

**Harvard Data Science Review • Special Issue 1: COVID-19:
Unprecedented Challenges and Chances**

Time Series Models Based on Growth Curves With Applications to Forecasting Coronavirus

Andrew Harvey¹ Paul Kattuman²

¹Faculty of Economics, University of Cambridge, Cambridge, England, United Kingdom,

²Cambridge Judge Business School, University of Cambridge, Cambridge, England, United Kingdom

Published on: Aug 25, 2020

DOI: <https://doi.org/10.1162/99608f92.828f40de>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

ABSTRACT

Time series models are developed for predicting future values of a variable that when cumulated is subject to an unknown saturation level. Such models are relevant for many disciplines, but here attention is focused on the spread of epidemics and the applications are for coronavirus. The time series models are relatively simple but are such that their specification can be assessed by standard statistical test procedures. In the generalized logistic class of models, the logarithm of the growth rate of the cumulative series depends on a time trend. Allowing this trend to be time-varying introduces further flexibility and enables the effects of changes in policy to be tracked and evaluated.

Keywords: generalized logistic, Gompertz curve, Kalman filter, negative binomial distribution, score-driven models, stochastic trend

Media Summary

As an epidemic takes hold, reliable projections of its path and of its eventual size are important to health care providers and policy makers, in order to help them meet the challenge of organizing medical resources to meet treatment requirements. The many unknowns surrounding the COVID-19 virus make it particularly important and urgent to have reliable estimates of the future course of the daily numbers of new cases and deaths.

Epidemiological models that seek to estimate epidemic trajectories fall into two broad classes. One class, called compartmental models, consists of models that seek to be faithful to the details of the routes and processes of disease transmission. More specifically, they project how individuals in an initially susceptible population come to be exposed to the virus, potentially infected, and, if infected, either recover from the disease or die of it. This class of models is used to understand the consequences of policies, for example, a lockdown. But many aspects of the transmission and progress of the disease are unknown, and so these models, which are inevitably complex, are typically forced to rely on a large number of assumptions.

Time series models, on the other hand, simply use historical data to make predictions. This article develops a new class of time series models that reflect epidemic trajectories even when they change over time. Unlike most standard time series models, the new models are able to make good forecasts even before new cases and/or deaths reach their peak. Furthermore they are able to track the epidemic as human behavior changes, and can be used to evaluate the effects of changes in policy. Applying these models to the study of COVID-19 dynamics in the United Kingdom and Germany, we illustrate their relative simplicity and transparency, as well as their ability to produce good forecasts. Although the focus of the article is on the spread of epidemics, and specifically COVID-19, the models are applicable in a wide range of disciplines.

1. Introduction

The progress of an epidemic typically starts off with the number of cases following an exponential growth path. Over time the growth rate falls and the total number of cases approaches a final level—the ‘leveling of the curve.’ Complex behavioral modeling of the progress of the disease, for example, by using the ‘semi-mechanistic Bayesian hierarchical model’ implemented by the team at Imperial College London, depends on many assumptions and unknowns; see [Flaxman et al. \(2020\)](#). Simple and transparent time series models may offer an alternative way of making predictions of the trajectory of the epidemic; see, for example, section 2 in [Chowell et al. \(2016\)](#), where such approaches are called “phenomenological.” Similar issues arise in economics where there is a contrast between calibrated and Bayesian models based on economic theory and data-based time series models. [Avery et al. \(2020\)](#) give an excellent review of these issues.

The progress toward an upper bound or saturation level can be taken on board with a sigmoid curve, such as the logistic

$$(1.1) \quad \mu(t) = \bar{\mu} / (1 + \gamma_0 e^{-\gamma t}), \quad \gamma_0, \gamma, \bar{\mu} > 0, \quad -\infty < t < \infty,$$

where $\bar{\mu}$ is the final level, γ is a rate of progress parameter, γ_0 takes account of the initial conditions, and e is Euler’s number. Logistic curves can be shown to arise from a model of a simple epidemic; see, for example, chapter 2 in [Daley & Gani \(1999\)](#). More generally, sigmoid curves are used in many disciplines for a variety of applications, such as estimating the demand for new products and population growth of mammals subject to space and resource limitations. An early and influential discussion of growth curves was given by [Gregg et al. \(1964\)](#). More recently [Panik \(2014\)](#) describes growth curves and statistical methods for fitting them. Here we concentrate on an extension of the logistic curve called the generalized logistic (GL) or Richards curve; see pp. 78-80 in [Panik \(2014\)](#). The Gompertz curve emerges as an important special case. Figure 1 shows two Gompertz growth curves together with the corresponding changes as given by the first derivatives. The effect of what turns out to be a key parameter, γ , on the peak is evident.

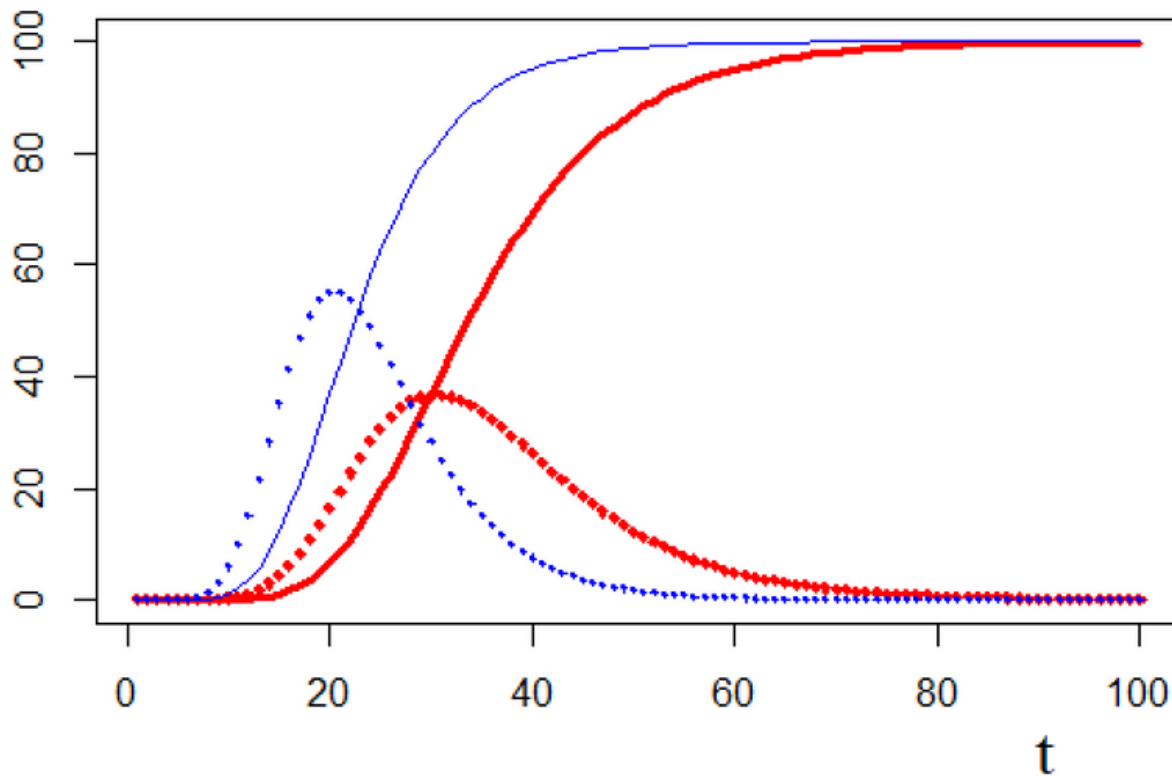


Figure 1. Gompertz functions with final level of 100. The solid red curve (bold) has a point of inflection equal to 30, with $\gamma_0 = 20$ and $\gamma = 0.1$. The solid blue curve, to the left, has a point of inflection equal to 20, with $\gamma_0 = 20$ and $\gamma = 0.15$. The dotted lines show the corresponding change curves multiplied by 10.

By formulating a statistical model, parameters such as γ_0 , $\bar{\mu}$ and γ can be estimated from observations, denoted Y_t , made at discrete times $t = 1, \dots, T$, on $\mu(t)$. One option is to work directly with the level, by basing a time series model on a deterministic trend, as in (1.1); for example, [Meade & Islam \(1995\)](#) report fitting a variety of growth curves to telecommunications data, while [Ciufolini & Paolozzi \(2020\)](#) report fitting a deterministic trend to total coronavirus cases in Italy. The methods discussed in [Panik \(2014\)](#) are restricted to deterministic curves, possibly with a first-order autoregressive disturbance term. Our view is that, as with most economic and social time series, a deterministic trend fitted to the level is too inflexible. Instead we prefer to work with the change or the growth rate, with specification of the model informed by an assumption that the total follows a growth curve. The saturation level can be continually updated as new observations become available. A second advantage of working with the difference or the growth rate is that when logarithms are taken, estimation of the basic models derived from the GL class can be carried out by least squares regression, as proposed in [Harvey \(1984\)](#). Thus, the method is viable even for a small number of observations. Finally the deterministic time trend in the estimating equations can be replaced by a stochastic one. This is particularly effective with a Gompertz function. The flexible trend, which can be estimated by the Kalman filter, as in the STAMP package of [Koopman et al. \(2020\)](#), allows parameters to evolve over time. As a result, the model can

adapt to significant events and changes in policy. Figure 2 shows predictions made for new cases of coronavirus in Germany. The predictions, which include a day of the week effect, were made using data up to the end of March 2020 and show a downward trend even though the series was only just reaching what subsequently turned out to be its peak. Moving beyond the peak signals that R_t , the effective reproduction number at time t , has dipped below one; see [Flaxman et al. \(2020\)](#) and [Aronson et al. \(2020\)](#).

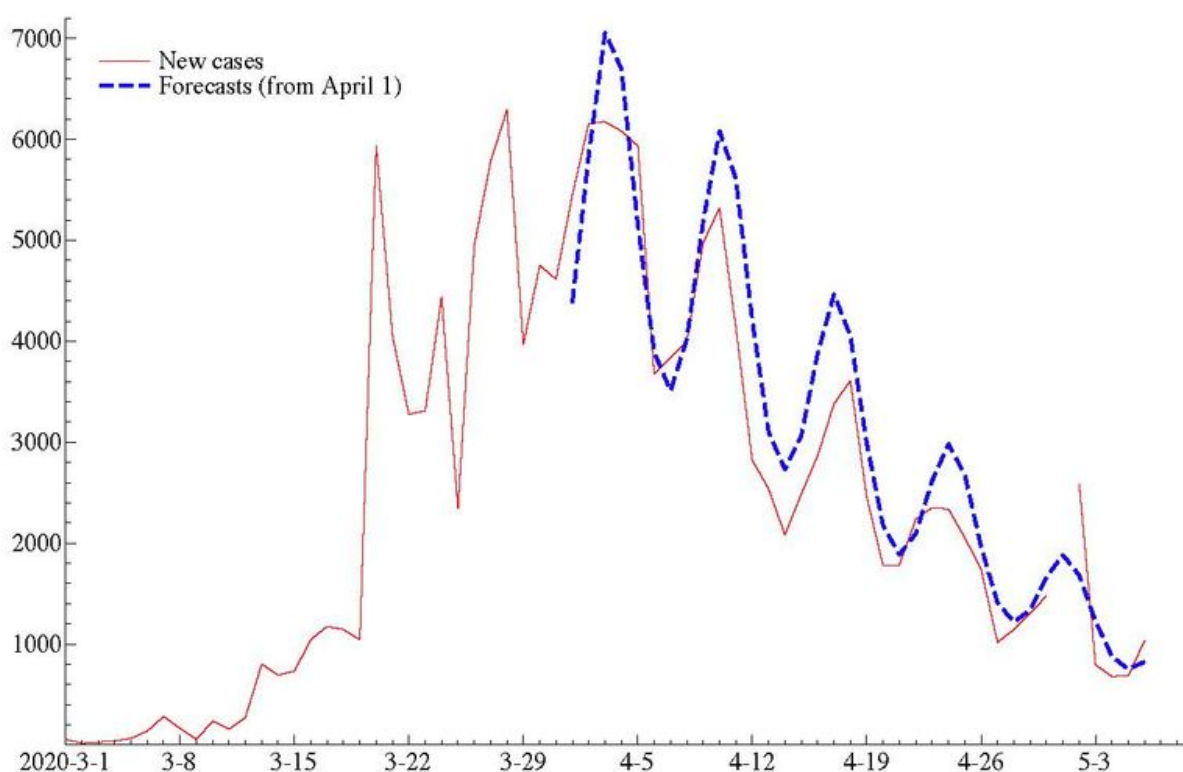


Figure 2. Forecasts (dashed - blue) for new cases in Germany using data (solid red) up to March 31. There was no observation reported for May 1. Data for May 2 can be regarded as the sum of observations for May 1 and 2.

When numbers are small, as is the case with deaths at the beginning or end of an epidemic, there is a strong argument for adopting a negative binomial distribution. Models formulated under an assumption of Gaussianity need to be modified accordingly and we show how this may be done using the score-driven approach described in [Harvey \(2013\)](#) and implemented in the TSL computer package of [Lit et al. \(2020\)](#).

Section 2 describes the GL class of growth curves, while Section 3 discusses estimation. These methods are then applied in Section 4 to data on coronavirus in U.K. hospitals and in Germany. Forecasts are set out and evaluated. The effects of policy, primarily the lockdown imposed on March 21 in the United Kingdom, are assessed in Section 5. There is concern about a potential second wave of infections as restrictions start to be eased and ways in which this might be monitored are discussed in Section 6. Section 7 concludes.

