UNIVERSITY OF CAMBRIDGE
Faculty of Economics

CAMBRIDGE WORKING PAPERS IN ECONOMICS

Time series modeling of epidemics: leading indicators, control groups and policy assessment

Andrew
Harvey

Abstract

This article shows how new time series models can used to track the progress of an epidemic, forecast key variables and evaluate the effects of policies. The univariate framework of Harvey and Kattuman (2020) is extended to model the relationship between two or more series, and the role of common trends is discussed. Data on daily deaths from Covid-19 in Italy and the UK provides an example of leading indicators when there is balanced growth. When growth is not balanced, the model can be extended by including a nonstationary component in the leading series. The viability of this model is investigated by examining the relationship between new cases and deaths in the Florida second wave of summer 2020. The balanced growth framework is then used as the basis for policy evaluation by showing how some variables can serve as control groups for a target variable. This approach is used to investigate the consequences of Sweden's soft lockdown coronavirus policy.

Reference Details

CWPE 2114
Published 22 February 2021

Key Words Balanced growth, Co-integration, Covid-19, Gompertz curve, Kalman filter, Stochastic trend
JEL Codes C22, C32

Website www.econ.cam.ac.uk/cwpe

Time series modeling of epidemics: leading indicators, control groups and policy assessment

Andrew Harvey
Faculty of Economics, Cambridge University

February 22, 2021

Abstract

This article shows how new time series models can used to track the progress of an epidemic, forecast key variables and evaluate the effects of policies. The univariate framework of Harvey and Kattuman (2020) is extended to model the relationship between two or more series, and the role of common trends is discussed. Data on daily deaths from Covid-19 in Italy and the UK provides an example of leading indicators when there is balanced growth. When growth is not balanced, the model can be extended by including a nonstationary component in the leading series. The viability of this model is investigated by examining the relationship between new cases and deaths in the Florida second wave of summer 2020. The balanced growth framework is then used as the basis for policy evaluation by showing how some variables can serve as control groups for a target variable. This approach is used to investigate the consequences of Sweden's soft lockdown coronavirus policy.

KEYWORDS: Balanced growth; Co-integration; Covid-19; Gompertz curve; Kalman filter; Stochastic trend.

JEL classification; C22, C32

Contact details: Faculty of Economics, Cambridge University, Cambridge CB3 9DD. Email: ach34@cam.ac.uk

1 Introduction

The aim of this article[1] is to show how time series models[2] can be used to track the progress of an epidemic, forecast key variables and evaluate the effects of policies. Developing effective techniques to accomplish these tasks is of some importance, because, as documented by Ioannidis et al (2020), the performance of many of the methods used to forecast the current Covid-19 epidemic has not been impressive. The new models draw much of their inspiration from time series econometrics. However, the characteristics of time series for epidemics are different from those of most time series in economics and these differences need to be taken into account.

Harvey and Kattuman (2020a) - hereafter HK - developed a class of univariate time series models for predicting future values of a variable which when cumulated is subject to an unknown saturation level. In these models, the logarithm of the growth rate of the cumulated series depends on a time trend. 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. Nowcasts and forecasts of the variables of interest, such as the daily number of cases, its growth rate and the instantaneous reproduction number, R_t, can be made. Estimation of the models is by maximum likelihood and goodness of fit can be assessed by standard statistical test procedures.

[1] This is a much revised version of my November 2020 National Institute of Economic and Social Research discussion paper, *NIESR* DP 517.

[2] The application of classical time series methods to data on epidemics is relatively undeveloped. Most of the emphasis has been on building 'semi-mechanistic' models to simulate the path of an epidemic under different assumptions about behaviour and policies; see Avery et al (2020)

Time series models can also be used to address other questions by exploring relationships between different series. One application concerns how the time path of an epidemic in a country which suffers an outbreak before another can be used as a leading indicator. The rationale for modeling the logarithm of the growth rate (of the cumulated series) comes from the properties of a Gompertz growth curve and when two such curves follow the same time path, but one lags the other, the trends in the series on the logarithms of the growth rate are a constant distance apart. This suggests that when the trends are stochastic, the same will be true. This situation, known as balanced growth, arises in macroeconomics and is a special case of what econometricians call co-integration; see, for example, Stock and Watson (1988). The situation is illustrated by showing how the time path of deaths in the UK in the first few months of the coronavirus epidemic follows the time path of deaths in Italy two weeks earlier.

The requirement that two series exhibit balanced growth, while highly desirable, is not necessary for one to be a good leading indicator of the other. The need for additional flexibility is explored with data from the 'second wave' of coronavirus in Florida in the early part of the summer of 2020 where it is shown how daily new cases can potentially offer improved forecasts of deaths in two to three weeks time. The forecasts are based on a bivariate unobserved component time series model that combines the dynamic information in the two series by a common trend specified as an integrated random walk but includes an independent random walk component for new cases.

Multivariate time series models can be used to assess the impact of policies using control groups. The impact of lockdown is explored by developing the ideas associated with balanced growth to try to estimate the number of coronavirus deaths in Sweden had a more stringent lockdown been imposed. The methodology draws on the study of control groups in time series by Harvey and Thiele (2020). It is argued that the fact that death rates in Sweden were roughly ten times those in neighbouring countries could be

misleading; the growth paths of the UK and Italy provide more relevant information. A comparison is made with studies based on the synthetic control method of Abadie et al (2010, 2015).

2 Growth curves and time series models

This section sets out the basic model in which the logarithm of the growth rate of the cumulated series consists of a stochastic trend plus an irregular term. It is then shown how the framework may be extended to model the relationship between two series.

2.1 Dynamic trend models

The observational model uses data on the time series of the cumulated total of confirmed cases or deaths, Y_t, $t = 0, 1, ..., T$, and the daily change, $y_t = \Delta Y_t = Y_t - Y_{t-1}$. HK show how the theory of generalized logistic growth curves suggests models for $\ln y_t$ and $\ln g_t$, where $g_t = y_t/Y_{t-1}$ or $\Delta \ln Y_t$. For the special case of the Gompertz growth curve

$$\ln y_t = \ln Y_{t-1} + \delta - \gamma t + \varepsilon_t, \quad \gamma > 0, \quad t = 1, ..., T, \tag{1}$$

and

$$\ln g_t = \delta - \gamma t + \varepsilon_t, \quad t = 1, ..., T, \tag{2}$$

where ε_t is a random disturbance term.

