T}\left(1+g_{T-k+\ell|T}\right) = \mu_{T+\ell-1|T}\left(1+\exp\delta_{T-k+\ell|T}\right), \qquad \ell = 1,2,\ldots,k \tag{13}$$

with $\mu_{T|T-k} = Y_T$, so

$$\mu_{y,T+\ell|T} = g_{T+\ell|T}\mu_{T+\ell-1|T}\,, \qquad \ell = 1,2,\ldots,k.$$

The recursions can be continued to give forecasts. The difference is that when $\ell > k$, the $\delta'_{T-k+\ell|T}s$ are forecasts, not smoothed estimates.

The bivariate model (12) was fitted to data available on 19th January. The observations start on 4th October and finish on 18th January for the published series and on 15th January for specimen series. The Christmas day and New Year specimen observations were treated as missing. The reasons for omitting the last three specimen dated figures were set out in Section 3. Estimation details can be found in Appendix B.

The estimates of $g_{y,T}$ and R^e_{4T} for the specimen dated series are -0.040 (0.026) and 0.85 (0.10), respectively. These are the same (to two decimal places) as the estimates for the univariate series. Figure 7 shows nowcasts, from 16th to 18th January, and forecasts, from 19th, of the trend in specimen data. The prediction interval for the observations on $\ln g_t$ is one standard deviation either side of its predicted trend. (A daily effect was not included in the model.)

5. Conclusions and future directions

This article has demonstrated the way in which our new time series models are able to track the progress of the COVID-19 epidemic in the UK in early 2021. The models are not only simple and transparent, but

https://doi.org/10.1017/nie.2021.12 Published online by Cambridge University Press

are able to adapt quickly to changes in key series. This ability to respond in a timely fashion is illustrated by the comparison of our estimates of the current R number with those produced by SAGE. The complexity of the behavioural response to lockdown and the restrictive measures imposed by the Tier system in different areas in late 2020 makes a formal structural modelling difficult. The roll out of the vaccine adds yet more complexity. Our models track these changes and project forward to make short-term forecasts of the situation over the next few weeks. Models are estimated for the four UK nations and for the regions within England. All the models have the same form.

We show how multivariate generalisations of our models can combine information in different series, some of which are subject to substantial revisions. The approach derives from econometric techniques for handling different vintages, but there are some novel technical features. The methods are new to epidemiology. We demonstrate that joint modelling of published and specimen dated observations on new cases can be accomplished without too much difficulty. The methods may be adapted to use some time series as leading indicators for others. Further work on using new cases as a leading indicator of admissions and deaths is currently underway.

Supplementary Materials. To view supplementary material for this article, please visit http://dx.doi.org/10.1017/nie.2021.12.

Acknowledgements. Comments and suggestions from Leopoldo Catania, Jagjit Chadha, Radu Cristea, Daniela de Angelis, Michael Höhle, Jonas Knecht, Rutger-Jan Lange, Franco Peracchi, Jerome Simons and a referee are gratefully acknowledged; of course, they bear no responsibility for opinions expressed or mistakes made. Some of the ideas were presented at the NIESR conference in November 2020 and at the University of Cambridge BSU seminar in January 2021. AH is grateful to the University of Cambridge Keynes Fund for support on the project Persistence and Forecasting in Climate and Environmental Econometrics.

References

Abbott, S., Hellewell, J., Thompson, R. N. *et al.* (2020), 'Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts'. *Wellcome Open Research*, 5:112. https://doi.org/10.12688/wellcomeopenres.16006.1.

Anesti, N., Galvão, A.B. and Miranda-Agrippino, S. (2021), 'Uncertain kingdom: Nowcasting GDP and its revisions', *Journal of Applied Econometrics*, to appear.

Avery, C., Bossert, W., Clark, A., Ellison, G. and Ellison, S. (2020), 'An economist's guide to epidemiology models of infectious disease', *Journal of Economic Perspectives*, **34**, 4, pp. 79–104.