2. Growth Curves

Let $\mu(t) \geq 0$ be a monotonically increasing function defined over the real line. The rate of change or ‘incidence curve’ is $d\mu(t)/dt \geq 0$. The generalized logistic is

$$(2.1) \quad \mu(t) = \bar{\mu} / (1 + (\gamma_0 / \kappa) e^{-\gamma t})^\kappa, \quad \gamma_0, \gamma, \kappa > 0,$$

where γ is a growth rate parameter. The parameter κ must be positive for there to be an upper asymptote; allowing κ to be negative gives the class of general modified exponential (GME) growth curves. The logistic is obtained by setting $\kappa = 1$, while letting $\kappa \rightarrow \infty$ yields the Gompertz curve. When γ_0 is determined by the value of the curve at $t = 0$, it is

$$(2.2) \quad \gamma_0 = \kappa [(\bar{\mu} / \mu(0))^{1/\kappa} - 1].$$

Differentiation yields

$$(2.3) \quad \ln d\mu(t)/dt = \rho \ln \mu(t) + \delta - \gamma t,$$

where $\delta = \ln(\gamma_0 \gamma / \bar{\mu}^{1/\kappa})$ and $\rho = (\kappa + 1)/\kappa$, so $0 < \kappa < \infty$ implies $1 < \rho < \infty$. Alternatively, because $d\mu(t)/dt = g(t)\mu(t)$,

$$(2.4) \quad \ln g(t) = (\rho - 1) \ln \mu(t) + \delta - \gamma t,$$

where $g(t)$ is the growth rate of $\mu(t)$. Note that $\rho - 1 = 1/\kappa$.

The generalized logistic differential equation implied by Equation 2.1 is

$$(2.5) \quad \frac{d\mu(t)}{dt} = \gamma \kappa \left[1 - \left(\frac{\mu(t)}{\bar{\mu}} \right)^{1/\kappa} \right] \mu(t).$$

The term in square brackets is less than one and tends to zero as $\mu(t) \rightarrow \bar{\mu}$. The growth rate implied by Equation 2.5 is

$$(2.6) \quad g(t) = \gamma \kappa \left[1 - \left(\frac{\mu(t)}{\bar{\mu}} \right)^{1/\kappa} \right].$$

When $\kappa = 1$, Equation 2.5 is a Riccati equation.

2.1. Where Is the Peak?

The point of inflection on the growth curve is the point at which the number of new cases, $d\mu(t)/dt$, peaks. Differentiating the logarithm of $d\mu(t)/dt = g(t)\mu(t)$ yields the condition

$$(2.7) \quad g(t) = -g_g(t),$$

where $g_g(t)$ is the growth rate of the growth rate. It follows from Equation 2.4 that the point of inflection in the GL occurs when $g(t) = \gamma/\rho$. The corresponding value of $\mu(t)$ is

$$(2.8) \quad \mu^* = \bar{\mu}\rho^{-\kappa} = \bar{\mu}(\kappa/(\kappa + 1))^{\kappa}.$$

The change declines more slowly than it ascends when $\kappa > 1$; see, for example, Figure 1. When γ_0 is determined by $\mu(0)$, as in Equation 2.2, the peak is at

$$(2.9) \quad t^* = \ln \gamma_0 / \gamma, \quad \gamma_0 > 1,$$

so it comes forward as γ increases.

2.2. The Gompertz Curve

The Gompertz curve is

$$(2.10) \quad \mu(t) = \bar{\mu} \exp(-\gamma_0 e^{-\gamma t}), \quad \gamma_0, \gamma > 0, \quad -\infty < t < \infty.$$

When the GL is parametrized as in Equation 2.1, letting $\kappa \rightarrow \infty$ gives Equation 2.10. Taking the logarithm of $\mu(t)$ gives

$$(2.11) \quad \ln \mu(t) = \ln \bar{\mu} - \gamma_0 e^{-\gamma t},$$

leading to Equation 2.4 with $\rho = 1$, that is,

$$(2.12) \quad \ln g(t) = \delta - \gamma t,$$

where $\delta = \ln(\gamma_0 \gamma)$. The point of inflection is given in Equation 2.9, which corresponds to $\mu^* = \bar{\mu}/e \approx 0.368\bar{\mu}$. The maximum value of the change is $0.368\gamma\bar{\mu}$, while $g(t) = \gamma$. Letting $\kappa \rightarrow \infty$ in Equation 2.8, so that $\rho \rightarrow 1$, also gives $\mu^* = \bar{\mu}/e$.

2.3. Statistical Distributions and Epidemics

Writing a GL growth curve as $\mu(t) = \bar{\mu}F(t)$ allows $F(t)$ to be interpreted as the cumulative distribution function (CDF) of the log of a Dagum distribution (see pp. 212-13 in [Kleiber & Kotz, 2003](#)). Hence, the corresponding probability density function (PDF), $f(t)$, is a special case of an Exponential Beta of the Second Kind (EGB2) distribution (see [McDonald & Xu, 1995](#)).

The Gompertz distribution written using Equation 2.10 has $f(t) = \gamma\gamma_0 F(t) \exp(-\gamma t)$. Figure 1 shows two Gompertz PDFs with the red bold dotted curve corresponding to $\gamma_0 = 20$ and $\gamma = 0.1$ and the blue dotted curve corresponding to $\gamma_0 = 20$ and $\gamma = 0.15$. The effect of increasing γ is to raise the peak and bring it forward. Since $d\mu(t)/dt = \bar{\mu}f(t)$ the peak is brought down by a lower $\bar{\mu}$.

When estimating a deterministic growth curve, some researchers, for example, [Ciufolini & Paolozzi \(2020\)](#), prefer to use a sigmoid defined in terms of the Gaussian error function, which is the CDF of a Gaussian variate.

Similarly, [Murray \(2020\)](#) use the Gaussian error function to model the logarithm of total coronavirus cases in U.S. states. It is difficult to see why the Gaussian error function, which has no closed form, might be preferred to the more flexible GL curve.

It follows from Equation 2.5 that

$$(2.13) \quad d\mu(t)/dt = \bar{\mu}F(t)(1 - F(t)^{1/\kappa}),$$

where $\mu(t) = \bar{\mu}F(t)$. In a simple epidemic, $d\mu(t)/dt$ is proportional to a logistic growth curve, $F(t)(1 - F(t))$. Allowing κ to be other than one gives more flexibility and is a useful generalization if the model provides a good fit to the data. Indeed, complex mechanistic models of epidemics, with the population assigned to compartments labeled susceptible, infected, and recovered (SIR), often produce incidence curves that are positively skewed. An example is the model of [Giordano et al. \(2020\)](#), which is based on a system of eight differential equations.

3. Statistical Modeling

In the observational model, the cumulative total at time $t - 1$ replaces $\mu(t)$ and the (positive) change, $y_t = \Delta Y_t = Y_t - Y_{t-1}$, replaces $d\mu(t)/dt$. Equation 2.3 leads to

$$(3.1) \quad \ln y_t = \rho \ln Y_{t-1} + \delta - \gamma t + \varepsilon_t, \quad \rho \geq 1, \quad t = 2, \dots, T,$$

where the disturbance terms ε_t are assumed to be independently and identically distributed with mean zero and constant variance, σ_ε^2 , that is, $\varepsilon_t \sim \text{IID}(0, \sigma_\varepsilon^2)$. Subtracting $\ln Y_{t-1}$ from both sides gives the form corresponding to Equation 2.4, namely,

$$(3.2) \quad \ln g_t = (\rho - 1) \ln Y_{t-1} + \delta - \gamma t + \varepsilon_t,$$

where $g_t = y_t/Y_{t-1}$, although it may also be defined as $\Delta \ln Y_t$. The parameters ρ , δ and γ can therefore be estimated by regression. If ρ takes a specified value, $\ln y_t - \rho \ln Y_{t-1}$ is simply regressed on a constant and time trend. The estimators of δ and γ are then efficient when the disturbance is Gaussian, that is, $\varepsilon_t \sim \text{NID}(0, \sigma_\varepsilon^2)$.

The observational model for the Gompertz curve is

$$(3.3) \quad \ln y_t = \ln Y_{t-1} + \delta - \gamma t + \varepsilon_t, \quad t = 2, \dots, T,$$

or the simple time trend regression

$$(3.4) \quad \ln g_t = \delta - \gamma t + \varepsilon_t, \quad t = 2, \dots, T.$$

3.1. A Time-Varying Trend: The Dynamic Gompertz Model

A stochastic trend may be introduced into Equation 3.4 to give the dynamic Gompertz model

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

where

$$(3.5) \quad \begin{aligned} \delta_t &= \delta_{t-1} - \gamma_{t-1} + \eta_t, & \eta_t &\sim \text{NID}(0, \sigma_\eta^2), \\ \gamma_t &= \gamma_{t-1} + \zeta_t, & \zeta_t &\sim \text{NID}(0, \sigma_\zeta^2), \end{aligned}$$

and the normally distributed irregular, level and slope disturbances, ε_t , η_t and ζ_t , respectively, are mutually independent. When $\sigma_\eta^2 = \sigma_\zeta^2 = 0$, the trend is deterministic, that is, $\delta_t = \delta - \gamma t$ with $\delta = \delta_0$. On the one hand, when only σ_ζ^2 is zero, the slope is fixed and the trend reduces to a random walk with drift. On the other hand, allowing σ_ζ^2 to be positive, but setting $\sigma_\eta^2 = 0$ gives an integrated random walk (IRW) trend, which when estimated tends to be relatively smooth. The degree of smoothness depends on the signal-noise ratio, $q_\zeta = \sigma_\zeta^2 / \sigma_\varepsilon^2$.

The STAMP package of [Koopman et al. \(2020\)](#) can estimate a stochastic trend using techniques based on state space models and the Kalman filter. The Kalman filter outputs 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, which is the estimate of the state at time t based on all T observations in the series. Estimation of the unknown variance parameters is by maximum likelihood (ML). Tests for normality and residual serial correlation are based on the standardized innovations, that is, one-step ahead prediction errors, $v_t = y_t - \delta_{t|t-1}$, $t = 3, \dots, T$.

A stochastic trend can be introduced into the more general GL model. However, unless ρ is fixed, it may be hard to estimate in small samples.

The Kalman filter can be bypassed by adopting the reduced form, which comes from the innovations form of the Kalman filter; see pp. 3, 71 in [Harvey \(2013\)](#). For the GL curve, the steady-state innovations form is

$$(3.6) \quad \ln y_t = \rho \ln Y_{t-1} + \delta_{t|t-1} + v_t, \quad t = 3, \dots, T,$$

or

$$(3.7) \quad \ln g_t = (\rho - 1) \ln Y_{t-1} + \delta_{t|t-1} + v_t, \quad t = 3, \dots, T,$$

where

$$\begin{aligned} \delta_{t+1|t} &= \delta_{t|t-1} - \gamma_{t|t-1} + \alpha_1 v_t, \\ \gamma_{t+1|t} &= \gamma_{t|t-1} + \alpha_2 v_t. \end{aligned}$$

The filter for an IRW has $\alpha_2 = \alpha_1^2 / (2 - \alpha_1)$ with $0 \leq \alpha_1 \leq 1$; (see pp. 78-9 in [Harvey, 2013](#)). The implied value of q_ζ is

$$(3.8) \quad q_\varsigma = \frac{\alpha_1^4}{(2-\alpha_1)^2(1-\alpha_1)}.$$

The parameter α_1 thus plays a role similar to that of the signal-noise ratio, q_ς , and can be estimated by ML. When it is zero, the model reverts to Equation 3.2.

3.2. Forecasts

Forecasts of future observations and an estimate of the final level can be obtained from the predictive recursions

$$(3.9) \quad \hat{y}_{T+\ell|T} = \hat{\mu}_{T+\ell-1|T}^\rho \exp(\delta_T) \exp(-\gamma\ell),$$

$$(3.10) \quad \hat{\mu}_{T+\ell|T} = \hat{\mu}_{T+\ell-1|T} + \hat{y}_{T+\ell|T}, \quad \ell = 1, 2, \dots$$

where δ_T is the level at time T , that is, $\delta_T = \delta - \gamma T$ in Equation 3.1, and $\hat{\mu}_{T|T} = Y_T$. As $\ell \rightarrow \infty$, $\hat{\mu}_{T+\ell|T} \rightarrow \bar{\mu}$.

Alternatively,

$$(3.11) \quad \hat{g}_{T+\ell|T} = \hat{\mu}_{T+\ell-1|T}^{\rho-1} \exp(\delta_T - \gamma\ell), \quad \ell = 1, 2, \dots,$$

$$(3.12) \quad \hat{\mu}_{T+\ell|T} = \hat{\mu}_{T+\ell-1|T} (1 + \hat{g}_{T+\ell|T}),$$

so that $\hat{y}_{T+\ell|T} = \hat{g}_{T+\ell|T} \hat{\mu}_{T+\ell-1|T}$ and $\widehat{Y}_{T+\ell|T} = \hat{\mu}_{T+\ell|T}$.