A stochastic, or time-varying, trend may be introduced into (2), to give the dynamic trend model

$$\ln g_t = \delta_t + \varepsilon_t, \quad \varepsilon_t \sim NID(0, \sigma_\varepsilon^2), \qquad t = 1, ..., T, \tag{3}$$

where

$$\begin{aligned} \delta_t &= \delta_{t-1} - \gamma_{t-1} + \eta_t, & \eta_t &\sim NID(0, \sigma_\eta^2), \\ \gamma_t &= \gamma_{t-1} + \zeta_t, & \zeta_t &\sim NID(0, \sigma_\zeta^2), \end{aligned} \tag{4}$$

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 an 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 $q_\zeta = \sigma_\zeta^2/\sigma_\varepsilon^2$; the trend is deterministic when $q_\zeta = 0$.

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 while a negative γ_t means that the growth rate is increasing.

Additional components, such as day of the week effects, can be added to (3). These may be deterministic or stochastic. Explanatory variables, including interventions, can also be included, as may stationary components, such as autoregressive processes. All these models can be handled using techniques based on state space models and the Kalman filter; see Durbin and Koopman (2012). Here the STAMP package of Koopman et al. (2020) is used. Estimation of the unknown variance parameters is by maximum likelihood (ML). Diagnostic tests for normality and residual serial correlation are based on the one-step ahead prediction errors, $v_t = \ln g_t - \delta_{t|t-1}$, $t = 3, ..., T$.

The KF outputs the estimates and forecasts of the state vector $(\delta_t, \gamma_t)'$. Estimates at time t conditional on information up to and including time t are denoted $(\delta_{t|t}, \gamma_{t|t})'$, while predictions j steps ahead are $(\delta_{t+j|t}, \gamma_{t+j|t})'$. The smoother, which estimates the state at time t based on all T observations in

the series, is denoted $(\delta_{t|T}, \gamma_{t|T})'$.

Remark 1 *When the observations are small, a negative binomial distribution for y_t, conditional on past observations, may be appropriate. HK show how the model may be modified to deal with this possibility. However, the numbers in the applications here are big enough to allow y_t to be treated as conditionally lognormal and hence for the conditional distribution of $\ln g_t$ to be considered normal.*

2.2 Forecasts

The forecasts of the trend in future values of $\ln g_t$ in the dynamic Gompertz model are given by $\delta_{T+\ell|T} = \delta_{T|T} - \gamma_{T|T}\ell$, $\ell = 1, 2, ..$, 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 are obtained from a recursion for the trend in their cumulative total, Y_t, namely

$$\mu_{T+\ell|T} = \mu_{T+\ell-1|T}(1 + g_{T+\ell|T}), \qquad \ell = 1, 2, .. \tag{5}$$

where $g_{T+\ell|T} = \exp \delta_{T+\ell|T}$ and $\mu_{T|T} = Y_T$. The trend in the daily figures is then

$$\mu_{y,T+\ell|T} = g_{T+\ell|T}\mu_{T+\ell-1|T} \quad , \qquad \ell = 1, 2, .. \tag{6}$$

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, $\delta_{T+\ell|T}$.

2.3 Comparing different growth curves

The Gompertz growth curve lies behind the notion of setting up time series models in which the logarithm of the growth rate of the cumulative total of a variable follows a trend. It is therefore able to provide insight on how to formulate and interpret models linking several series.

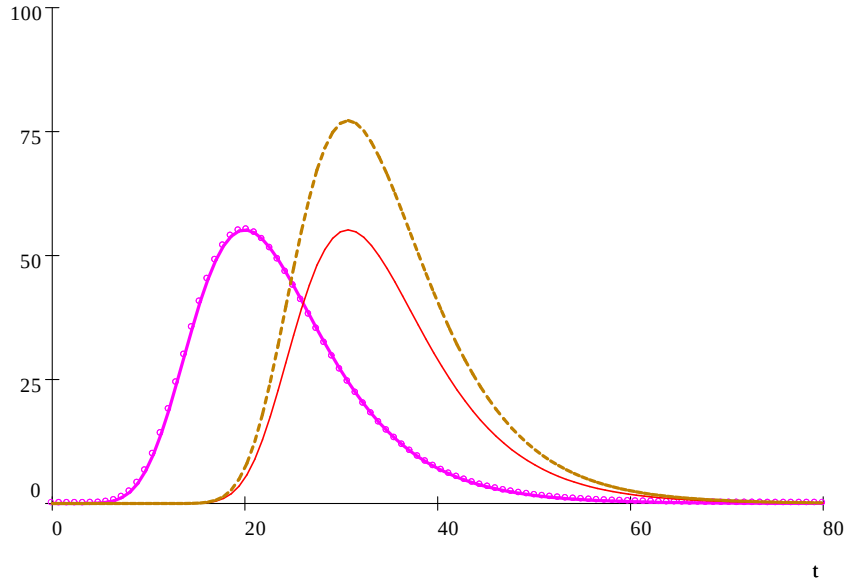

Figure 1: Gompertz incidence curves with $\gamma = 0.15$, $\alpha_1 = 20$ for the left hand curve and $\alpha_2 = 100$ for the right hand curves.

The Gompertz growth curve is

$$\mu(t) = \overline{\mu}\exp(-\alpha e^{-\gamma t}), \qquad \alpha, \gamma > 0, \qquad -\infty < t < \infty, \tag{7}$$

where γ is a growth rate parameter, $\overline{\mu}$ is the upper bound or saturation level and α reflects initial conditions. The associated incidence curve is

$$d\mu(t)/dt = \mu'(t) = \gamma\alpha\mu(t)\exp(-\gamma t),$$

with a peak at $t = \gamma^{-1}\ln\alpha$. Figure 1 shows an incidence curve with a peak at $t = 19.97$, together with the same curve shifted to the right so the peak is at 30.71. A curve above the right hand curve is also shown; this is higher because the value of $\overline{\mu}$ is 1400 rather than 1000 as it is for the other two curves. In all cases $\gamma = 0.15$, but for the left hand curve α is 20 whereas for the right hand curves it is 100.

