Forecasting the COVID-19 Epidemic: The Case of New Zealand[*]

Paul Ho
Federal Reserve Bank of Richmond[†]

Thomas A. Lubik
Federal Reserve Bank of Richmond[‡]

Christian Matthes
Indiana University[§]

October 21, 2020

Abstract

We estimate a statistical model for COVID-19 cases and deaths in New Zealand. New Zealand is an important test case for statistical and theoretical research into the dynamics of the global pandemic since it went through a full cycle of infections. We choose functional forms for infections and deaths that incorporate important features of epidemiological models but allow for flexible parameterization to capture different trajectories of the pandemic. Our Bayesian estimation reveals that the simple statistical framework we employ fits the data well and allows for a transparent characterization of the uncertainty surrounding the trajectories of infections and deaths.

JEL Classification: C32, C51

Key Words: Bayesian Estimation, Forecasting, COVID-19, New Zealand

[*] James Lee provided exceptional research assistance. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[†] Research Department, P.O. Box 27622, Richmond, VA 23261. Email: paul.ho@rich.frb.org.

[‡] Research Department, P.O. Box 27622, Richmond, VA 23261. Email: thomas.lubik@rich.frb.org.

[§] Wylie Hall, 100 South Woodlawn Avenue, Bloomington, IN 47405. Email: matthesc@iu.edu.

1 Introduction

The COVID-19 pandemic has affected the world economy to a degree not seen since the Great Depression. GDP has collapsed by 10% or more in almost all countries affected by the coronavirus, driven by government-mandated lockdowns or voluntary measures by households and businesses to avoid in-person transactions. Similarly, in most countries unemployment has reached historically high levels. Collapses of such economic magnitude are rarely seen outside of major calamities such as natural disasters and wars and therefore present a momentous challenge for policymakers.

In order to stop the as-of-now still temporary decline in activity from turning into a full-blown depression, the spread of the virus has to be stopped. However, this has proved to be vexing to epidemiologists, economists, and policymakers alike. Early on, many epidemiological models struggled to make plausible forecasts of how COVID-19 would evolve. For one, the coronavirus was sufficiently different from other contagions that the basic characteristics needed to calibrate theoretical models were difficult to establish. In addition, data were sufficiently sparse at the beginning of the outbreak that statistical inference proved problematic. Finally, because of the exponential nature of the growth of infections, miscalculations at the start of the epidemic can have dramatic effects on longer-run forecasts.

In this paper, we estimate a statistical model of the spread of the coronavirus and forecast the path of infections and deaths caused by COVID-19 in New Zealand. We focus on documenting the uncertainty surrounding the estimates and projections. We argue that understanding the source of uncertainty is an important step in making public health decisions. In addition, the Bayesian approach utilized in our framework allows us to characterize the inherent uncertainty in modelling a pandemic in a consistent manner. New Zealand is of special interest for modelers and forecasters since it went through a full epidemiological cycle.[1] It therefore offers the opportunity to evaluate the fit of any model and its real-time forecast performance given a full set of relevant data.

Our statistical model for the number of infections over the course of the pandemic is based on Ho et al. (2020). The model's key feature is that it almost exclusively relies on the statistical properties of the observed data. While we do not impose the specific relationships that are implied by theoretical models drawn from epidemiology (e.g., Atkeson (2020); Eichenbaum et al. (2020); Fernández-Villaverde and Jones (2020)), we let our choice of functional forms be guided by the typical behavior of an infectious disease over time. This allows us to gain flexibility in modeling the epidemic and thereby avoid the potential pitfalls

[1] At the time of writing in the middle of August 2020, an apparently isolated cluster of new infections has appeared in Auckland, with reports of up to 30 new cases. It is not clear as of this point whether this is the start of a "second wave" similar to what has occurred in the U.S.

of imposing strict behavior of the contagion.

Epidemiologists have long studied the spread of infectious diseases, using both increasingly complex theoretical models and also more purely empirical frameworks. We contribute to the latter by utilising the toolkit prevalent in the analysis of economic data. In that, our work is similar to Harvey and Kattuman (2020), Li and Linton (2020), and Liu et al. (2020), who also use statistical models to forecast the pandemic. A recent empirical contribution to modelling the pandemic that is similar in spirit to ours is Gibson (2020). He uses a statistical model because of identification issues inherent in theoretical epidemiological models that are especially concerning when samples are small as in the case of New Zealand. In contrast to our paper, Gibson (2020) uses inferences from the dispersion of Covid-19 cases in U.S. counties, whereas we focus on aggregate New Zealand data.