For the Gompertz model, where $\rho = 1$, the forecasts for the growth rate are simply

$$(3.13) \quad \hat{g}_{T+\ell|T} = \exp(\delta_T) \exp(-\gamma\ell),$$

so Equations 3.11 and 3.12 yield

$$(3.14) \quad \hat{\mu}_{T+\ell|T} = \hat{\mu}_{T|T} \prod_{j=1}^{\ell} (1 + \exp(\delta_T) \exp(-\gamma j)).$$

A future point of inflection is given at $\ell^* = (\delta_T - \ln \gamma)/\gamma$. When $\ell \rightarrow \infty$,

$\hat{\mu}_{T+\ell|T} \simeq Y_T \exp((\exp \delta_T)/(\exp \gamma - 1))$. This follows from writing $\ln [\prod_{j=1}^{\ell} (1 + g_{T+j})]$ as $\sum_{j=1}^{\ell} \ln(1 + g_{T+j}) \simeq \sum_{j=1}^{\ell} g_{T+j}$. Higher order terms can be neglected when g_{T+j} is small, so $\sum_{j=1}^{\ell} g_{T+j} = \exp(\delta_T) \sum_{j=1}^{\ell} \exp(-\gamma j) \rightarrow \exp(\delta_T)/(\exp \gamma - 1)$ as $\ell \rightarrow \infty$.

An assumption of normality for the disturbances in Equation 3.1 implies that, conditional on information at time $t-1$, y_t is log-normal. Thus there may be a case for adding $0.5\sigma^2$ to δ_T , where σ^2 is the prediction error variance. (When there is no stochastic trend, $\sigma^2 = \sigma_\varepsilon^2$.) Predictive distributions of future observations may be obtained by simulation.

3.3. Models for the Growth Rate

A stochastic model for the logistic curve can be based on Equation 2.6, as in [Levenbach & Reuter \(1976\)](#), by adding a serially independent Gaussian disturbance term so that

$$(3.15) \quad g_t = \gamma - (\gamma/\bar{\mu})\hat{\mu}_{t|t-1} + \varepsilon_t, \quad t = 1, \dots, T,$$

where $\hat{\mu}_{t|t-1}$ is an estimate of μ_t , such as Y_{t-1} , based on information at time $t - 1$ and ε_t is a serially independent Gaussian disturbance with mean zero and constant variance, σ_ε^2 , that is, $\varepsilon_t \sim \text{NID}(0, \sigma_\varepsilon^2)$.

Regressing g_t on Y_{t-1} gives estimates of the key parameters γ and $\bar{\mu}$. Generalization of this approach is based on Equation 2.6 but leads to a nonlinear equation

$$(3.16) \quad g_t = \gamma\kappa - \left(\frac{Y_{t-1}}{\bar{\mu}}\right)^{1/\kappa} + \varepsilon_t, \quad t = 1, \dots, T,$$

which requires a search over the range of κ if it is unknown. In the Gompertz case, estimation is as straightforward as for the logistic model because

$$(3.17) \quad g_t = \gamma \ln \bar{\mu} - \gamma \ln Y_{t-1} + \varepsilon_t, \quad t = 1, \dots, T.$$

Models based on the logarithm of the growth rate are preferred because they have better statistical properties. For example, the disturbance term is less likely to be heteroscedastic.

3.4. Small Numbers: The Negative Binomial Distribution

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, which, when parameterized in terms of a time-varying mean, $\xi_{t|t-1}$, and a fixed positive shape parameter, ν , has probability mass function (PMF)

$$p(y_t | y_{t-1}, y_{t-2}, \dots) = \frac{\Gamma(\nu + y_t)}{y_t! \Gamma(\nu)} \xi_{t|t-1}^{y_t} (\nu + \xi_{t|t-1})^{-y_t} (1 + \xi_{t|t-1}/\nu)^{-\nu}, \quad y_t = 0, 1, 2, \dots,$$

with $\text{Var}_{t-1}(y_t) = \xi_{t|t-1} + \xi_{t|t-1}^2/\nu$. An exponential link function ensures that $\xi_{t|t-1}$ remains positive and at the same time yields an equation similar to Equation 3.1:

$$(3.18) \quad \ln \xi_{t|t-1} = \rho \ln Y_{t-1} + \delta - \gamma t, \quad \rho \geq 1, \quad t = 3, \dots, T.$$

A stochastic trend may be introduced into the model, as in subsection 3.2, by developing a filter for the time-varying trend similar in structure to that of Equation 3.6. Because the observations are not Gaussian, the dynamic conditional score (DCS) framework described in [Harvey \(2013\)](#) is used to give

$$(3.19) \quad \ln \xi_{t|t-1} = \rho \ln Y_{t-1} + \delta_{t|t-1}, \quad t = 3, \dots, T,$$

where $\delta_{t|t-1}$ is like the filter for the IRW in the Gaussian model, that is,

$$\begin{aligned}\delta_{t+1|t} &= \delta_{t|t-1} - \gamma_{t|t-1} + \alpha_1 u_t, & 0 \leq \alpha_1 \leq 1, \\ \gamma_{t+1|t} &= \gamma_{t|t-1} + \alpha_2 u_t, & \alpha_2 = \alpha_1^2 / (2 - \alpha_1),\end{aligned}$$

but with $u_t = y_t / \xi_{t|t-1} - 1$, which is the conditional score for $\ln \xi_{t|t-1}$, that is $v(y_t - \xi_{t|t-1}) / (v + \xi_{t|t-1})$, divided by the information quantity. The dynamic Gompertz model has $\rho = 1$. Estimation is by maximum likelihood.

Predictions of future observations and the final level can be obtained from Equations 3.9 and 3.10.

4. Forecasting Coronavirus in the United Kingdom and Germany

We began working on this project at the beginning of April 2020. At that time coronavirus was not as far advanced in the United Kingdom as in Italy, and our initial exploration was focussed on Italy. Figure 3 shows new cases, as measured by hospital admissions on the European Centre for Disease Prevention and Control (ECDC) website, in the United Kingdom and Italy; see the Appendix. Italy is seen to be clearly ahead and it is apparent that the decline is slower than the rise. This asymmetry is not consistent with a simple logistic model, Equation 1.1, and later analysis confirmed this to be the case for most other countries, including the United Kingdom.

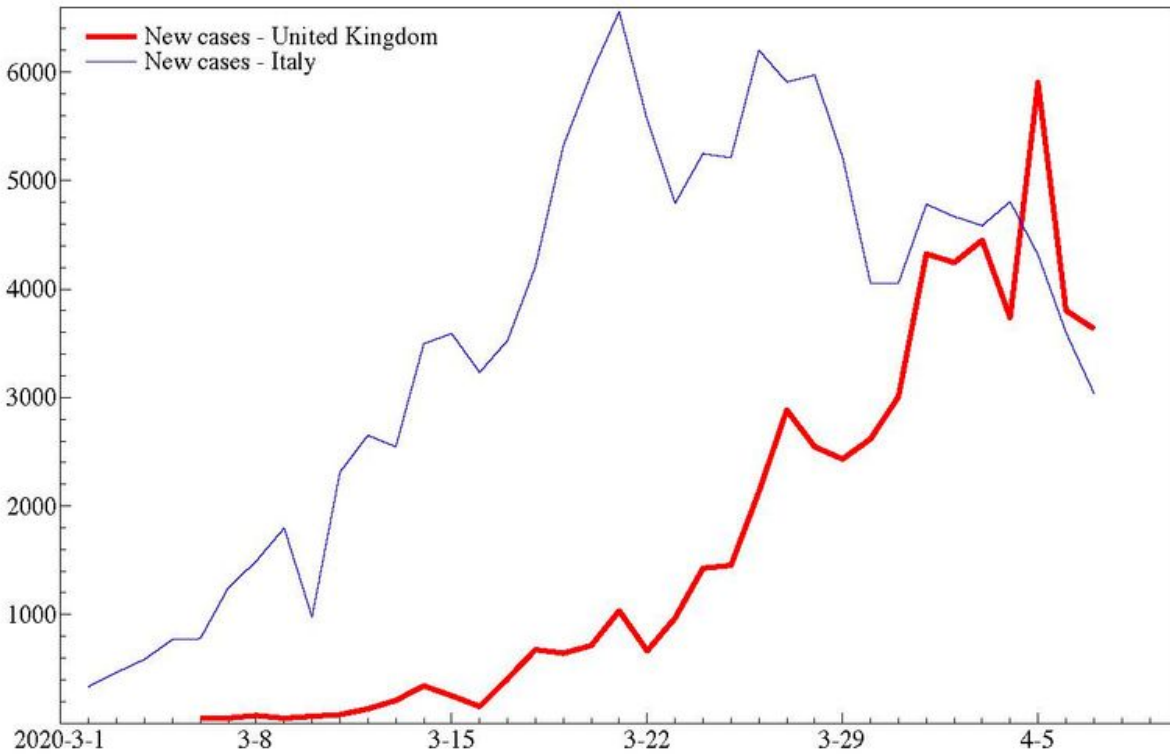


Figure 3. New cases in the United Kingdom (red, bold) and Italy (blue) up to April 10.

Economic and social time series are typically subject to periodic variation, due to seasonal, day of the week, and other effects. Preliminary analysis of data for hospital admissions and deaths in Italy indicate a day of the week pattern and this is confirmed for the United Kingdom; one reason is that laboratory confirmation for the virus tends to slow down during weekends. Day of the week effects are initially included in the equations by the introduction of a single harmonic cycle with period seven; this possesses only two parameters as opposed to the six needed for a full set of dummy variables.

Although we estimated many models in early April, the first results reported here for the United Kingdom are those obtained on April 13, that is, with data up to April 12, starting on March 5. The models were updated on April 20 and 27. Unfortunately, the data available to us after April 29 were not consistent with the earlier data as they were not confined to just tests carried out on hospital admissions.

4.1. Models Fitted to New Cases in the United Kingdom

Table 1 shows April 13 estimates for the GL, Equation 3.2, the Gompertz, where ρ is set to one, and the dynamic Gompertz where the trend is an IRW. The results for the logistic model are not reported because it gives an inferior fit and is rejected by a t test on $\hat{\rho}$; the standard error is in parentheses. The diagnostics presented are the Durbin-Watson (DW) statistic for residual serial correlation, a portmanteau Q —statistic for serial correlation based on P autocorrelations, the Bowman-Shenton normality test statistic and a heteroscedasticity statistic consisting of the squares of the last third of the residuals over the first third.

There is a substantive reduction in the prediction error variance for all models when daily effects are included. The peak day is Friday, the same day as was found for Italy. The results for the GL are shown in the penultimate column and may be compared with those in the first column. The likelihood ratio statistic is which is significant at the 5% level of a χ^2_2 distribution. Figure 4 shows the fit and residuals from the GL. Figure 5 shows the histogram and correlogram for the GL residuals but using data up to April 20.

Table 1. Estimates for U.K. coronavirus cases (hospital admissions) as of April 13 and, in the last column, as of April 20.

Statistic	GL	Gompertz	Dynamic Gompertz	GL	Dynamic Gompertz*
Y	0.081	0.033	0.054	0.091	0.061
δ_T	-5.62	-2.48	-2.63	-6.32	-3.15
ρ	0.264 (0.153)	1	1	0.323 (0.129)	1
q_c	0	0	0.001	0	0.001

$\ln L$	38.350	37.765	39.368	42.590	55.595
DW	1.62	1.43	1.67	2.10	2.01
Q(6)	7.32	5.30	8.25	Q(9)=9.38	Q(10)=9.32
Normality	0.47	0.46	1.09	1.29	1.64
Hetero F	0.37	.72	.45	.52	0.45

Note. Fixed parameters are in italics. The maximized log-likelihood is $\ln L$; note that the last column is based on a different set of data and so its likelihood value should not be compared to those in the other columns.

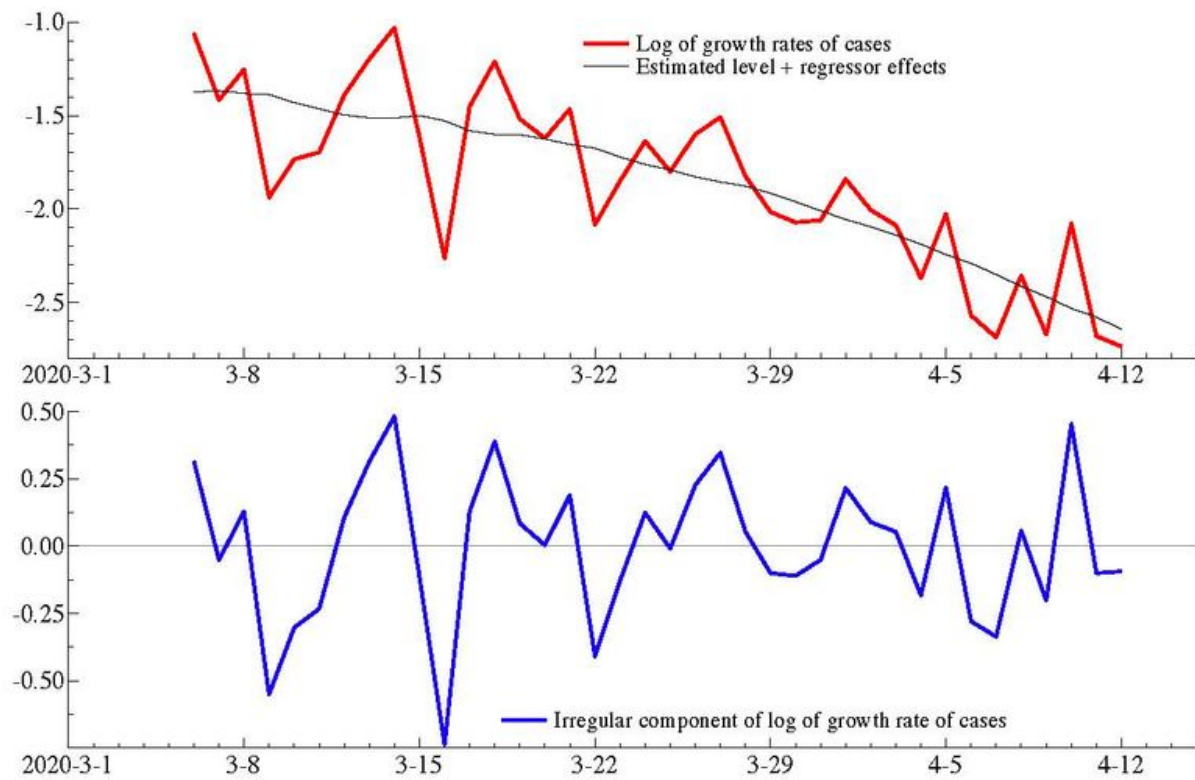


Figure 4. Fit of generalized logistic model to the logarithm of growth rates of cases in the United Kingdom (using data up to April 13).