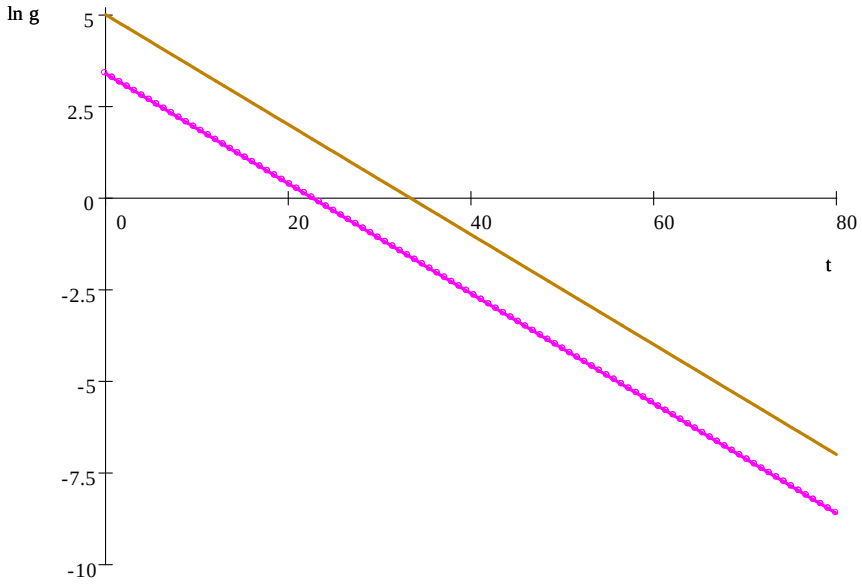

Figure 2: Logarithms of the growth rates for incidence curves in Figure 1; $\gamma = 0.15$, $\alpha_1 = 20$ and $\alpha_2 = 100$ (upper line).

Although the right hand curves in Figure 1 clearly lag the left hand one, it is not immediately evident how to model the relationship. However, the logarithms of the growth rates of $\mu(t)$ are

$$\ln g(t) = \delta - \gamma t, \qquad t \geq 0, \tag{8}$$

where $\delta = \ln \alpha\gamma$; compare (2). Figure 2 shows the two lines for $\ln g(t)$ running in parallel. The distance between them depends on the intercepts, δ, which in turn depend on the initialization parameter, α. The height of the incidence curve, which depends on the saturation level, $\overline{\mu}$, is irrelevant; as a result the lines corresponding to the two right hand incidence curves in Figure 1 are identical. This is important because it means that small populations can be compared with big ones: size does not matter.

When two lines are parallel, the upper line lags the lower one by

$$k = \frac{\delta_2 - \delta_1}{\gamma} = \frac{\ln \alpha_2 - \ln \alpha_1}{\gamma}, \tag{9}$$

where δ_1 and δ_2 are the intercepts of the lower and upper lines respectively and α_1 and α_2 are the corresponding initial conditions. In Figure 2 the lag is $k = 10.73$. When the $\gamma's$ are different, the epidemic progresses at different speeds. The lines for $\ln g(t)$ are no longer parallel and the time lag is no longer constant.

3 A model for leading indicators

Now consider observational models of the form (2) for two time series which are on the same growth path because $\gamma_1 = \gamma_2$ but the first series leads the second by k time periods. The observations run from $t = 1$ to T but when the first series is lagged by k time periods, $\ln g_{1,t-k}$ runs from $t = k + 1$ to $T + k$. Subtracting the first series from the second gives

$$\ln g_{2t} = \delta + \ln g_{1,t-k} + \varepsilon_t, \tag{10}$$

where $\delta = \ln(\alpha_2/\alpha_1)$ and the disturbance term is $\varepsilon_t = \varepsilon_{2t} - \varepsilon_{1,t-k}$. The equation takes the same form when the trends are stochastic, so long as there is balanced growth. The disturbance, ε_t, can replaced by any stationary process.

When the two series are not on the same growth path, there is no longer a value of k for the contrast in (10) that makes it stationary. The stationarity test of Kwiatkowski et al (1992) - the KPSS test- can be used to test for this possibility.

A bivariate time series model combines the dynamic information in the target series with that in the leading indicator. It is set up by lagging the observations on the leading indicator so that they are aligned with the target.

Hence defining $g_{1,t}^{(k)} = g_{1,t-k}$ for $t = k+1, .., T+k$ gives

$$\begin{aligned} \ln g_{1,t}^{(k)} &= \delta_t + \psi_t + \varepsilon_{1t}, \qquad t = k+1, ..., T+k, \\ \ln g_{2t} &= \overline{\delta} + \delta_t + \varepsilon_{2t}, \qquad\qquad t = k+1, ..., T. \end{aligned} \tag{11}$$

The k future values of $\ln g_{2,T+j}$, $j = 1, .., k$ are treated as missing observations[3]. The trend, δ_t, is an IRW that is designed to capture the growth path of the target series. Its initial level has been (arbitrarily) assigned to the first series; hence the need for a constant term, $\overline{\delta}$, in the second. The role of the other stochastic component, ψ_t, is to allow for deviations of the leading indicator from the balanced growth path. A convenient specification for it is the first-order autoregressive process, $\psi_t = \phi\psi_{t-1} + \zeta_t$, where ζ_t is $NID(0, \sigma_\zeta^2)$. All disturbances, including ε_{1t} and ε_{2t}, are Gaussian and assumed to be mutually as well as serially independent. Only a single lag is present in (11). More lags could be included, but the aim is find a viable leading indicator for movements in the trend rather than to estimate a distributed lag for the observations. Estimation of (11) is by state space methods. As new observations become available, nowcasts and forecasts are updated by the Kalman filter.

When $|\phi| < 1$, the series are co-integrated with balanced growth. In the absence of balanced growth, the suggestion is to let ψ_t be a random walk, by setting $\phi = 1$. The value of k is then based on experimentation and prior information about what might constitute a reasonable lag. The hope is that the RW specification for ψ_t enables its movements to be separated from those in the IRW trend.

In a univariate model, Harvey and Kattuman (2020b) use filtered estimates of $g_{y,t}$, given by $g_{y,t|t} = g_{t|t} - \gamma_{t|t}$, to track the progress of an epidemic. A corresponding estimator of the instantaneous reproduction number, R_t, can be constructed in a number of ways, as in Wallinga and Lipsitch (2008).

[3]If the first k observations on the second series are reliable they could be used by treating the first k values of the first series as missing.