Bettencourt, L.M., and Ribeiro, R.M. (2008), 'Real time Bayesian estimation of the epidemic potential of emerging infectious diseases'. *PLoS One*, **3**, e2185. https://doi.org/10.1371/journal.pone.0002185.

Chowell, G., Nishiura, H. and Bettencourt, L.M.A. (2007), 'Comparative estimation of the reproduction number for pandemic influenza from daily case notification data', *Journal of the Royal Society Interface*, **4**, 12, pp. 155–66.

Durbin, J. and Koopman, S. (2012), *Time Series Analysis by State Space Methods*, Oxford: Oxford University Press.

Harvey, A. (1989), *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge: Cambridge University Press.

Harvey, A. (2020), 'Time series models for epidemics: Leading indicators, control groups and policy assessment', Cambridge Working Papers in Economics 2114. https://doi.org/10.17863/CAM.65417.

Harvey, A. and Chung, C. (2000), 'Estimating the underlying change in unemployment in the UK, with discussion', *Journal of the Royal Statistical Society: Series A*, **163**, pp. 303–39.

Harvey, A. and Kattuman, P. (2020a), 'Time series models based on growth curves with applications to forecasting coronavirus', Harvard Data Science Review, Special Issue 1—COVID-19, available online at https://hdsr.mitpress.mit.edu/pub/ozgjx0yn.

Harvey, A. and Kattuman, P. (2020b), 'A farewell to R: Time series models for tracking and forecasting epidemics', CEPR Working paper, Issue 51, 7 October, available online at https://cepr.org/content/covid-economics.

Höhle, M. and an der Heiden, M. (2014), 'Bayesian nowcasting during the STEC O104:H4 outbreak in Germany', *Biometrics*, **70**, pp. 993–1002.

Ioannidis, J., Cripps, S. and Tanner, M. (2020), 'Forecasting for COVID-19 has failed', *International Journal of Forecasting*, in press. 10.1016/j.ijforecast.2020.08.004.

Koopman, S., Lit, R. and Harvey, A. (2020), *STAMP 8.4: Structural Time Series Analyser, Modeller and Predictor*. 5th ed., London: Timberlake Consultants.

Lit, R., Koopman, S.J. and Koopman, A.C. (2020), 'Time Series Lab - Score edition'. https://timeserieslab.com.

Wallinga, J. and Lipsitch, M. (2007), 'How generation intervals shape the relationship between growth rates and reproductive numbers', *Proceedings of the Royal Society B*, **274**, pp. 599–604.

Appendix A. Data sources

The data for the UK were obtained from Public Health England's (PHE) Coronavirus toolkit (https://coronavirus.data.gov.uk/developers-guide). Archived, or, real-time data are currently not available in the first release, 'v1', of their application programming interface (API). Discussion of access to archived data is summarised in this GitHub ticket: https://github.com/publichealthengland/coronavirus-dashboard/issues/241. Towards the latter part of 2020 PHE released an experimental version, 'v2', of their API which has archived data from 12 August 2020 onwards. All archived data are taken from this endpoint: https://api.coronavirus.data.gov.uk/v2. It is worth restating their disclaimer that this is an experimental endpoint and 'subject to active development and may become unstable or unresponsive without prior notice'.

Appendix B. Estimation for bivariate model for publication and specimen data series

The prediction error variances for specimen and published were 0.0134 and 0.0172, respectively, with a correlation of 0.077. The slope variances were constrained to be the same, and q was set at 0.015 for the specimen series. The estimated AR coefficient in the published series was 0.672, and its variance was 0.0056. The irregular variances for specimen and published were 0.0082 and 0.0091, respectively, with a correlation of −0.190.