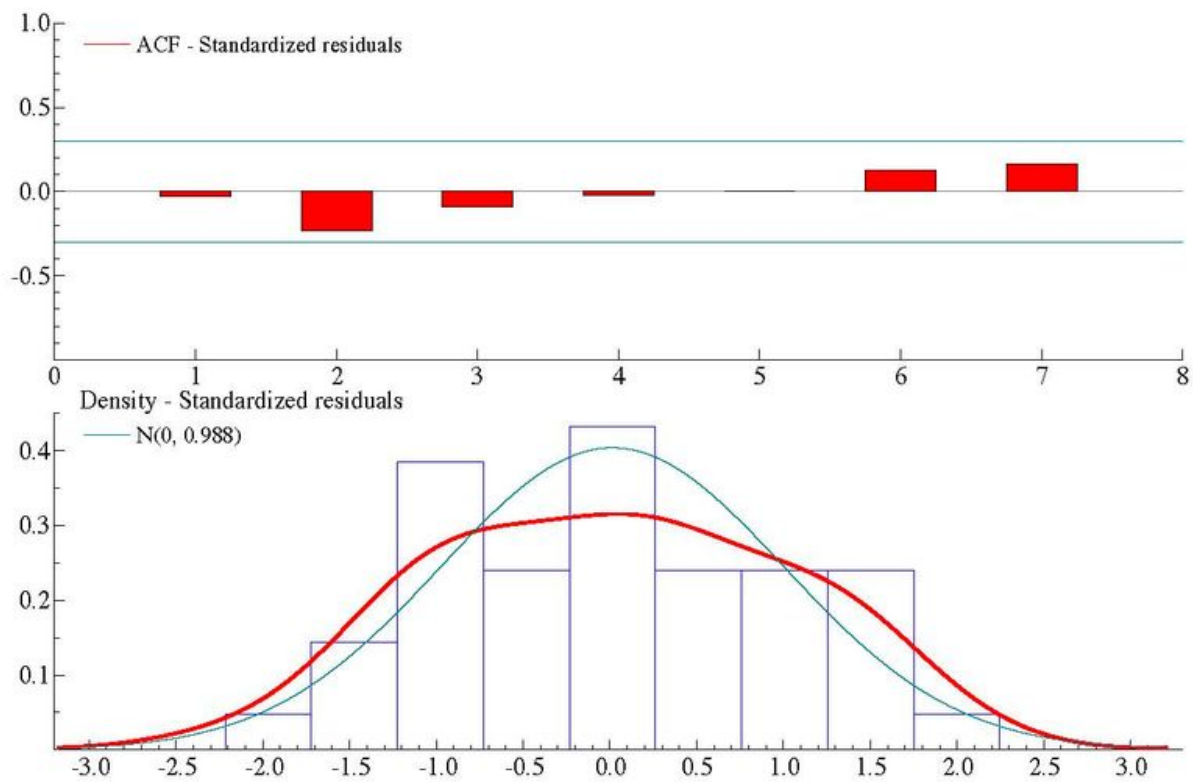


Figure 5. Residuals from generalized logistic model fitted to logarithm of growth rates of cases in the United Kingdom (using data up to April 20).

4.2. Forecasts and Forecast Evaluation

The forecasts of new cases made on April 13 are shown in Figure 6, together with the actual values up to May 29. The upper line is the dynamic Gompertz, while the lower lines are the GL with and without the daily effect. The actual outcome turns out to lie between the dynamic Gompertz and the GL. As with the German predictions shown in Figure 2, the important point is that for both models the forecasts are moving downward even though the observations have barely peaked.

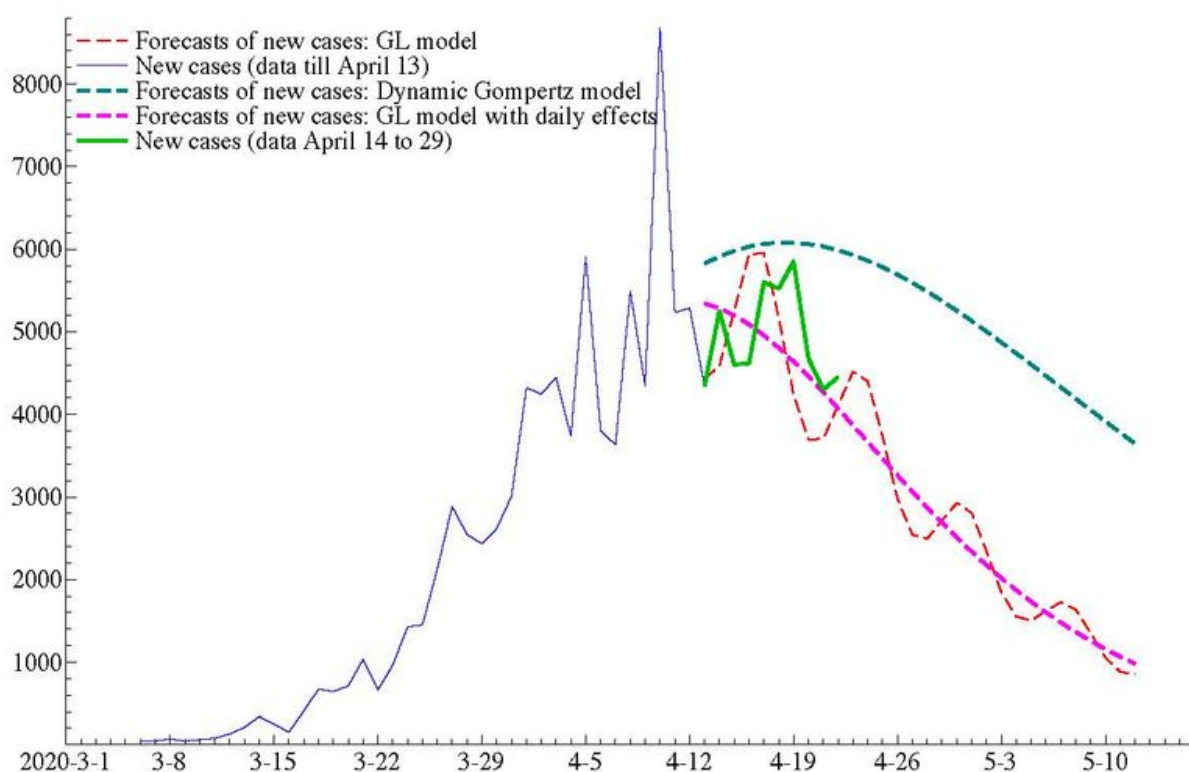


Figure 6. Forecasts (dashed) for new cases in the United Kingdom (using data up to April 13).

Predictions from the GL made on April 20 show little change from those made on April 13. The final level under current policies is estimated to be 186,000. By contrast the dynamic Gompertz predictions change significantly. They are still higher than the GL predictions, but the final level predicted (by the dynamic Gompertz model) is now 237,790, whereas on April 13 it was 308,960. The dynamic Gompertz predictions made on April 27 show little movement; the final predicted level is 253,800. Figure 7 shows the dynamic Gompertz predictions without the daily component, but the daily component was included when the models were estimated. The flexibility of the dynamic model resides in its ability to adapt to a situation in which the observations are falling less rapidly than indicated by the static GL.

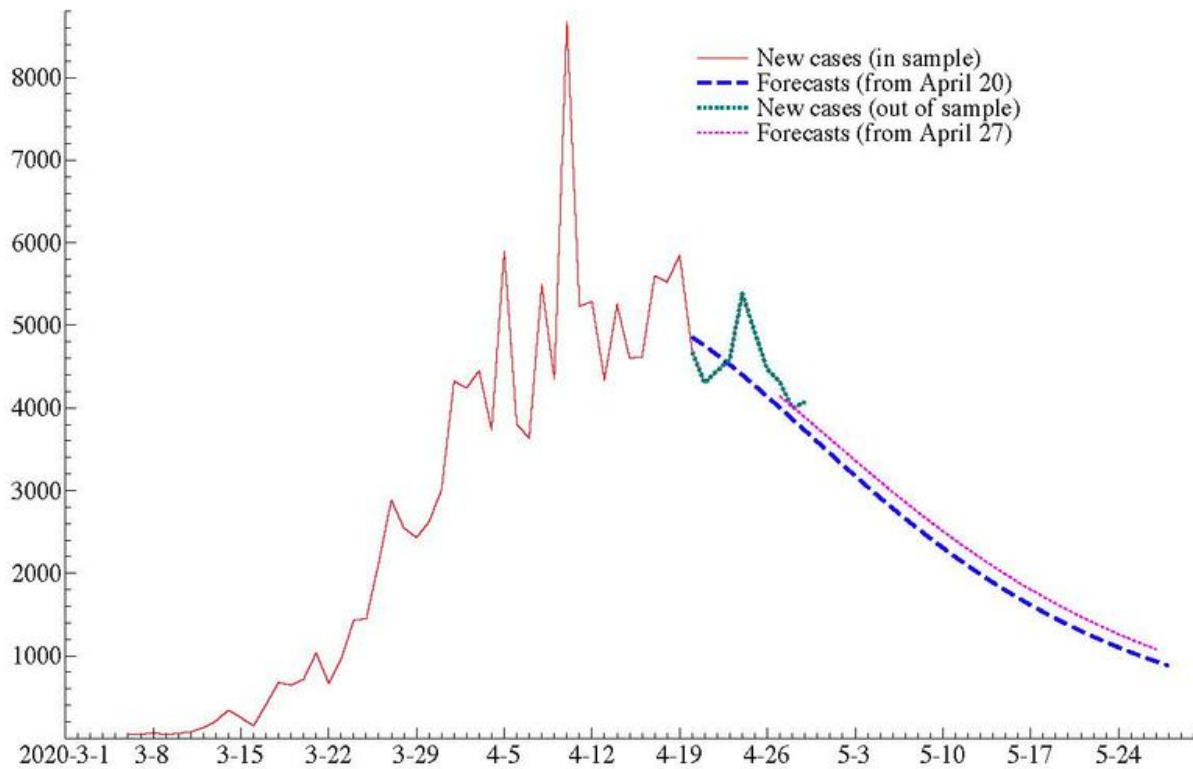


Figure 7. Forecasts (dashed) for new cases in the United Kingdom made by Dynamic Gompertz model (using data up to April 20, and up to April 27).

4.3. Germany

The data for new cases in Germany are from the Robert Koch Institut (RKI) and constitute the confirmed cases of COVID-19 in all national hospitals and testing centers.¹

Fitting with data for all of March give the results in Table 2 for a dynamic Gompertz model with daily effects. Estimation of a GL give $\tilde{\rho} = 1.21$ (0.33), so the t statistic is only 0.64. The likelihood is smaller at 16.43. There is no evidence of residual serial correlation in either model and although the heteroscedasticity statistic indicates a diminishing variance, its value seems to be heavily influenced by just one observation (March 9) near the beginning of the series. The fit is shown in Figure 8.

Table 2. Germany: Estimation results for dynamic Gompertz models with daily effects.

Statistic	Dynamic Gompertz (April 1)	Dynamic Gompertz (May 7)
γ	0.077	0.062

δ_T	-2.41	-5.15
q_c	0.0015	0.0015
$\ln L$	18.43	65.43
DW	2.02	2.01
$Q(9) + Q(11)$	10.02	16.83
Normality	0.86	6.11
Hetero F	0.42	0.22

The forecasts shown earlier in Figure 2 are quite remarkable in their accuracy over the next 36 days, that is, up to May 6; see also Figure 9. The observations had not yet started to go down on April 1, yet the sigmoid nature of the underlying growth curve means that the forecasts pick up the subsequent downward movement.