The most practical for Covid-19 are

$$\widetilde{R}_{t,\tau} = 1 + \tau g_{y,t|t} \quad \text{and} \quad \widetilde{R}^{e}_{t,\tau} = \exp(\tau g_{y,t|t}), \tag{12}$$

where τ is the generation interval, which is the number of days that must elapse before an infected person can transmit the disease; setting $\tau = 4$ is a good choice. In the bivariate leading indicator model, the filtered estimates, $g_{t|t}$ and $\gamma_{t|t}$, for the second series give the nowcast of $g_{y,t|t}$ at time $t = T$ and the forecast at $t = T + k$. Forecasts can also be made beyond $t = T + k$, but without the benefit of corresponding values of the leading indicator. The Kalman filter and smoother implicitly weights observations in both series in order to compute $g_{t|t}$ and $\gamma_{t|t}$ for the target.

3.1 Italy and the UK

Figure 3 shows the daily deaths in Italy and the UK from March 2nd to June 20th, 2020; after that the numbers for Italy start to become small. The figures are for when the deaths were recorded rather than when they occured. Series based on date of death would not have the daily pattern but were difficult to obtain at that time. Data sources are given in the Appendix.

Italy clearly leads the UK but the relationship is captured more precisely in Figure 4 which shows the logarithms of the growth rates (LDL) of total deaths. The UK numbers are small at the beginning of March and so there are missing observations. A lag of 14 is not inconsistent with prior information and it has the attraction of lining up the days of the week in the two countries. Figure 5 shows the LDL series with Italy lagged by 14 days together with the contrast between the two countries obtained by subtracting Italy from the UK. The contrast series appears to be stationary with a mean close to zero; without the lag for Italy the values at the end of March and the beginning of April tend to be higher than the others, reflecting the later UK lockdown. Estimating a regression model with daily dummy variables removed most

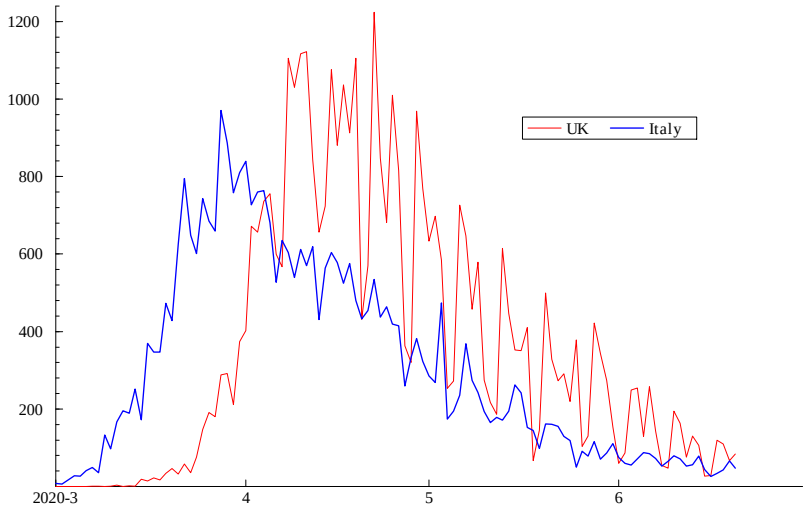

Figure 3: Daily deaths in Italy and UK

of the serial correlation and gave a mean of $\widetilde{\delta} = -0.083$, with a standard error on 0.035. The diagnostic statistics were[4]: $r(1) = -0.06$, $Q(14) = 13.40$, $BS = 1.85$ and $H = 1.24$.

Fitting a bivariate time series model of the form (11), starting on March 16th and finishing on July 5th, gave a slowly changing trend that was close to being deterministic. The δ_{1t} term was excluded but a daily component was included. The estimate of the daily growth rate of UK deaths 14 days beyond the final observation on June 20th was $g_{y,T+k|T} = -0.058$, giving a forecast of $\widetilde{R}_{T+k,4} = 0.77$.

[4] 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.

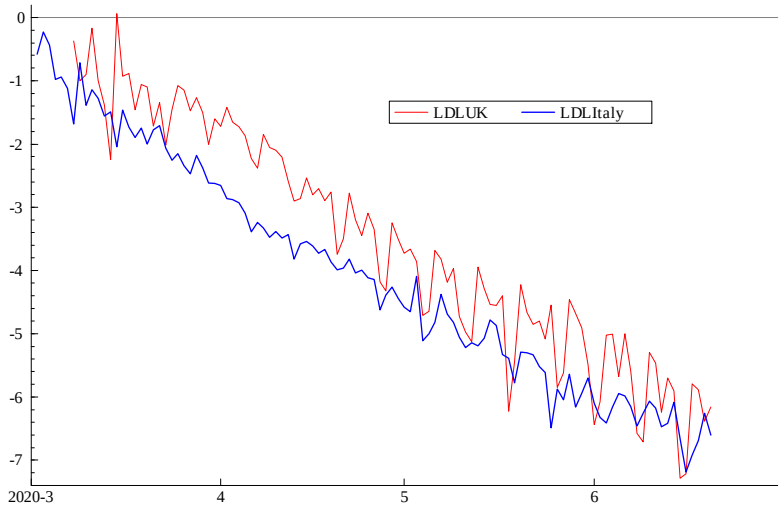

Figure 4: Logarithms of the growth rates (LDL) of total deaths in UK and Italy

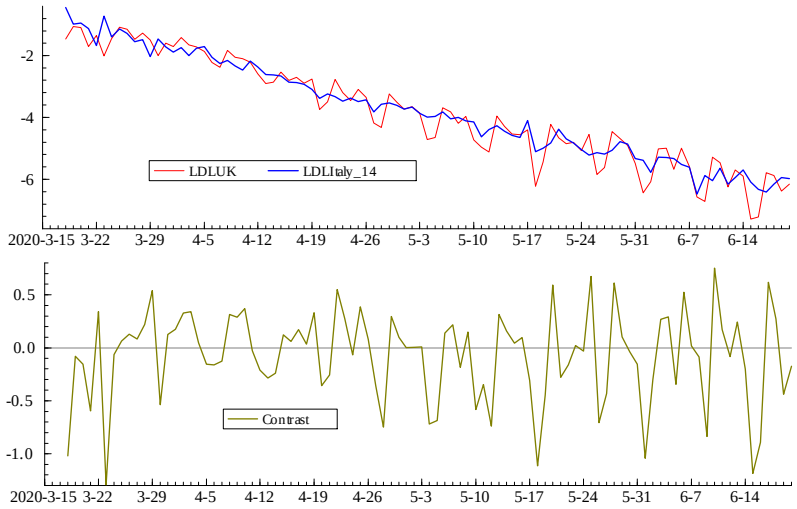

Figure 5: LDL series from March 16th to June 20th with Italy lagged by 14 days together with the contrast LDLUK-LDLItaly

3.2 Deaths and New Cases in Florida

Daily cases of Covid-19 in the US state of Florida peaked in early April. There was then a decline following a lockdown during April. After April restrictions were eased and there was a leveling out in May, followed by a sharp rise in June. This second wave poses a challenge for a model in which new cases are used as a leading indicator for deaths. The model deals with the second wave by allowing $\gamma_{t|t}$ to become negative; estimates of R_t can still be obtained from $g_{y,t|t}$, as in (12).