We proceed as follows. In the next section, we present our statistical model for infections and deaths, discuss our modelling choices and describe the data and estimation procedures. Section 3 contains the presentation of the results. Section 4 concludes.

2 An Empirical COVID-19 Model

We begin by developing a general specification for the evolution of infections and deaths over the course of an epidemic. In contrast to much of the recent literature focusing on structural models of the pandemic, we take a statistical approach. Our model is almost entirely data-driven, in that it tries to match the underlying time series properties of the data at hand while, at the same time, relying on guidance from epidemiological insights on how an epidemic runs its course. To that end, we specify flexible functional forms consistent with the typical dynamics of epidemics.

Model for Infections The time path of the number of infections during an epidemic follows a typical pattern. When a pathogen enters a population that is susceptible to infection, the number of infected cases is initially low. However, the growth rate of new infections is high and tends to rise sharply at an exponential rate since each infected person creates a chain of new infections. At some point, however, the pathogen runs out of susceptible hosts, either because they are already infected, are immune, or they are simply not physically present because of health policies such as social distancing. At this inflection point, the growth rate of infections falls until it eventually declines to zero.

Our statistical model replicates these broad patterns using a flexible functional form that describes the path of infections over time as depending on the current and the lagged levels of the number of infections. The model is loosely parameterized, whereby the parameters are estimated to provide best fit of the model specification to the available data. In contrast to theoretical epidemiological models, our specification has more leeway to go where the data tell it to and is not constrained by precise theoretical relationships that may be specified incorrectly.

We specify a model in which the growth rate of the cumulative number of cases depends on the current cumulative number of cases. In particular, denoting the cumulative number of cases normalized by population in period t by C_t, we consider:

$$\Delta \log C_{i,t} = \log(1+\gamma)\frac{\phi(C_{t-1};\alpha,\zeta,\eta)}{\phi(10^{-5};\alpha,\zeta,\eta)}\exp(\varepsilon_t^C) \tag{1}$$

$$\phi(C;\alpha,\zeta,\eta) \equiv \exp[-C^{-\alpha} - (\zeta^\eta - C^\eta)^{-2}] \tag{2}$$

where $\varepsilon_t^C \sim \mathcal{N}(0,\sigma^2)$.[2]

The model is set up to flexibly match the trajectory of cases.[3] Initially, the rate of growth is approximately exponential. When a fraction 10^{-5} of the population has been infected, the growth rate in the absence of shocks is γ. The parameter α determines how the growth rate of C_t increases or decreases in the early stages of the pandemic, capturing the appearance of large clusters or the effects of social distancing measures. The parameters ζ and η determine the long-run number of cumulative cases and the speed at which the population converges to that number. This terminal state of the pandemic likely depends on factors such as demographic variables or the effect of policies that attempt to mitigate the spread of infections. Finally, ε_t^C is a shock that allows for deviations from the model predictions, arising due to randomness in how the virus spreads.

Identification of the model parameters is based on the growth rate and changes in the growth rate of infections. Early in an epidemic, the data typically show exponential growth, rapid and increasing. After some time, as the stock of susceptible hosts starts getting smaller, the rise in the growth rate decelerates until it reaches a peak. Subsequently, the growth rate of new infections declines. These three distinct phases of an epidemic can be associated with distinct parameters in our model, which are thus identified from the data flow.

This is also where a problematic aspect of any epidemiological model lies. At first, data are sparse, but the underlying course of the infection is such that it should be easy to forecast. Put differently, the epidemic develops a very strong trend with exponential growth. Simply extrapolating from this growth trend would produce good forecasts for a while – until the

[2] Ho et al. (2020) allow for AR(1) errors in the estimation on U.S. data. For reasons of parsimony, given the small sample available in New Zealand, we impose i.i.d. errors, especially since preliminary estimates did not suggest strong serial correlation

[3] Ho et al. (2020) show simulation results that detail how the various model parameters affect and determine the pattern of the infection path.

spread starts slowing down and gravitates towards an inflection point. While it is known from epidemiological models based on the course of previous epidemics that there is an inflection point, estimates from the sparse initial data are highly uncertain. Moreover, theoretical and statistical epidemiological models are sensitive to small variations in parameters and suffer from identification problems as highlighted and documented by Koroloev (2020). It is in this sense that model estimates and forecasts should be interpreted with much caution at the beginning of the pandemic, and that uncertainty at this stage should explicitly be considered when making public health policy decisions.