Figure 8 shows the residual autocorrelation functions (ACFs) and histograms. The diagnostic statistics were[5] as follows for specimen: $r(1) = 0.20$, $Q(18) = 25.41$, $BS = 9.87$ and $H = 3.95$, and for

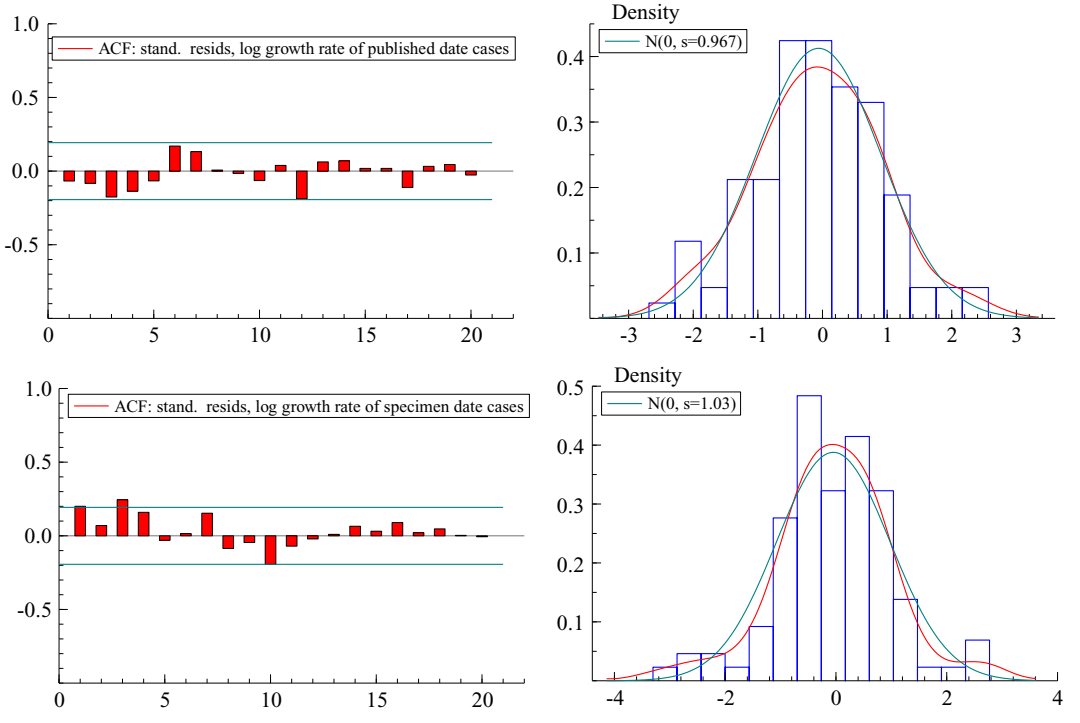

Figure 8. (Colour online) Residuals from bivariate model fitted to data up to 19 January 2021

[5]r(1) is the autocorrelation at lag one, Q(P) is Box–Ljung statistic with P autocorrelations, BS is the Bowman–Shenton normality statistic and H is a heteroscedasticity statistic constructed as the ratio of the sum of squares in the last third of the sample to the sum of squares in the first third.

https://doi.org/10.1017/nie.2021.12 Published online by Cambridge University Press

published: $r(1) = -0.07$, $Q(18) = 20.21$, $BS = 0.38$ and $H = 0.90$. There is some remaining serial correlation, but not a great deal. Fitting an AR(1) to the specimen series as well as to the published series may reduce $r(1)$. The greater stability in the published series is reflected in the smaller BS normality test statistic.

The output for the state vector shows that the slopes on 18th January are almost identical for the two series; for specimen data, it is -0.0506 (0.0260), and for published, it is -0.0502 (0.0261). The difference arises because, although STAMP is able to constrain the variances of the slopes to be the same, it is currently unable to set the deterministic parts of the slope to be equal.

Cite this article: Harvey, A., Kattuman, P. and Thamotheram, C. (2021), 'Tracking the mutant: Forecasting and nowcasting COVID-19 in the UK in 2021', *National Institute Economic Review*, **256**, pp. 110–126. https://doi.org/10.1017/nie.2021.12