Aside from the model having to deal with a situation where new cases and deaths rise and fall, there is the problem that the basis on which new cases are recorded changes over time. At the beginning of the pandemic, new cases in many countries were primarily hospital admissions, but over time testing became more widespread. A balanced growth model assumes that the growth rate in deaths is the same as the growth rate in new cases. When this does not hold the inclusion of a stochastic trend in the model offers a way of dealing with the discrepancy. In the case of Florida there was an increase in testing in May, although the growth rate in tests was roughly constant from the end of May onwards. This suggests that the growth rate of confirmed new cases may still be a good indicator of the path of new infections.

The observations, particularly deaths, have a strong weekly pattern. A clearer impression of the underlying trend is given by Figure 6 which shows a seven day moving average of the logarithms of the growth rates of total new cases and deaths from March 29th to July 19th 2020 inclusive. The fact that cases are leading is clear with the gap increasing over time because of increased testing. New cases peak some time before the end of the sample whereas deaths appear to be at their peak, something confirmed by later observations.

The lag in (11) is chosen so as to get maximum benefit for new cases as a leading indicator. It is not trying to model the distribution of days from

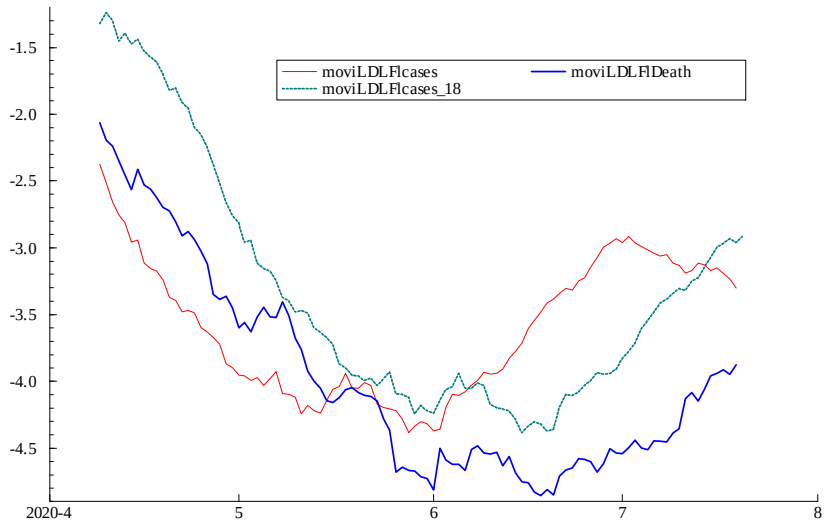

Figure 6: Seven day MA of LDL Deaths in Florida, New Cases and New Cases lagged 18 days (dotted line)

infection to death although the choice of k may be roughly aligned with the mean time to death. After some experimentation it was decided to fix the lag at 18.

The model, including day of the week variables, was fitted to the Florida data from March 29th till July 19th, with the new cases shifted forward 18 days so as to end on August 6th; thus $k = 18$ in (11). Specifying ψ_t as a first-order autoregressive process gave an estimated ϕ of 0.998, so a RW seems appropriate. The smoothed estimates of the daily component for deaths and the RW are shown in Figure 7; the high variation in the daily component coincides with relatively low numbers of deaths. The size and variability of the daily component in deaths was much bigger than for new cases. Similarly the prediction error variance of 0.115 for new cases was less than half the 0.253 obtained for deaths. Little serial correlation remained in the residuals for deaths: the Box-Ljung Q-statistic for the first 18 residual

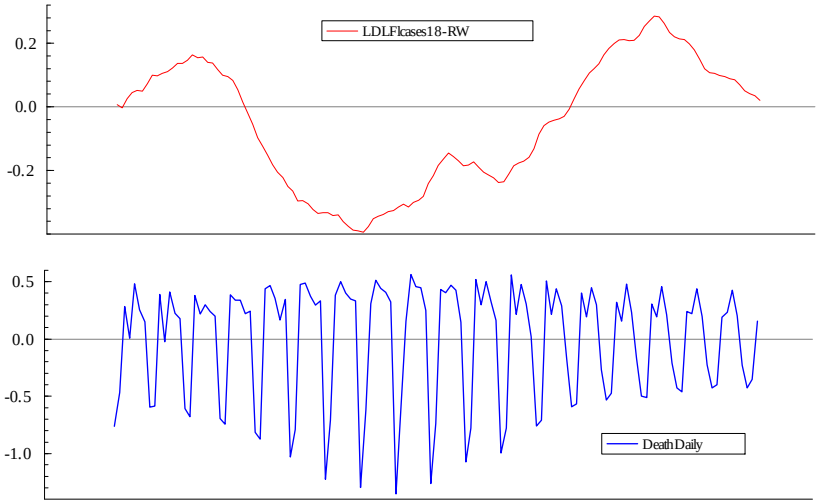

Figure 7: Additional RW component in cases and daily component in deaths

autocorrelations was 8.16, while the corresponding figure for cases was a little higher[5] at 25.01. The signal-noise ratio was estimated as 0.00037, so the trend changes relatively slowly but is still able to adapt to changes in direction.

Figure 8 shows the forecasts of of the logarithm of the growth rate deaths, obtained by using the leading indicator, together with the actual observations from July 20th to August 6th. The dotted line is the trend in deaths. As can be seen, the model forsees the turning point.