Modeling Deaths In addition to modeling infections, we also consider the mortality rate. Fundamentally, the number of deaths is a function of the number of infections. Not all infections are fatal. Moreover, an observed death is the outcome of a process that can vary over time. We thus assume that the number of deaths at any given day is proportional to the average number of observed infections over a time period. This captures the idea that there is a minimum number of days that pass after an initial infection can result in a fatality.

We assume that the number of deaths is proportional to the average number of cases over some window of n days ending h days ago:[4]

$$D_t = \lambda \frac{C_{t-h} - C_{t-h-n}}{n} + \varepsilon_t^D \tag{3}$$

$$\varepsilon_t^D \sim \mathcal{N}(0, \frac{C_{t-h} - C_{t-h-n}}{n}\omega^2). \tag{4}$$

The constant λ is the death rate. The variance of the shock ε_t^D is proportional to the average number of cases in the window, reflecting a higher variability in the absolute number of deaths from day to day when there are more infections. ω is a scale parameter. We take h and n as parameters to be estimated, reflecting uncertainty about the time between testing positive for Covid-19 and dying from the virus.

Data and Estimation The New Zealand Covid-19 data used in the estimation are obtained from the Institute of Environmental Science and Research's Covid-19 Dashboard (`https://nzcoviddashboard.esr.cri.nz/#!/`). We report new cases and total cumulative cases from February 25, 2020 to August 10, 2020 for the entire country. Total cumulative

[4] Ho et al. (2020) use a more general mortality function given the wider variability and availability of U.S. data, where mortality depends on testing. Alternatively, the death rate λ could be specified as some function of time. The cycle in New Zealand is very short, however, so that the time trend may not be as important as in the U.S. The somewhat simpler specification that we implement here arguably provides a good enough description, especially given the small number of fatalities in New Zealand.

cases includes what is classified as probable cases along with confirmed cases.[5]

We estimate the model using Bayesian methods. In terms of computational algorithms we use a standard Metropolis-Hastings algorithm with a random walk proposal (see, for example, Gelman et al. (2013)). Further details of the estimation can be found in Ho et al. (2020). We obtain 2.5 million Monte Carlo draws from posterior distributions of each of the parameters $\{\gamma, \alpha, \zeta, \eta, \sigma\}$.

We use the estimated models for infections and death to forecast these respective variables. Since New Zealand had passed through a full epidemiological cycle by our sample end date of August 10, 2020, the forecasting exercise is conducted as a real-time (pseudo) out-of-sample exercise, where we estimate the model up to a specific date and forecast from that date out. We do so in a rolling window fashion where we show how the forecasts and its surrounding uncertainty have changed over time. We show the fit of the respective model and its error bands by choosing some initial condition for the start of the sample and then iterate forward without shocks (i.e., taking $\varepsilon_t^C = 0$) from a fixed initial condition. The initial condition is chosen to minimize the squared error between the data and the model-implied trajectory of cases under the mean parameter estimate and is fixed across all parameter draws. The error bands thus capture the amount of parameter uncertainty. For our forecasts, we simulate paths for $\{\varepsilon_t^C\}$ for a sub-sample of the Monte Carlo draws. We take the respective last data point as an initial condition and iterate forward for each parameter draw, together with an associated sequence of shocks, in order to account for the additional uncertainty from measurement errors.

3 Results

We present empirical results for infections and deaths, reporting both daily incidence and cumulative numbers. We report estimation results and forecasts in Figure 1 for the epidemic cycle that New Zealand experienced between February 25 and May 8, 2020. That is, the model is estimated on data up to May 8, from which we then forecast out until August 10.[6]

[5] There is likely to be measurement error with a possible undercount in this variable since case numbers depend on the amount of testing. Moreover, it is a priori unclear in which direction the measurement error goes for recorded deaths. This is a general problem that researchers have to confront at this stage of the pandemic. In the case of New Zealand, at least probable and confirmed cases are reported. In principle, this could be addressed by including measurement error in the specification explicitly, but we likely pick up some aspects via the shocks in our model.