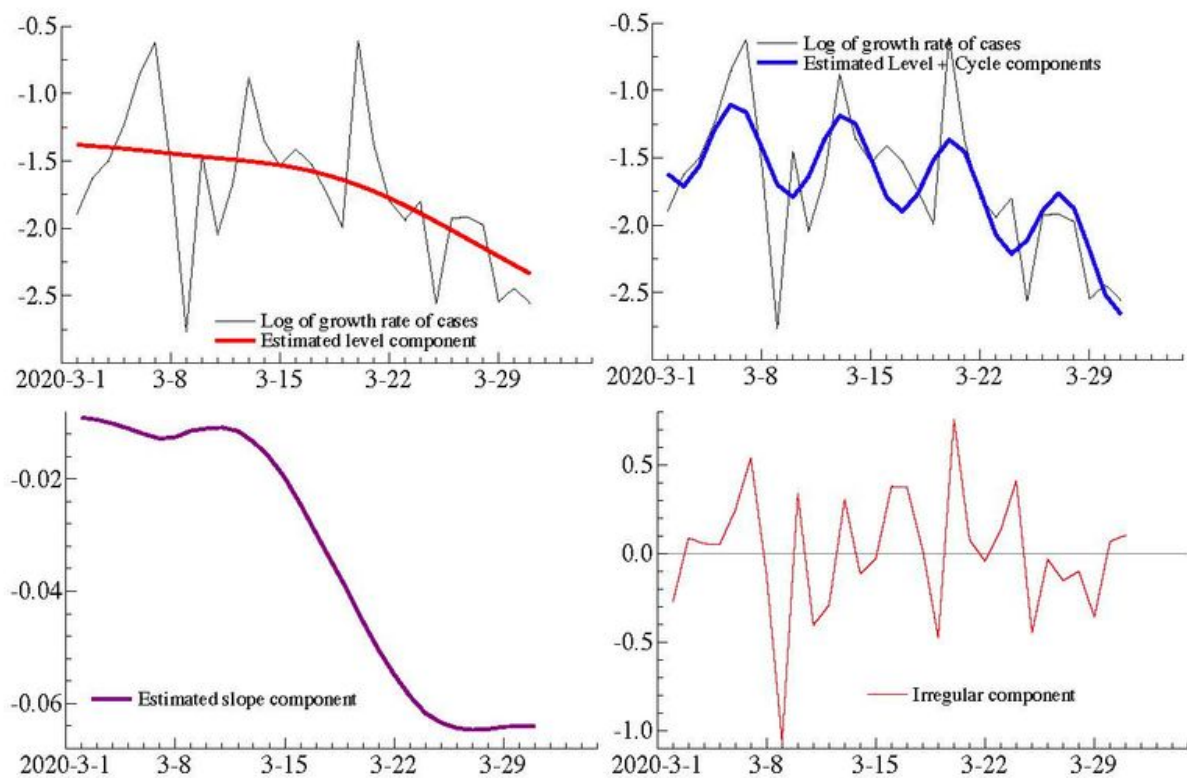


Figure 8. Forecasts for logarithm of growth rate of cases in Germany made (using data up to March 31).

Fitting the dynamic Gompertz using data up to and including May 6 shows the model to be stable; see Table 2 and Figure 10. In particular, the signal-noise ratio is almost unchanged. The observations in April are less variable than in March so the residuals are much smaller. The slope appears to fall in early April but it then returns to a value similar to what it was at the end of March. In summary, the dynamic Gompertz model works extremely well.

4.4. Deaths

Deaths in Germany behave in a way similar to new cases: Figure 11 shows the log growth rates. There are some zero values in the first three weeks of March, so the dynamic Gompertz is estimated using data from March 22. Because the numbers are smaller it is perhaps not surprising that the signal-noise ratio is estimated to be zero. In other words, we end up with the static Gompertz model.

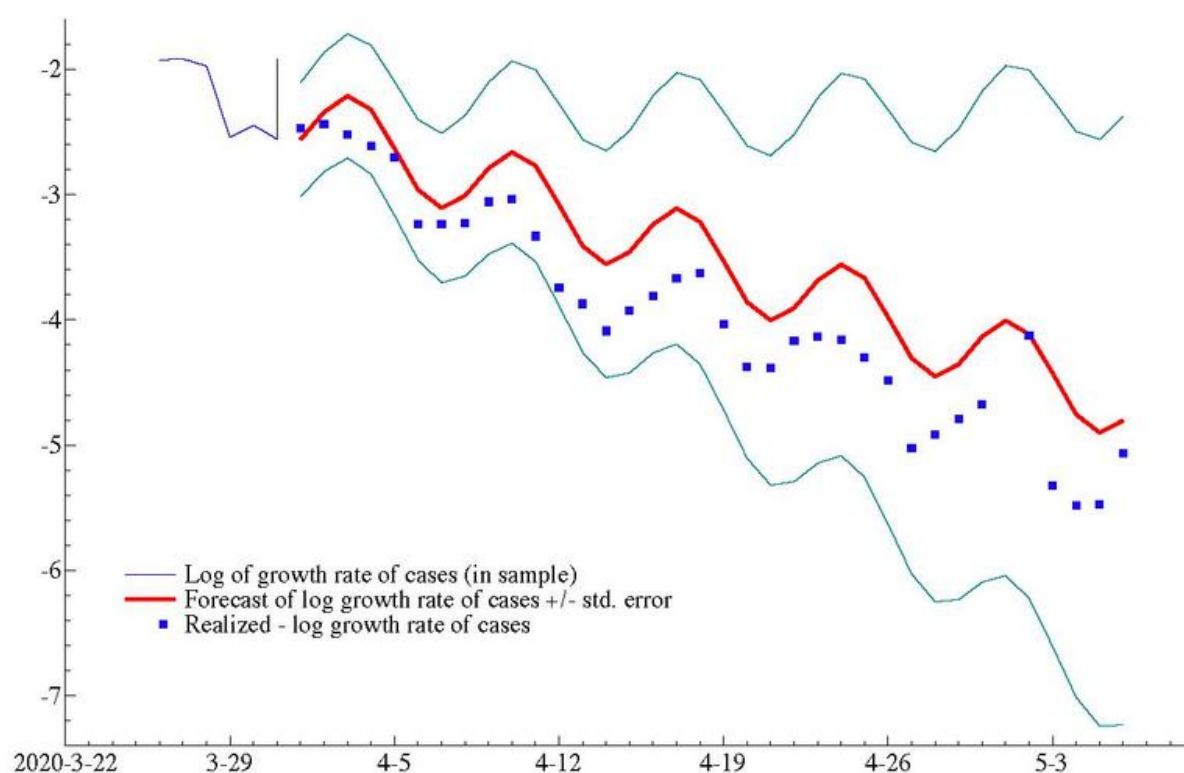


Figure 9. Forecasts for logarithm of growth rate of cases in Germany made (using data up to March 31).

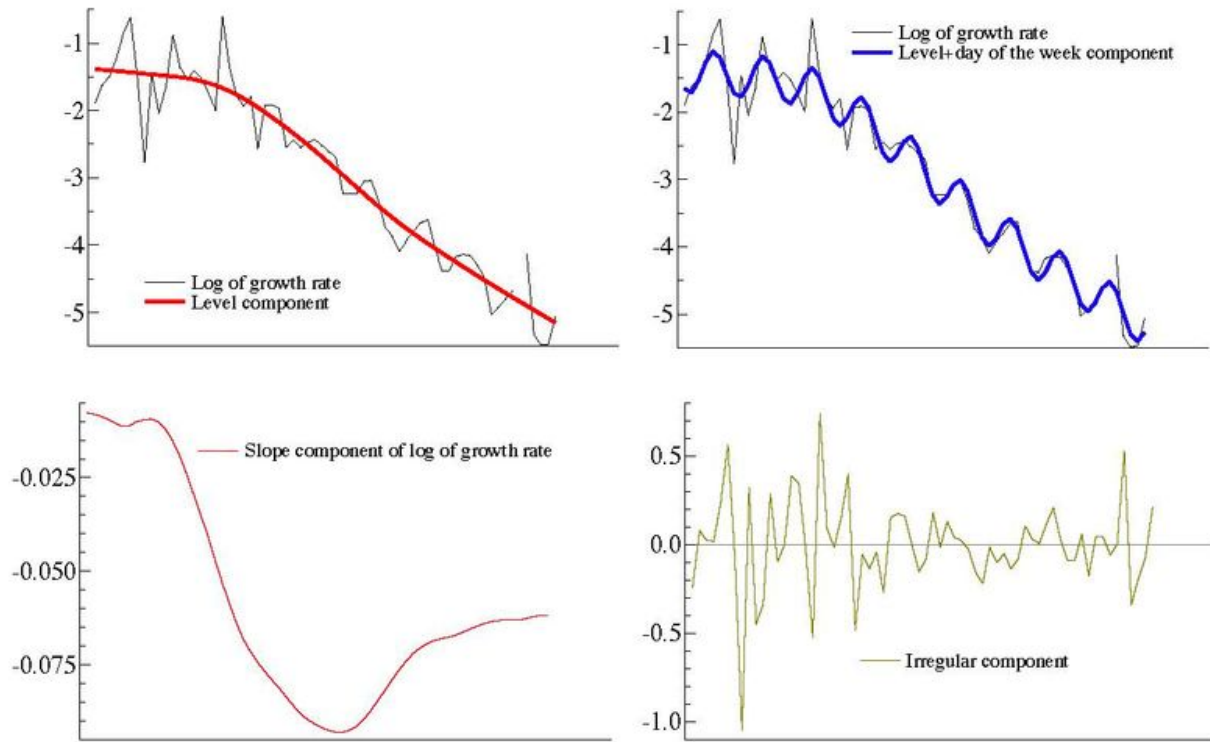


Figure 10. Smoothed estimates from the dynamic Gompertz model of the level and slope components of the logarithm of growth rate (using data up to May 6).

The score-driven negative binomial model, introduced in subsection 3.4, can be estimated with the TSL package of [Lit et al. \(2020\)](#). Observations from March 11 up to, and including, May 6, are used. There are some zero values. Setting the ρ parameter to unity gives $\tilde{\alpha} = 0$ —corresponding to a deterministic trend—and $\tilde{\gamma} = 0.071$. The correlogram of the scores has a relatively high value at lag 7, indicating a daily effect. Modeling a fixed daily effect with dummy variables produces the fit in Figure 12, with the significantly increased log-likelihood. The parameter estimates are $\tilde{\gamma} = 0.070$, $\tilde{\delta}_T = -4.14$, and $\tilde{\upsilon} = 13.25$. It is reassuring that the values of $\tilde{\gamma}$ and $\tilde{\delta}_T$ are consistent with those reported for new cases. The total number of deaths on May 6 was 6,996 and the final total is predicted to be 8,714. When we revised our article on June 19 the total was 8,946 and daily deaths had fallen to very low numbers.

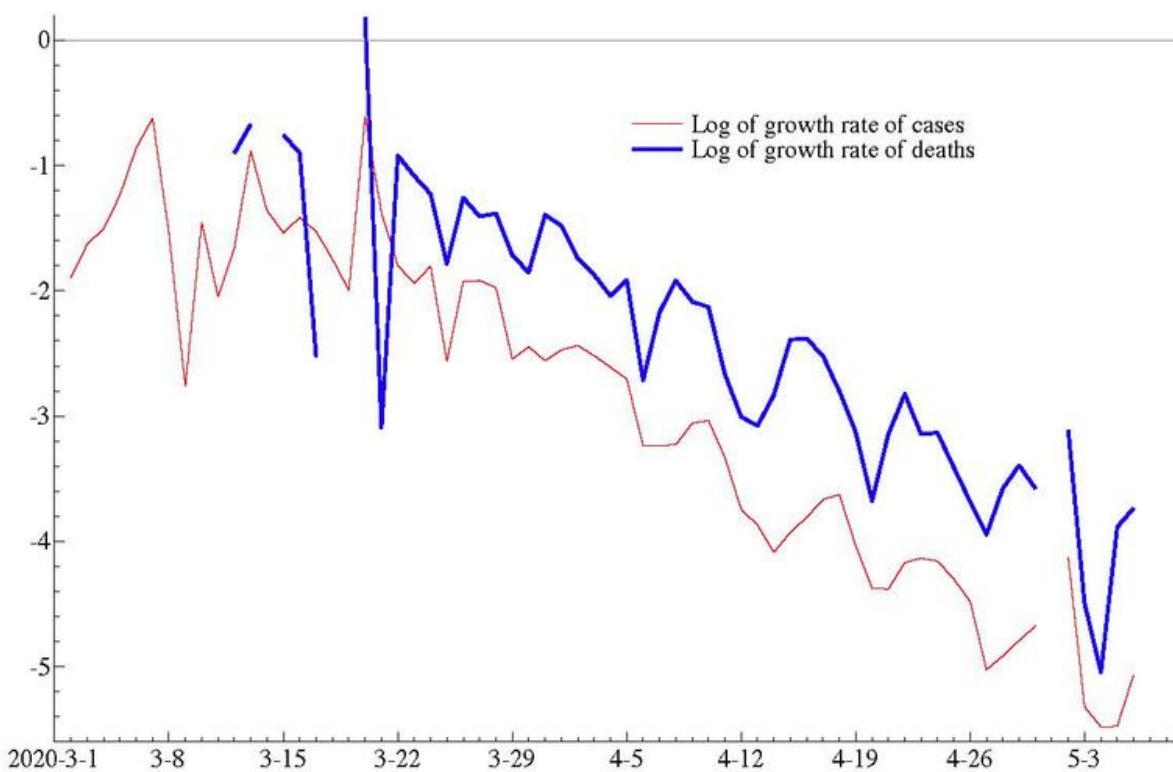


Figure 11. Logarithm of growth rate for infected cases and deaths in Germany.