An estimate of the growth rate of the epidemic based on deaths can be computed from the estimates of the level and slope of the LDL Death series on August 6th. These were -3.945 and -0.0105 respectively giving an estimated growth rate, g_y, of $\exp(-3.945) - 0.0105 = 0.0089$ and, correspondly, $R^e_{t|t,4} = 1.036$. It can be seen from the graph that the growth rate of LDLFlDeath

[5]The suggestion in the STAMP manual is to test against a chi-square variable with allowance made for the loss in degrees of freedom due to estimated parameters which here is six. Thus the chi-square may be taken to have 12 degrees of freedom

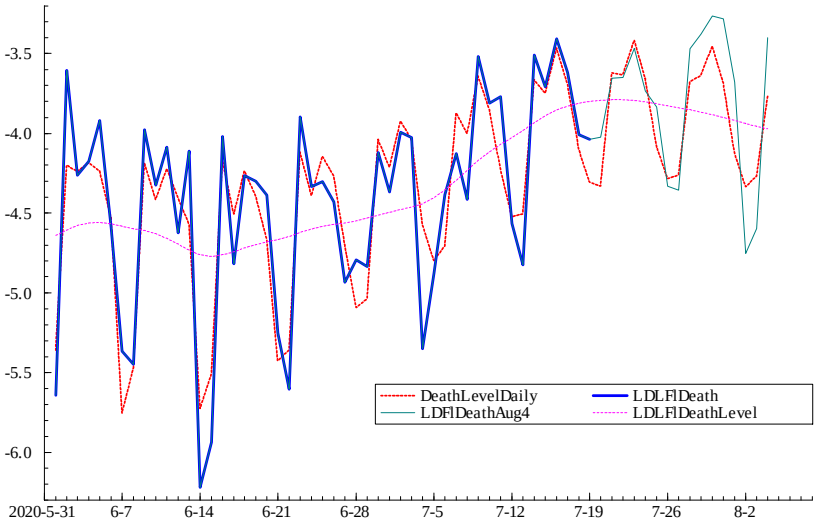

Figure 8: Forecasts and trend of the logarithm of the growth rate of deaths (dots), obtained by using the leading indicator, together with the actual observations from July 20th to August 6th; observations before July 20th shown by thick line.

on July 19th is still positive, and estimating a univariate model up to this point gave $R^e_{t|t,4} = 1.287$ with the forecasts contining on a upward path, overshooting the actual observations.

4 Policy interventions and control groups

The balanced growth framework can be used as the basis for policy evaluation by showing how some variables can serve as control groups for a target variable. This approach is used to investigate the consequences of Sweden's soft lockdown coronavirus policy in the early part of 2020. A comparison is then made with studies based on the method of synthetic control.

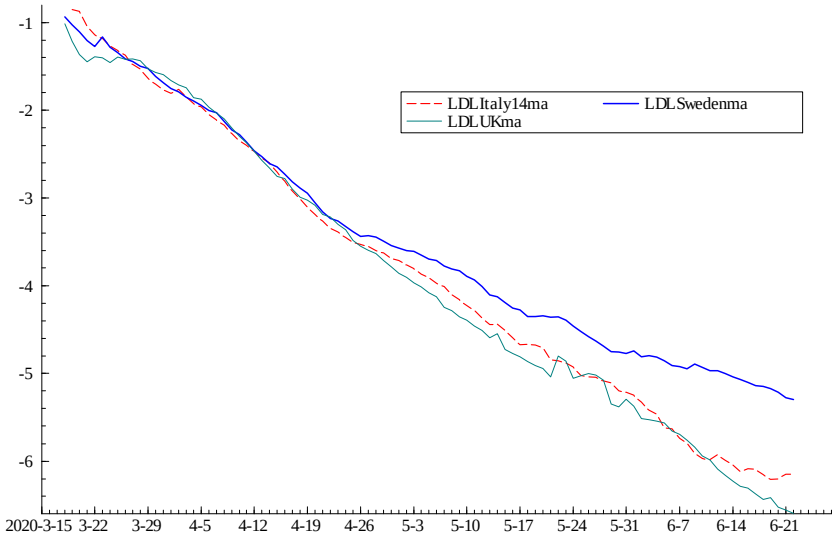

Figure 9: Seven day moving averages of the logarithms of the growth rate from March 18th to July 22nd

4.1 Fewer deaths in Sweden with a full lockdown ?

Sweden did not opt for the full lockdown that other European countries imposed in March. Restrictions were minimal: the government recommended frequent handwashing, working from home, self-isolation for those who felt ill or were over 70 and social distancing[6]; see, for example, Kamerlin and Kasson (2020). Did this policy lead to the number of deaths being significantly higher than it might have been under a full lockdown? To answer this question we need to determine the growth path that Sweden would most likely have followed under a hard lockdown.

The analysis is based on daily deaths in Sweden, UK and Italy (lagged 14 days) from 18th March to 22nd July; by the end of July numbers had become small. A comparison of actual and potential growth paths is best

[6]Carl Bildt, a former prime minister, was quoted as saying "Swedes, especially of the older generation, have a genetic disposition to social distancing anyway."

carried out with the logarithms of growth rates of the cumulative total for the reasons discussed earlier. Although Sweden is much smaller than the UK and Italy, there is no need to take deaths per 100,000 because it follows from the discussion in sub-section 2.3 that standardizing in this way leaves the the growth rate, g_t, unchanged. Because the day of the week effect is very strong, particularly in the UK, the logarithms of growth rates were smoothed with a seven day moving average, centred on the fourth day. The graph in Figure 9 shows that Sweden initially fell at the same rate as the UK and Italy but then started to diverge[7] around 24th April, about a month after the UK lockdown began on March 23rd.

If Sweden had kept on the same growth path as the UK and Italy there would have been fewer deaths. An estimate of the number of deaths under this alternative scenario is given by reference to the forecasting equations in sub-section 2.2. Let $t = m$ denote the date of divergence and let $\widehat{\delta}_t$ denote the values of δ_t estimated for the lockdown growth path using the data on UK and Italy. Since the moving averages are quite smooth, $\widehat{\delta}_t$ was constructed as a simple average of the two countries, rather than by restricted least squares (RLS) as[8] in Harvey and Thiele (2020). Then

$$\widehat{\mu}_{m+j} = \widehat{\mu}_{m+j-1}(1 + \widehat{g}_{m+j}) \simeq \widehat{\mu}_{m+j-1} \exp \widehat{\delta}_{m+j}, \quad j = 1, 2, .., T - m. \tag{13}$$

The initial value is $\widehat{\mu}_m = Y_m$, or a weighted average around that point.

[7] The growth path of deaths in the UK and Italy differs somewhat from the growth path of new cases. The growth rate of $\ln g_t$ for new cases, that is γ_t, drops significantly within a little over two weeks from the start of lockdown; HK estimate the UK fall by fitting intervention variables. A corresponding sharp drop in γ_t is less evident in the deaths data. The divergence of Sweden from Italy and the UK is more a consequence of the Swedish γ_t increasing, rather than the $\gamma_t's$ falling for the other countries.