[6] After May 8 the data record a few numbers in the single digits of additional new cases. However, as the graph shows, and as news reports indicated, the epidemic was largely contained by then. We also estimated the model for data up to August 10; its fit, however, deteriorated markedly, giving an arguably incorrect picture of the path of the epidemic. The underlying reason is that we assume shocks can only scale the number of cases up and down. Consequently, when the model indicates that the epidemic cycle is over, a

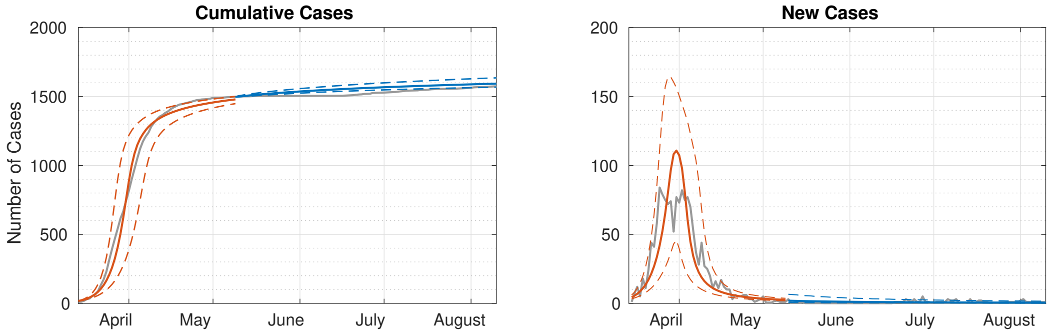

Figure 1: Model fit and forecasts (median and 95% error bands) for cumulative and new cases in New Zealand

We show the posterior median and the outer bands of the 95% coverage region.

The figure shows that the model fits the data well in tracking the time path of total cases. Uncertainty widens as more cases are being found around the middle of March as can be seen from the panel with new cases. There is a peak around March 20, but new case numbers drop before rising again to a second peak on April 5. The model interprets a large part of the trough on March 30 as measurement error, which widens the uncertainty around this inflection point of the infection. Notably, the model estimates March 30 as the peak infection date, after which new cases decline at essentially the same rate as they increased. The distribution of new cases appears almost bell-shaped, displaying a more rapid decline in new cases than many other countries. Arguably, this is due to the aggressive lockdown measures ordered by the New Zealand government.[7] Finally, the forecasts show that the epidemic has tightly stabilized around 1,500 total cases, with the remaining uncertainty deriving from measurement error shocks.

We now look at how the estimates and forecasts have evolved over time, from the onset of the epidemic when few data were available until it ran its course. Figure 2 shows the 95% error bands for forecasts of new cases and cumulative cases, using data ending April 10, April 17, and May 8, 2020. These expanding window forecasts confirm that forecasts perform reasonably well and that the model is able to adapt to new data. Specifically, error

huge shock is needed to fit any incoming data, even if there are only a small number of new cases. As a result, the estimation likely puts too much weight on these end-of-cycle observations, therefore biasing the inference. These additional results are available from the authors upon request.

[7] Our estimated peak of new cases follows the government's introduction of Alert Level 2 on March 21, Alert Level 3 on March 23 and Alert Level 4 on March 25. As Figure 1 shows, the estimated peak is not reflected in the actual data. That is, the model specification provided enough momentum from rising case numbers to impute a higher peak than actually occurred. It is tempting to speculate that this is due to the rapid imposition of alert levels and ever tighter lockdowns. An analysis of the value of lockdowns and social distancing policies is beyond the scope of this paper since it requires a more structural framework or much richer data on social distancing measures than is available for New Zealand.