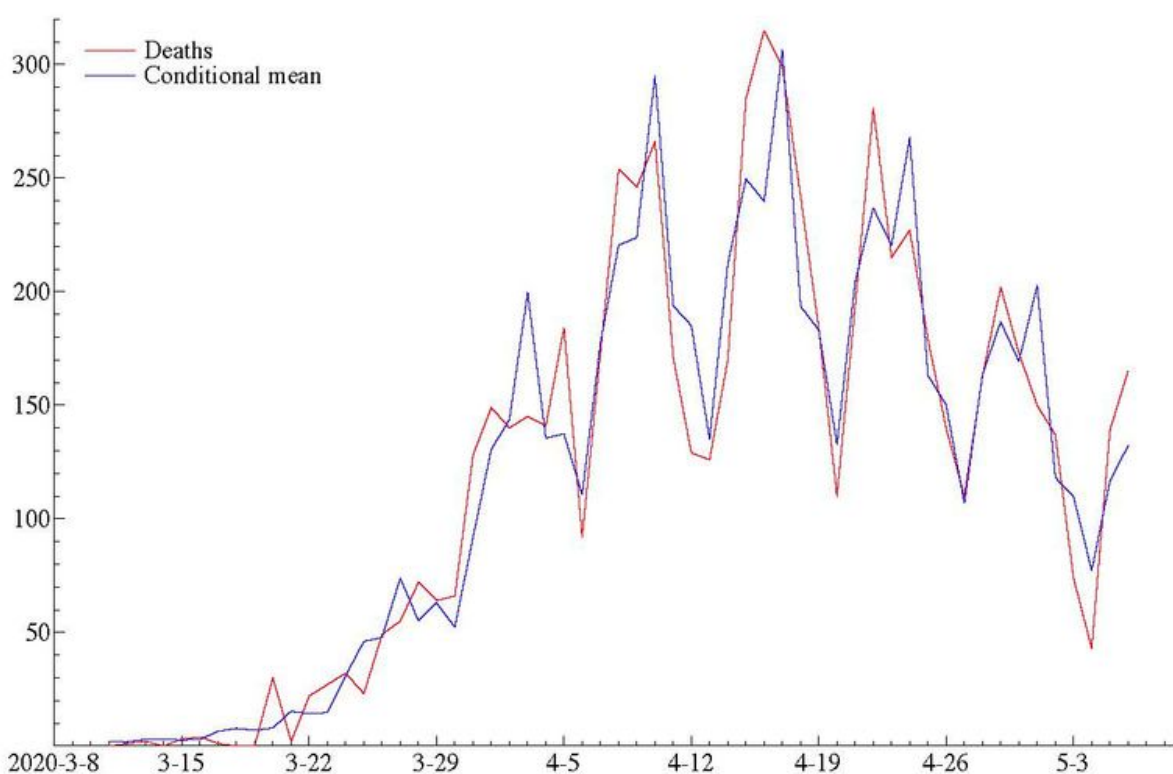


Figure 12. Deaths in Germany and negative binomial filter with day of the week effect.

Up to April 27, the U.K. data set on deaths from coronavirus was restricted to deaths in hospitals. After April 28, deaths in the community were included and the earlier figure revised. Although the relationship between deaths and new cases before April 28 is similar to that in Germany, there is a disconnect after that date and the data on new cases are of very little help in predicting deaths.

Fitting a negative binomial model, with daily effects, to the series on U.K. deaths from March 17 to May 14 lead to an estimated α of 0.34 (0.05). The changing trend is consistent with the Gaussian models fitted to new cases; the implied value of q_c is 0.007. The slope at the end of the series is 0.042, while $\tilde{\gamma} = 29.91$ (6.13). Figure 13 shows the forecasts of the underlying trend up to one month ahead. New observations after the forecasts were made are marked by dots.

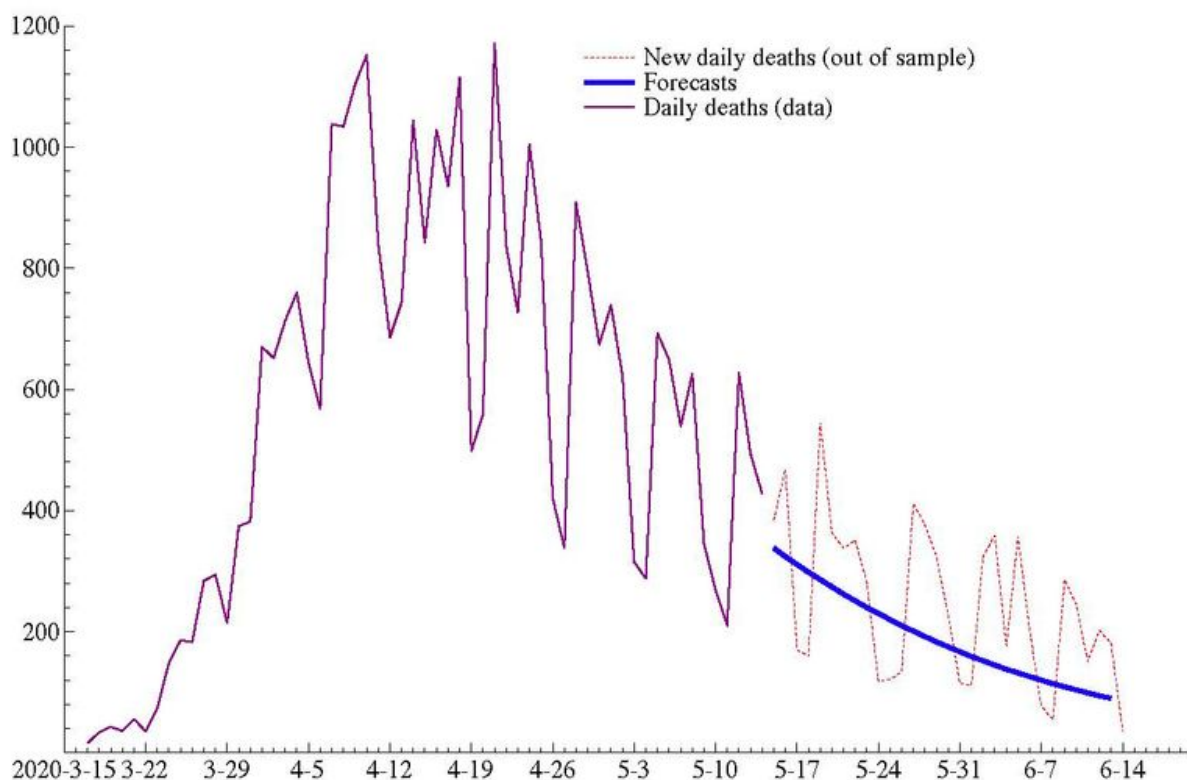


Figure 13. Forecasts (bold - blue) of daily deaths in the United Kingdom for one month ahead (using data up to and including May 14).

5. The Effect of Policy Interventions

The effect of significant interventions, which may be a policy change, such as the introduction of a lockdown, or an external event, such as the arrival of a cruise ship in a small port, may be modeled by intervention (dummy) variables. The difficulty is that the pattern of the response is rarely known and so it becomes difficult to obtain a meaningful estimate of the final effect. Nevertheless, some notion of the response can be obtained by making forecasts at the time the policy is thought to have become effective and comparing these forecasts with the actual outcome. In order to investigate this possibility, the dynamic Gompertz model was estimated on U.K. data up to and including March 31 (10 days after lockdown) using a fixed q_L of 0.001 (as estimated later with a larger sample and reported in Table 1). No day-of-the-week effect was included, because the series is rather short. Figure 14 shows there is considerable overprediction. However, if the signal-noise ratio is increased to $q_L = 0.01$, so the most recent observations receive more weight, the predictions are excellent. The final level is now approximately 290,000 as against 1.8 million. New cases are at their maximum, with a value of 5,400, on April 18, whereas with $q_L = 0.001$, they do not peak until May 20, when they reach 21,500. Although the variation in predictions is huge, the same is true of many of the large models where the output can be very sensitive to the assumptions made.

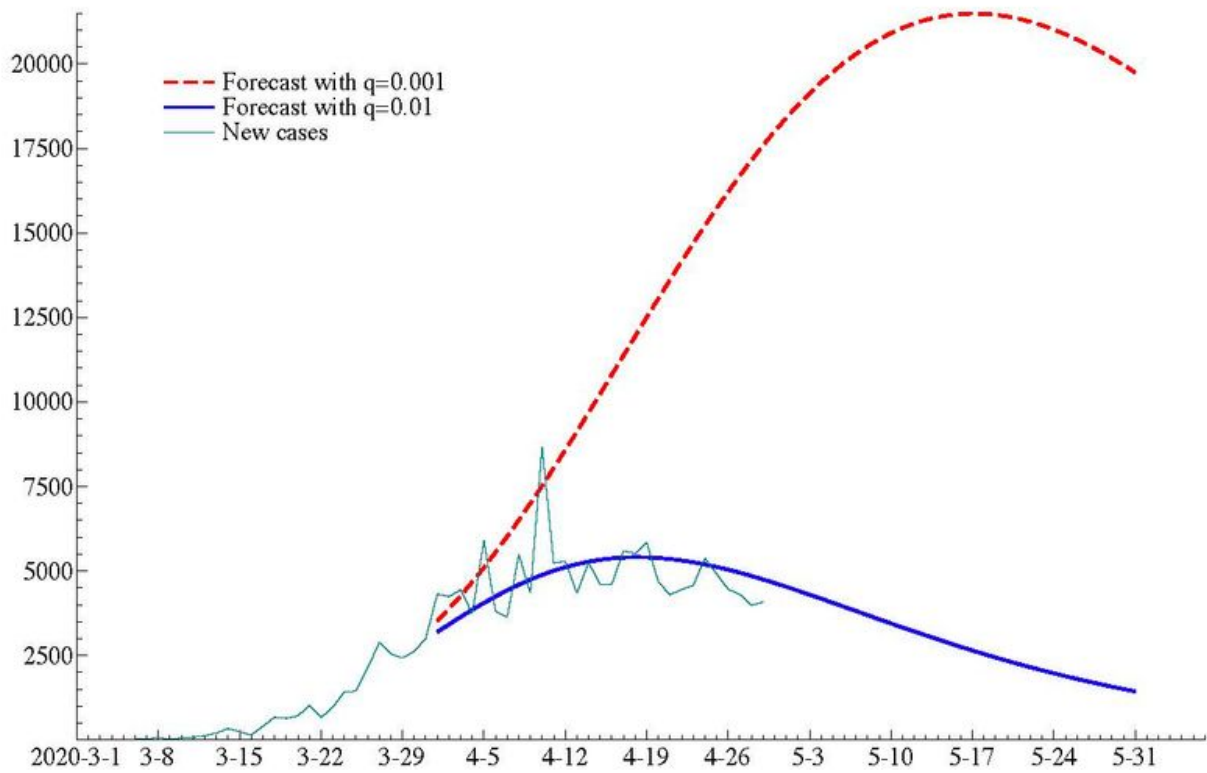


Figure 14. Forecasts of new cases in the United Kingdom, with $q = 0.01$ and $q = 0.001$ (dashed), using data up to April 1.

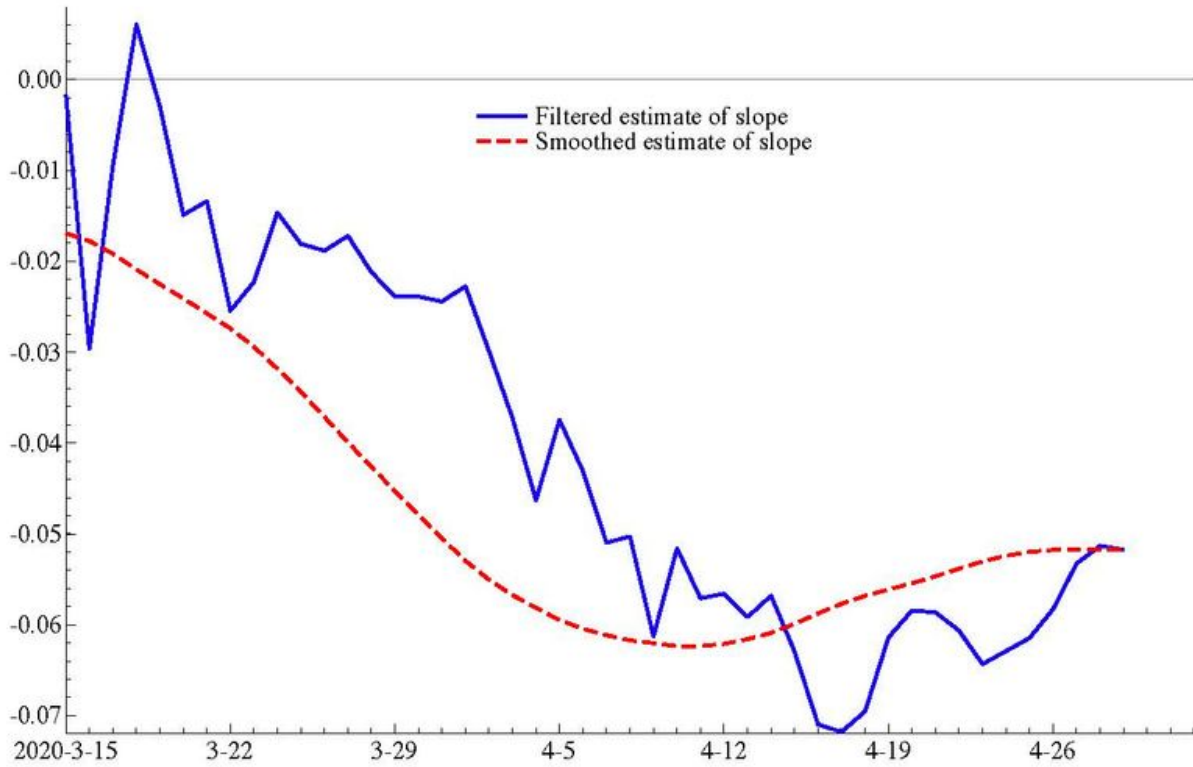


Figure 15. Filtered (solid blue) and smoothed (dashed - red) estimates of the slope in dynamic Gompertz model with daily effects (using data up to April 29).