[8] The general methodology is to select a set of controls from a donor pool by using the KPSS test to determine which series are on a balanced growth path with the target. The control group weighting is then determined by RLS. The complication here is that when there is an intervention balanced growth may require lagging some of the series.

Solving the recursion gives

$$\widehat{Y}_T = \widehat{\mu}_T = Y_m \prod_{j=1}^{T-m} (1 + \widehat{g}_{m+j}) \ \simeq Y_m \exp \sum_{j=1}^{T-m} \widehat{\delta}_{m+j} \tag{14}$$

as the estimated total number of deaths, up to time T, under the lockdown scenario. The estimated number of deaths after time m is $\widehat{Y}_T - Y_m$ while the actual is $Y_T - Y_m$. Here T is July 22nd; the number of deaths after that is relatively small.

The total on April 24th was 2236 and using formula (14) gives an estimate of 4062 for July 22nd as opposed to an actual figure of 5722, a difference of 1660. The sensitivity to the initial value can be gauged by noting that the estimates using the totals two days before and two days after April 24th are 3808 and 4378 respectively.

One way of reducing the dependence on the starting value is to estimate the underlying total for Sweden using formula (14) with $\widehat{g}_{m+j}$ replaced by the actual Swedish values. This gave a total of 5657. The ratio of $\widehat{Y}_T$ for the lockdown control group to that of Sweden is $1.816/2.530 = 0.718$. For $\widehat{Y}_T - Y_m$ it is $0.816/1.530 = 0.533$. This implies that the actual increase from April 24th, which was 3486, could have been 1902. The first method gave $4062 - 2236 = 1826$. The overall conclusion is that, between April 24th and July 22nd, there were perhaps forty to forty-five per cent more deaths than there might have been under a more stringent lockdown of the kind implemented in the UK and Italy.

It is worth noting that although Sweden may have had more deaths under its soft lockdown, this does not mean a higher death rate than countries which had a hard lockdown. On Sept 4th, the figures for deaths per one million for Sweden were 577 as against 611 for the UK and 587 for Italy. The rates for Denmark, Norway and Finland were 108, 49 and 61 respectively, but this should not lead one to infer that the soft Swedish lockdown resulted in a death rate of perhaps ten times what it might have been.

The number of deaths in Denmark is too small to allow a full analysis based on the logarithms of growth rates. The variability is high and after mid-May there are often days when no deaths occur. Numbers in Norway and Finland are lower still. However, up to the end of April the logarithm of the growth rate for Denmark is informative. Figure 10 shows the logarithms of the growth rates for Sweden, Italy, UK and Denmark. Denmark is on a similar growth path to that of the other countries but it is lower than the UK because coronavirus may have arrived earlier and lockdown was imposed on March 13th; the gap is consistent with Denmark leading the UK by about a week. During this period deaths in Denmark were much lower than in Sweden even though they were on the same growth path until close to the end of April. This difference therefore seems to be for reasons not directly connected to the policies of the two countries on lockdown.

On April 30th 2714 deaths had been recorded in Sweden as against 443 in Denmark, a ratio of 6.13. On April 24th the figures were 2236 and 394, a ratio of 5.68. (But bear in mind that the population of Sweden is 1.76 times that of Denmark so in per capita terms the ratio is closer to three.) On July 22nd the ratio of Swedish to Danish deaths had risen to 9.36. However, the ratio of the lockdown estimate of 4062 to the 611 Danish deaths is only 6.64 which is not far from the ratio at the end of April. Thus the estimate of the number of deaths obtained using the control group seems quite plausible. The conclusion is that for reasons unconnected with lockdown policy the death rate per head in Sweden was about three and a half times that in Denmark. The less stringent lockdown then raised this ratio to nearly five and a half.

4.2 Synthetic control

A number of researchers have anaysed the Swedish experience using the method of synthetic control (SC). The recent paper by Cho (2020) is a careful and thoughtful analysis, containing a number of references to earlier pa-

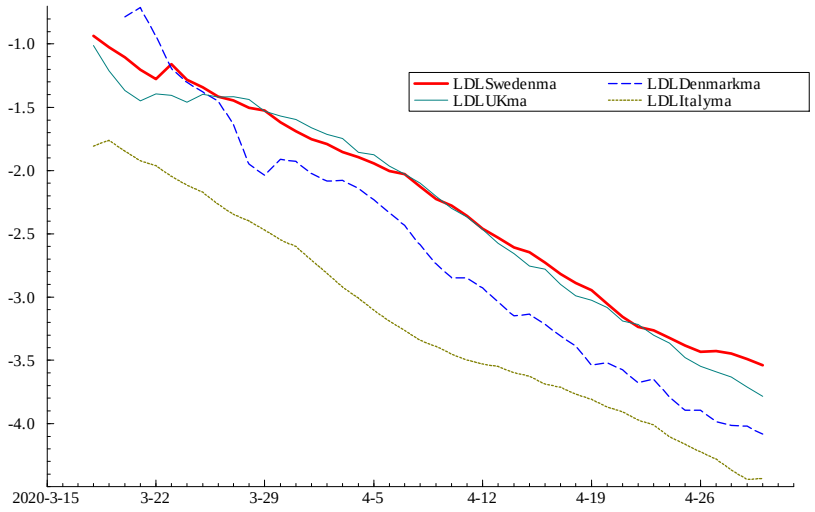

Figure 10: Seven day moving averages of the logarithms of the growth rate from March 18th to April 30th.

pers on the topic. Cho uses daily infection case data per million people to construct a synthetic control variable for Sweden using observations from February 29th to March 24th. The countries and their SC weights were: Finland (0.49), Greece (0.24), Norway (0.22), Denmark (0.03) and Estonia (0.02). The choice of these countries, with the exception of Greece, is not unexpected[9]. Cho concludes that, for the 75 days post-lockdown days, from March 25th until early June, synthetic Sweden is 75% lower than actual Sweden. The SC method cannot be applied directly to deaths because, as note above, the numbers for the key control group candidates are too small so Cho goes on to examine excess deaths by combining the analysis of new cases with weekly data on excess mortality. He concludes that excess deaths were about 25% less in synthetic Sweden as compared with actual Sweden. What is striking is that in the balanced growth analysis the reduction in

[9]In an earlier study, Born et al (2020) selected a somewhat different group, namely the Netherlands (0.39), Denmark (0.26), Finland (0.19), Norway (0.15) and Portugal (0.01).

deaths is quite close, at 29%, and converting to excess deaths might end up with a figure that is closer still.