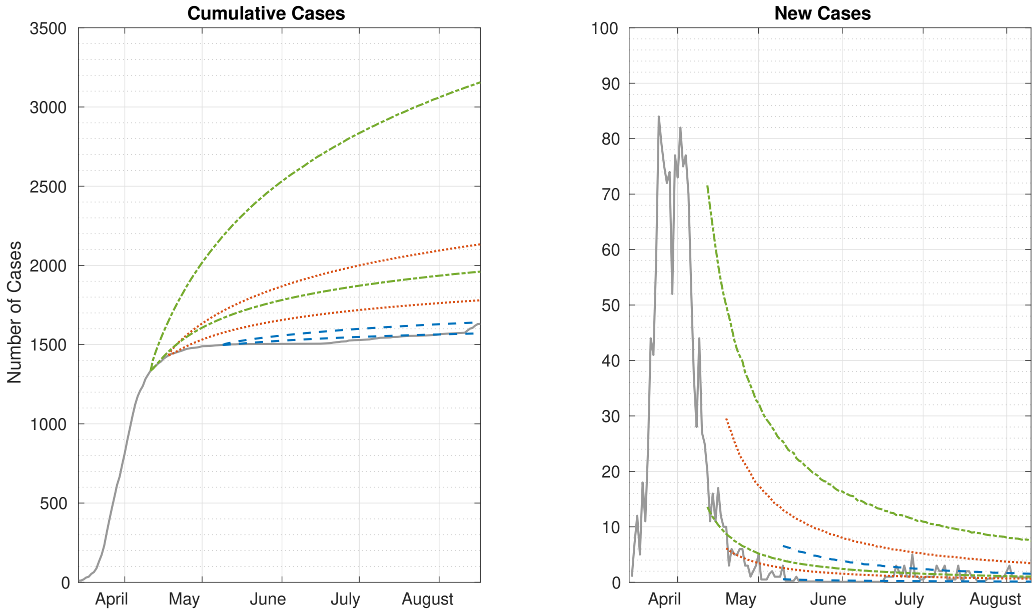

Figure 2: Forecasts (95% error bands) for cumulative and new cases in New Zealand using data ending April 10, April 17, and May 8, 2020.

bands tighten as more data become available. April 10 is a date that is already past the peak and thus after the inflection point of the epidemic curve - in hindsight. The model has not fully incorporated this since there is still a chance that the drop in new cases could be due to a measurement error. This is reflected in very wide coverage regions, ranging from 600 to 1,800 total cases over the full projection horizon.[8] However, one week later, the overall picture becomes much clearer as the error bands tighten and shift downward.

Finally, we also consider estimates and forecasts of deaths. One aspect of the epidemic in New Zealand is that the number of fatalities has been extremely low, in absolute but also in relative terms compared to population. From a statistical perspective this is a challenge since the number of observations is small and likely unevenly distributed over the sample. In that context it is of note how the estimates of the parameters that represent the time lags between cases and deaths change across the samples ending April 10, April 17, and May 8, 2020. Our posterior median estimates of h and n for these sampling periods are 6, 9, and 12, and 4, 4, and 5, respectively, which indicates that the time until death from infection is

[8] When estimated on data up to March 30, the error bands are extremely wide, ranging from zero to 200,000 total cases. At that point, the cumulative number of infections was barely 700. As discussed above, the high growth rates at the onset of an epidemic are a challenge for modellers and forecasters as predictions can be considerably off since even small parameter uncertainty can lead to large long-run forecast uncertainty. This is especially exacerbated in non-linear environments. Hence, a proper treatment of uncertainty and regular updating of estimates should be of foremost importance.

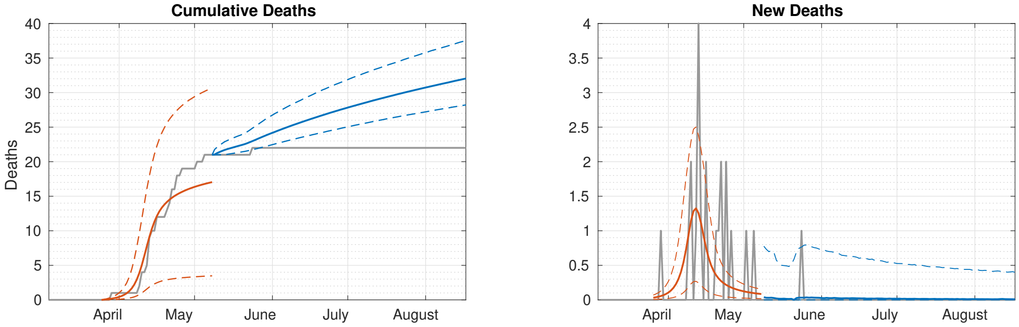

Figure 3: Model fit and forecasts (median and 95% error bands) for cumulative and new deaths in New Zealand

lengthening as time passes, presumably reflecting better treatment.

Figure 3 shows estimates and projections for deaths given a sub-sample that ends on May 8. The median number of daily new deaths predicted by the model after May 8 is around zero and the 90% quantile is below one. However, the estimation and parameter uncertainty is wide over the sample period, reflective of the small number of observations. The model overpredicts the total number of deaths, which again may reflect the strong policy response of the New Zealand authorities. Figure 4 parallels Figure 2 and shows the 95% posterior bands for forecasts of deaths, using data ending April 10, April 17, and May 8, 2020. While the coverage region for cumulative deaths based on April 10 data is quite wide, it quickly tightens once additional data points are reported. Nevertheless, all three forecast horizons miss overall deaths by the end of August.