Information about the pattern of the response can also be obtained by fitting a dynamic Gompertz model and graphing estimates of the slope. Figure 15 shows filtered and smoothed estimates of slope in such a model based on data up to April 29. The filtered estimates are most informative as they show the evolving changes in the slope after the lockdown of March 21. As can be seen, the big falls occur at the beginning of April, with little or no change after mid-April. Figure 8 shows a very similar movement in Germany, but taking place a few weeks earlier.

Growth curves can be used to parameterize a gradual response to an intervention. A permanent change is captured with the CDF and a temporary one with the PDF. A logistic CDF gives a response curve $W(t) = 1/(1 + \gamma_0^I \exp(-\gamma^I(t - t^I)))$, where t^I is the median. The I superscript distinguishes the parameters γ_0^I and γ^I from the ones used to model the time series itself. With t^L and t^U denoting the beginning and the end of the time span during which gradual response to the intervention occurs, the intervention dummies are defined by $w_t = 0$ for $t < t^L$, $w_t = W(t)$ for $t = t^L, t^L + 1, \dots, t^I, \dots, t^U$ and $w_t = 1$ for $t = t^U + 1, \dots, T$, or even just by $w_t = W(t)$ for $t = 1, \dots, T$. With the Gompertz CDF the response is $W(t) = \exp(-\gamma_0^I \exp(-\gamma^I(t - t^I)))$; in this case the point of inflection, corresponding to the maximum change, comes before the median of the time span between t^L and t^U . Figure 16 shows the CDFs

for logistic and Gompertz distributions with the median set to 7; for logistic $\gamma^I = 0.5$ and for Gompertz $\gamma^I = 0.2$.

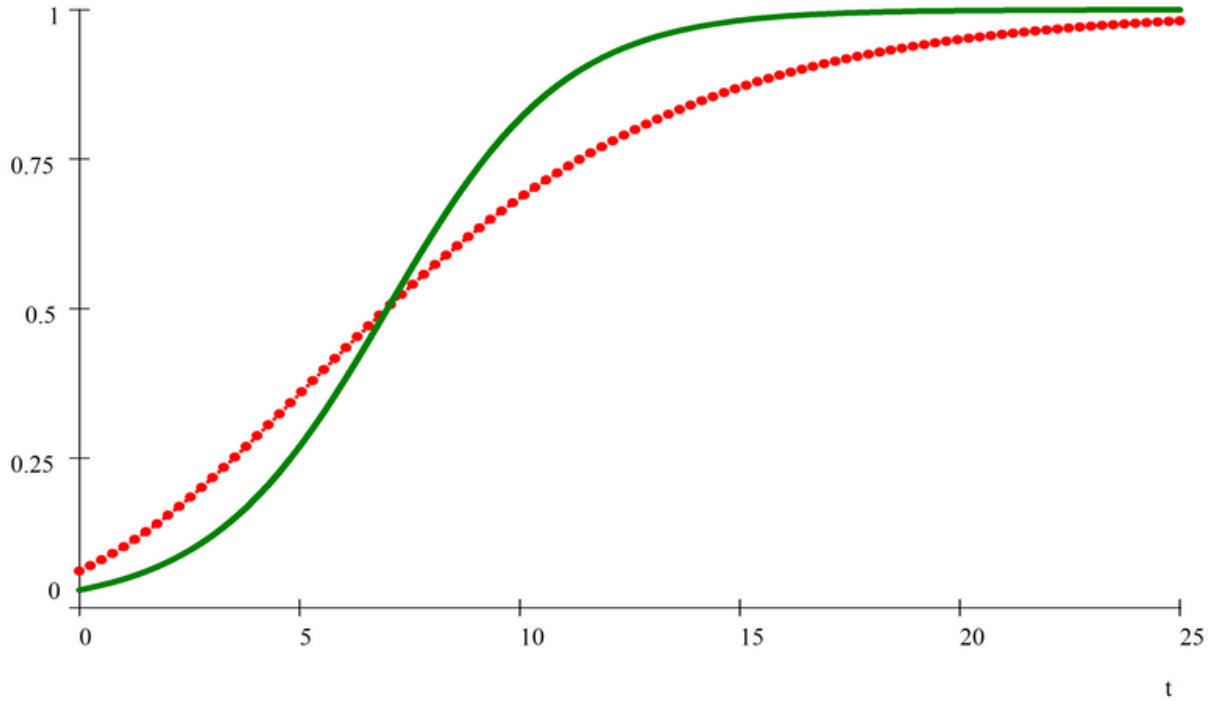


Figure 16. Cumulative distribution for logistic (green - solid) and Gompertz (red - dotted) functions with median set to 7. For logistic $\gamma = 0.5$ and for Gompertz $\gamma = 0.2$.

If the effect of a policy is to change γ in the model, a slope intervention is needed in Equation 3.2. Thus

$$(5.1) \quad \ln g_t = (\rho - 1) \ln Y_{t-1} + \delta - \gamma t - \beta t w_t + \varepsilon_t, \quad t = 1, \dots, T,$$

but unless the sample is moderately large, ρ will need to be fixed. When the full effect is realized, the slope on the time trend will have moved from γ to $\gamma + \beta$. A positive β will lower the growth rate, g_t , the peak of the incidence curve, and the final level.

Fitting the static model in Equation 5.1 to new cases in the United Kingdom with $\rho = 1$ and a logistic intervention, starting on March 26 and ending on April 12, gives an estimate of β equal to 0.020 (0.004) and an estimate of γ also equal to 0.020. The picture in Figure 17 is not unreasonable, but the estimate of the overall effect is 0.041, which may be a slight underestimate because minus one times the final slope in Figure 15 is close to 0.05. When the slope is allowed to be stochastic, $\beta_T + \gamma$ is estimated at 0.054. However, a stochastic slope risks some confounding with the intervention variable and in fact the estimate of β is reduced to 0.014 (0.006). Although neither model is completely satisfactory, both give a significant coefficient for the intervention variable, albeit after a degree of data mining.

6. A Second Wave?

With the relaxation of a lockdown, the unwelcome prospect of a second wave of infections arises. Dynamic GL models can monitor this possibility by tracking changes in the growth rate of the incidence curve. This growth rate, $g_y(t)$, depends on a potentially time-varying γ parameter, $\gamma(t)$, and the growth rate, $g(t)$. Differentiating the logarithm of $d\mu(t)/dt$ gives

$$(6.1) \quad g_y(t) = g(t) + g_g(t),$$

and for GL curves, it follows from Equation 2.4 that $g_g(t) = (\rho - 1)g(t) - \gamma(t)$. In the discrete time dynamic Gompertz model, Equation 3.5, the negative of the growth rate of the growth rate, $g_g(t)$, is tracked by the filtered estimates of the slope, that is, $\gamma_{t|t}$, while the growth rate itself is tracked by the exponent of the filtered level, that is, $g_{t|t} = \exp(\delta_{t|t})$. Figure 18 shows $\gamma_{t|t}$ and $g_{t|t}$ for Germany from a dynamic Gompertz model, together with the daily number of new cases. The maximum is identified as April 3 and after this date $g_{t|t} < \gamma_{t|t}$; see Equation 2.7. The possible onset of a second wave is raised if at some point $\gamma_{t|t}$ starts to fall below $g_{t|t}$, or below $\rho g_{t|t}$ in the general case. This would signal that the reproduction number, R_t , has moved up above one.

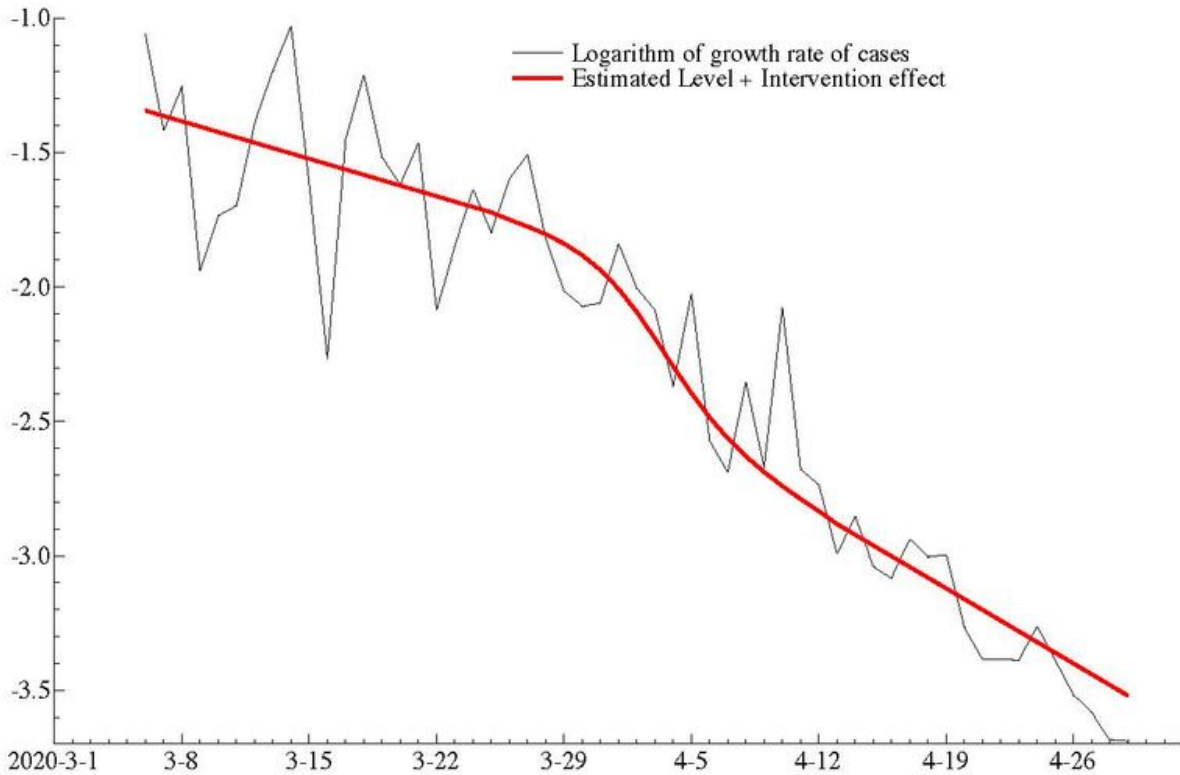


Figure 17. Estimates of logarithm of growth rate of total cases in United Kingdom with a logistic intervention and daily effect.

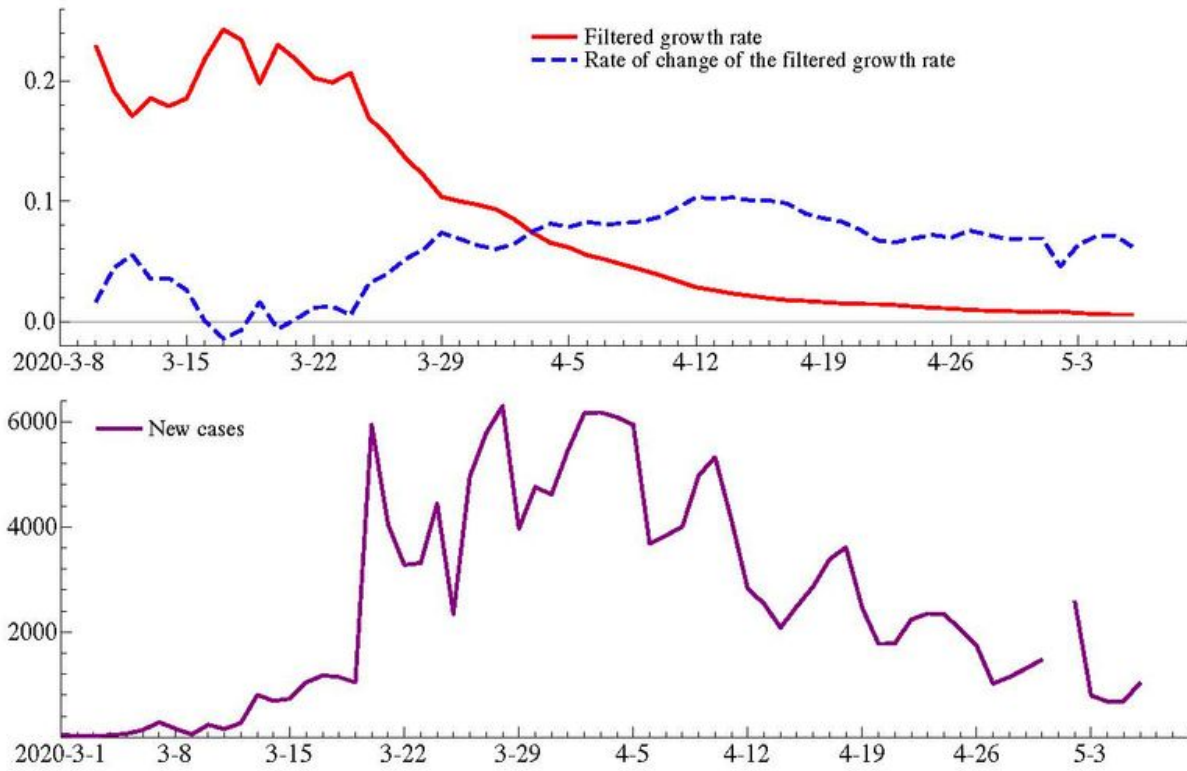


Figure 18. New cases in Germany together with filtered growth rate, $g_{t|t-1}$; and its rate of change, $\gamma_{t|t-1}$ (using data up to May 7).