Cho, in common with other SC researchers like Born et al (2020), uses raw cases numbers, standardized for population. However, the logarithm of the growth rate could also be used and since this yields better behaved time series it is the more appropriate path to take. It would be interesting to see if it yields the same SC group. This seems unlikely. Overall the balanced growth approach is simpler and more transparent. Harvey and Thiele (2020) reach the same conclusion in their analysis of the seminal SC applications in Abadie et al (2010, 2015).

5 Conclusion

The aim of this article has been to provide a methodological framework for the statistical analysis of the relationship between time series of the kind that are relevant for tracking and forecasting epidemics and analysing the effects of policy. The examples illustrate how the methods may be applied in practice, although a degree of caution is needed in interpreting the results because of data revisisions and different definitions of what constitutes a Covid-19 death.

The growth path of an epidemic is best captured by fitting a stochastic trend to the logarithm of the growth rate of the cumulated series. When two series are on a balanced growth path, the difference between them is stationary. The relationship between deaths from coronavirus in the UK and Italy in the first half of 2020 is a good example of balanced growth, with deaths in Italy fourteen days earlier providing a leading indicator for deaths in the UK. A bivariate state space model takes full account of the dynamics in both series and, by extracting the common underlying trend, yields estimates of the daily growth rate of an epidemic and the associated value of R_t.

The balanced growth model was extended by including a random walk

component. This allows the growth path of the leading indicator to deviate from the growth path of the target series. A model of this kind linking deaths to new cases in Florida was estimated for the period covering the second wave in early summer 2020. The forecasts made for deaths while they were still rising are remarkably successful in picking up the subsequent downward movement.

Policy evaluation can be carried out by using some series as control groups for others. A common trend or, better still, balanced growth is the key ingredient. The Swedish policy response to coronavirus provides an example of the methodology. It is shown that the average of the growth paths of deaths in the UK and Italy yields a suitable control group for deaths in Sweden. The Swedish growth path is initially the same as those of the UK and Italy but it diverges towards the end of April. The difference in the growth paths then enables the implications of the Swedish soft lockdown policy to be assessed. The analysis suggests that, between April 24th and July 22nd, there were perhaps forty to forty-five per cent more deaths than there might have been under a more stringent lockdown of the kind implemented in the UK and Italy.

Acknowledgements

Comments and suggestions from Leopoldo Catania, Stanley Cho, Jagjit Chadha, Radu Cristea, Paul Kattuman, Michael Höhle, Christopher Howe, Peter Kasson, Jonas Knecht, Rutger-Jan Lange, Daniel Mackay, Franco Peracchi, Jerome Simons and Herman van Dijk are gratefully acknowledged; of course they bear no responsibility for opinions expressed or mistakes made. Some of the ideas were presented at a Keynote talk at the (virtual) 40th International Symposium on Forecasting in October 2020 and at a NIESR conference in November 2020. The work was carried out as part of the University of Cambridge Keynes Fund project Persistence and Forecasting in Climate and Environmental Econometrics.

6 References

Abadie, A., Diamond, A., and J. Hainmueller. (2010) Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105, 493-505.

Abadie, A., Diamond, A., and J. Hainmueller. (2015) Comparative politics and the synthetic control method. *American Journal of Political Science*, 59, 495-510

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

Born, B., A. M. Dietrich and G. J. Müller (2020). Do lockdowns work? A counterfactual for Sweden. *Covid Economics,* 16, 1–22.

Cho, S-W. (2020). Quantifying the impact of nonpharmaceutical interventions during the COVID-19 outbreak: the case of Sweden. *Econometrics Journal*, 23, 323–344. doi: 10.1093/ectj/utaa025

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

Harvey, A. and P. Kattuman (2020a). Time Series Models Based on Growth Curves with Applications to Forecasting Coronavirus. *Harvard Data Science Review*. Special issue 1 - COVID -19. https://hdsr.mitpress.mit.edu/pub/ozgjx0yn

Harvey, A. and P. Kattuman (2020b). A Farewell to R: Time Series Models for Tracking and Forecasting Epidemics. *CEPR working paper*, Issue 51, 7th October. https://cepr.org/content/covid-economics

Harvey, A.C. and S. Thiele (2021). Co-integration and Control: assessing the impact of events using time series data. *Journal of Applied Econometrics,* 36, 71-85.

Ioannidis, J.P.A. , S. Cripps and M.A. Tanner (2020) Forecasting for COVID-19 has failed. *International Journal of Forecasting*, https://doi.org/10.1016/j.ijforecast.202

Kamerlin, S. C. L. and P. M. Kasson (2020). Managing Coronavirus

Disease 2019 Spread With Voluntary Public Health Measures: Sweden as a case study for pandemic control. *Clinical Infectious Diseases*, ciaa 864. https://doi.org/10.1093/cid/ciaa864

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

Kwiatkowski, D., Phillips, P.C.B., Schmidt P. and Y. Shin (1992). Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root: How sure are we that economic time series have a unit root? *Journal of Econometrics* 44, 159-78.

Onder, G. (2020). Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy. *Journal of the American Medical Association*, 323, 1775-6.

Stock, J. and M. Watson (1988). Testing for Common Trends. *Journal of the American Statistical Association*, 83, 1097-1107.

Wallinga, J. and M. Lipsitch (2006). How Generation Intervals Shape the Relationship Between Growth Rates and Reproductive Numbers. *Proc.R. Soc. B.*, 274, 599-604.

A Data sources

The data for European countries was obtained from the European Centre for Disease Prevention and Control (ECDC) website, https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide,. For Florida the source was : https://covidtracking.com/data. The data were obtained at the end of August and the beginning of September. Data can be subject to revisions. For example the UK definition of deaths was changed in August to include only people who had a laboratory-confirmed positive COVID-19 test and had died within 28 days of the date the test result was reported. Before that it included anybody who had ever tested positive for

COVID-19 no matter how long before the actual death.

Case-fatality statistics in Italy are based on defining COVID-19–related deaths as those occurring in patients who test positive for SARS-CoV-2 viaRTPCR, independently of pre-existing diseases that may have caused death. This method may have resulted in overestimation; see Onder (2020).