The filtered estimates, $g_{t|t}$ and $\gamma_{t|t}$, are obtained by discounting of past observations, with the rate of discounting depending on the signal-noise ratio, q_c . When a new policy is implemented, q_c may need to increase so that past observations are discounted at a faster rate, as illustrated in Figure 14. In these circumstances, the only viable tracking option may be to try a range of q_c values, bearing in mind the risk of triggering a false alarm. However, if the effects of a policy are spread over a period of time, as with a gradual relaxation of a lockdown, a value of q_c estimated with the complete sample may be perfectly satisfactory.

If $\gamma_{t|t}$ decreases, some idea of sampling variability is needed to assess the implications. This may be obtained from the standard deviation of $\gamma_{t|t}$, as given by the Kalman filter. The variation in $g_{t|t}$ is harder to determine, but since it is typically much smoother than $\gamma_{t|t}$, it is likely to be small in comparison to the variation in $\gamma_{t|t}$.

7. Conclusions

A new class of time series models is developed for predicting future values of a variable that when cumulated is subject to an unknown saturation level. Such situations arise in a wide range of disciplines. The models provide a simple and viable alternative, or complement, to the forecasts produced by large-scale mechanistic models.

Generalized logistic growth curves provide the basis for our models. Estimating equations for the logarithm of the growth rate of the cumulative variable provide a good fit to the data, as assessed by standard statistical tests. Such models feature a time trend that can be made time-varying using the Kalman filter. The Gompertz model is a special case that works particularly well with a stochastic trend and, when this modification is made, the fit is often better than with an unrestricted generalized logistic model with a deterministic trend. The dynamic Gompertz can adapt to changing conditions and tracking the slope, especially the filtered slope, can be informative about the effect of interventions.

Additional components can be included in the models. These include seasonal or day-of-the week effects. The latter turned out to be relevant for new cases and deaths in our application to coronavirus in the United Kingdom and Germany. The dynamic Gompertz model worked well for both countries but was particularly impressive for German new cases. Deaths were successfully modeled by a Gompertz model with a negative binomial conditional distribution and a dynamic equation driven by the conditional score.

Estimating a model up to the point at which the effect of an intervention is likely to make itself felt and then making (unconditional) forecasts provides information about the effects of the intervention. The effectiveness of this approach is illustrated by fitting a dynamic Gompertz model to U.K. data and then investigating the effect of the coronavirus lockdown. Further insight into the effect of the lockdown is given by tracking the filtered estimate of the growth rate of the growth rate. The impact of the lockdown can be estimated ex post by including a growth curve as an explanatory variable, thereby allowing the intervention to have a gradual response.

Finally, we suggest that the possibility of a second wave can be monitored by tracking the filtered estimates of new cases or deaths given by our model. Of course, if these methods are to be useful in practice they require reliable up-to-date observations on new cases, preferably at a disaggregated level. Implementing viable tracking procedures and relating them to current methods for tracking the reproduction number, R , is the next phase in our research program.

Acknowledgments

We would like to thank Yucheng Bian, Anthony Kattuman, Dario Palumbo, Tom Pape, Mark Salmon, Stefan Scholtes, Jerome Simons, Robert Taylor, Steven Thiele, and Qingyuan Zhao for helpful comments and suggestions. We are also grateful to the Editor, Associate Editor, and three referees for their constructive comments. Jonas Knecht supplied valuable research assistance.

Disclosure Statement

Andrew Harvey and Paul Kattuman have no financial or non-financial disclosures to share for this article.

References

- Angelopoulos, A. N., Pathak, R., Varma, R., & Jordan, M. I. (2020). On identifying and mitigating bias in the estimation of the COVID-19 case fatality rate. *Harvard Data Science Review*, (Special Issue 1). <https://doi.org/10.1162/99608f92.f01ee285>
- Aronson, J. K., Brasset, J., & Mahtani, K. R. (2020). *When will it be over? An introduction to viral reproduction numbers*. The Centre for Evidence-Based Medicine. <https://www.cebm.net/oxford-covid-19-evidence-service/>
- Avery, C., Bossert, W., Clark, A., Ellison, G., & Ellison, S. F. (2020). *Policy implications of models of the spread of coronavirus: Perspectives and opportunities for economists*. Working Paper 27007. National Bureau of Economic Research. <https://doi.org/10.3386/w27007>
- Chowell, G., Sattenspiel, L., Bansal, S., & Viboud, C. (2016). Mathematical models to characterize early epidemic growth. *Physics of Life Reviews*, 18, 66–97. <https://doi.org/10.1016/j.plrev.2016.07.005>
- Ciufolini, I., & Paolozzi, A. (2020). Mathematical prediction of the time evolution of the COVID-19 pandemic in Italy by a Gauss error function and Monte Carlo simulations. *The European Physical Journal Plus*, 135, Article 355. <https://doi.org/10.1140/epjp/s13360-020-00383-y>
- Daley, D., & Gani, J. (1999). *Epidemic modelling: An introduction*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511608834>
- Flaxman, S., Mishra, S., Gandy, A., Unwin, H. J. T., Coupland, H., Mellan, T. A., Zhu, H., Berah, T., Eaton, J. W., Guzman, P. N. P., Schmit, N., Callizo, L., Team, I. C. C.-19 R., Whittaker, C., Winskill, P., Xi, X., Ghani, A., Donnelly, C. A., Riley, S., ... Bhatt, S. (2020). *Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries*. Report 13, Imperial College London, MRC Centre for Global Infectious Disease Analysis. <https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-13-europe-npi-impact/>
- Giordano, G., Blanchini, F., & Bruno, R. (2020). Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nature Medicine*, 26(6), 855–860. <https://doi.org/10.1038/s41591-020-0883-7>
- Gregg, J. V., Hossell, C. H., & Richardson, J. T. (1964). *Mathematical trend curves: An aid to forecasting*. Imperial Chemical Industries Monograph no. 1. Oliver and Boyd.
- Harvey, A. C. (1984). Time series forecasting based on the logistic curve. *Journal of the Operational Research Society*, 35(7), 641–646. <https://doi.org/10.1057/jors.1984.128>

Harvey, A. C. (2013). *Dynamic models for volatility and heavy tails: With applications to financial and economic time series*. Econometric Society Monograph. Cambridge University Press.

<https://doi.org/10.1017/CBO9781139540933>

Kleiber, C., & Kotz, S. (2003). *Statistical size distributions in economics and actuarial sciences*. Wiley.

<https://doi.org/10.1002/0471457175>

Koopman, S. J., Lit, R., & Harvey, A. C. (2020). *STAMP 8.4: Structural time series analyser, modeller and predictor*. Timberlake Consultants.

Levenbach, H., & Reuter, B. E. (1976). Forecasting trending time series with relative growth rate models.

Technometrics, 18(3), 26–72. <https://doi.org/10.1080/00401706.1976.10489446>

Lit, R., Koopman, S. J., & Harvey, A. C. (2020). *Time series lab - Score edition*. <https://timeserieslab.com>

McDonald, J. B., & Xu, Y. J. (1995). A generalization of the beta distribution with applications. *Journal of Econometrics*, 66(1–2), 133–152. [https://doi.org/10.1016/0304-4076\(94\)01612-4](https://doi.org/10.1016/0304-4076(94)01612-4)

Meade, N., & Islam, T. (1995). Forecasting with growth curves: An empirical comparison. *International Journal of Forecasting*, 11(2), 199–215. [https://doi.org/10.1016/0169-2070\(94\)00556-R](https://doi.org/10.1016/0169-2070(94)00556-R)

Murray, C. J. L. (2020). Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. *MedRxiv*. <https://doi.org/10.1101/2020.03.27.20043752>

Panik, M. J. (2014). *Growth curve modeling: Theory and applications*. Wiley.

Appendix

Data

Surveillance data is the primary source of information for study of the dynamics of the pandemic. The data we use in this article on time series of confirmed cases and deaths is sourced from the European Centre for Disease Prevention and Control (ECDC), which is a European Union agency with responsibility for EU-level surveillance of infectious diseases. Since the beginning of the pandemic, through an Early Warning and Response System, ECDC has obtained daily numbers of confirmed cases of COVID-19 and deaths attributed to the virus from EU/EEA Member States and the United Kingdom.

ECDC's data collection involves systematic screening of daily reports from health authorities by a team of epidemiologists, and an extensive validation process. The screened and validated daily reports are documented in a database, and the daily extract consisting of the total number of diagnosed COVID-19 infections and the number of deceased patients for each country is published on the ECDC website.

It is important to understand the possible sources of biases in the COVID-19 surveillance data. [Angelopoulos et al. \(2020\)](#) provide a systematic and insightful review of the many sources of bias in the time series of deaths, cases, and recoveries relating to COVID-19. Their particular focus is on implications for the estimation of relative case fatality ratios between groups and countries. They show how some, though not all, biases can be addressed. Even when biases cannot be addressed, it is important to know their sources and implications.

In terms of the time series of cases, the definition of a ‘case’ can vary from country to country, and may also change over time. Confirmation of cases is affected by the testing regime—test availability, and the criteria under which an individual is eligible for testing—which is liable to change over time. In the early days of the pandemic, severe cases will have been diagnosed more often than milder cases. True numbers of infections are likely to be much higher than reported in countries where testing is limited. Biases also arise from imperfect or delayed reporting and errors in reporting.

Overall, deaths are more likely to be reported accurately by health care providers, but sources of bias remain. Time delays in reporting and the misattribution of deaths are not uncommon. In addition, national reporting guidelines during the early days are often focused on hospital deaths to the exclusion of deaths in community and care homes.

We focus on the United Kingdom and Germany due to our familiarity with the health data infrastructures and validation procedures in these countries. The data we use in this article can be accessed from:

<https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide>

©2020 Andrew Harvey and Paul Kattuman. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](#), except where otherwise indicated with respect to particular material included in the article.

Footnotes

1. See https://www.rki.de/DE/Home/homepage_node.htm ↵

References

- Angelopoulos, A. N., Pathak, R., Varma, R., & Jordan, M. I. (2020). On identifying and mitigating bias in the estimation of the COVID-19 case fatality rate. *Harvard Data Science Review*. <https://doi.org/10.1162/99608f92.f01ee285> ↵
- Aronson, J. K., Brasset, J., & Mahtani, K. R. (2020). *When will it be over? An introduction to viral reproduction numbers*. <https://www.cebm.net/oxford-covid-19-evidence-service/> ↵
- Avery, C., Bossert, W., Clark, A., Ellison, G., & Ellison, S. F. (2020). *Policy implications of models of the spread of coronavirus: Perspectives and opportunities for economists*. <http://www.nber.org/papers/w27007> ↵

- Chowell, G., Sattenspiel, L., Bansal, S., & Viboud, C. (2016). Mathematical models to characterize early epidemic growth. *Physics of Life Reviews*, 18, 66-97. <https://doi.org/10.1016/j.plrev.2016.07.005> ↵
- Ciufolini, I., & Paolozzi, A. (2020). Mathematical prediction of the time evolution of the COVID-19 pandemic in Italy by a Gauss error function and Monte Carlo simulations. *The European Physical Journal Plus* 135: 355. <https://doi.org/10.1140/epjp/s13360-020-00383-y> ↵
- Daley, D., & Gani, J. (1999). *Epidemic modelling: An introduction*. Cambridge University Press. ↵
- Flaxman, S., Mishra, S., Gandy, A., Unwin, H. J. T., Coupland, H., Mellan, T. A., Zhu, H., Berah, T., Eaton, J. W., Guzman, P. N. P., Schmit, N., Callizo, L., Team, I. C. C.-19 R., Whittaker, C., Winskill, P., Xi, X., Ghani, A., Donnelly, C. A., Riley, S., ... Bhatt, S. (2020). *Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries*. Report 13, Imperial College London, MRC Centre for Global Infectious Disease Analysis. <https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-13-europe-npi-impact/> ↵
- Giordano, G., Blanchini, F., & Bruno, R. (2020). Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nature Medicine*, 26, 855–860. <https://doi.org/10.1038/s41591-020-0883-7> ↵
- Gregg, J. V., Hossell, C. H., & Richardson, J. T. (1964). *Mathematical trend curves: An aid to forecasting*. Imperial Chemical Industries Monograph no. 1. Oliver. ↵
- Harvey, A. C. (1984). Time series forecasting based on the logistic curve. *Journal of the Operational Research Society*, 35, 641-646. ↵
- Harvey, A. C. (2013). *Dynamic models for volatility and heavy tails: With applications to financial and economic time series*. Econometric Society Monograph. Cambridge University Press. ↵
- Kleiber, C., & Kotz, S. (2003). *Statistical size distributions in economics and actuarial sciences*. Wiley. ↵
- Koopman, S. J., Lit, R., & Harvey, A. C. (2020). *STAMP 8.4: Structural time series analyser, modeller and predictor*. Timberlake Consultants. ↵
- Levenbach, H., & Reuter, B. E. (1976). Forecasting trending time series with relative growth rate models. *Technometrics*, 18, 26-72. ↵
- Lit, R., Koopman, S. J., & Harvey, A. C. (2020). *Time series lab - score edition*. <https://timeserieslab.com> ↵
- McDonald, J. B., & Xu, Y. J. (1995). A generalization of the beta distribution with applications. *Journal of Econometrics*, 66, 133-152. ↵
- Meade, N., & Islam, T. (1995). Forecasting with growth curves: An empirical comparison. *International Journal of Forecasting*, 11, 199-215. ↵
- Murray, C. J. L. (2020). Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. *medRxiv*. <https://doi.org/10.1101/2020.03.27.20043752> ↵
- Panik, M. J. (2014). *Growth curve modeling: Theory and applications*. Wiley. ↵