4 Conclusion

In this paper, we present a statistical time series model to describe and forecast the dynamics of the spread of an infectious disease applied to the case of COVID-19 in New Zealand. The model is constructed to capture the typical pattern of the evolution of infections, which can then be used to forecast the future path of the pandemic. Because of the efforts of New Zealand to contain the COVID-19 outbreak, we have infection and mortality data over a full epidemic cycle, which allow researchers to study the validity of their models. A key aspect of our analysis is a focus on the uncertainty that surrounds both estimates of the model and its forecasts. To that end we rely on Bayesian methods, which provide a straightforward way to quantify uncertainty.

Our analysis shows that a simple statistical forecasting model works remarkably well in capturing the evolution of the epidemic. A distinct advantage of our approach is that

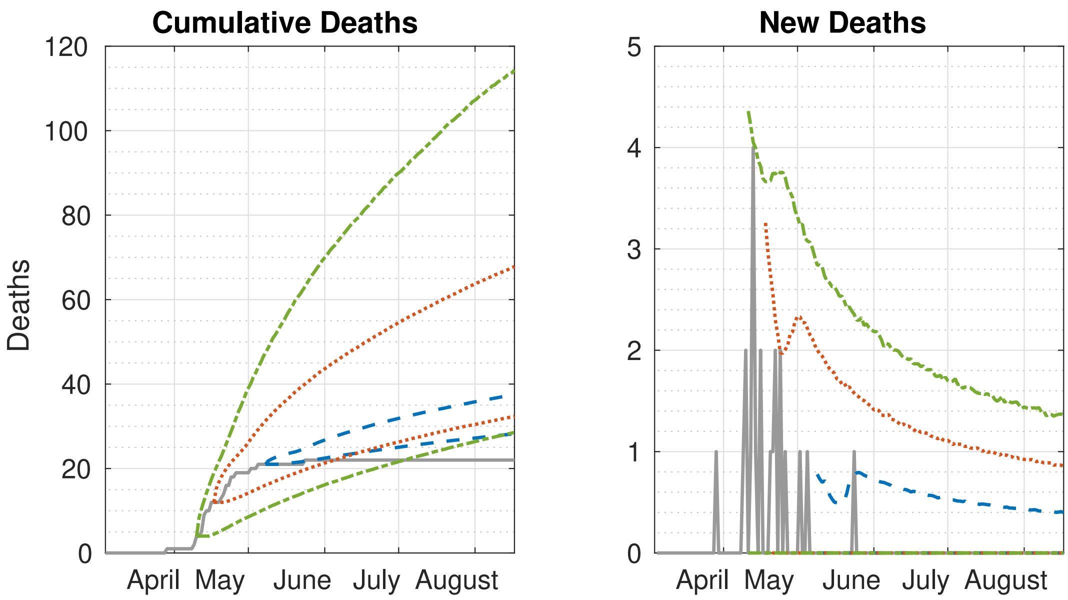

Figure 4: Forecasts (95% error bands) for deaths in New Zealand using data ending April 10, April 17, and May 8, 2020.

it is straightforward to execute and much easier to communicate to policymakers and the public than large-scale statistical models or complicated theoretical models. In addition, quantifying uncertainty as to possible future paths of an epidemic is crucially important. At the beginning of an infection cycle available data are sparse and possibly contaminated by measurement error. Consequently, the degree of uncertainty is potentially enormous, which should be taken into account by decision makers.

Nevertheless, our model has distinct shortcomings. First and foremost, it does not allow for modelling the effects of policy measures to contain the virus spread. At best, frequent re-estimation of the model on incoming data allows us to gain insight into how such measures may have changed the trajectory of the epidemic. In terms of modeling, this can also be addressed by introducing endogenous time variation in the model parameters, which can be made to change as functions of other variables such as metrics for social distancing. Finally, inference could be improved by bringing additional information to bear, such as regional variation in the virus spread as in a panel setting. Ho et al. (2020) introduce these modifications into a specification for the 50 U.S. states and D.C.

References

Atkeson, Andrew (2020), "On Using SIR Models to Model Disease Scenarios for COVID-19." *Quarterly Review*, 41, 1–35.

Eichenbaum, Martin S., Sergio Rebelo, and Mathias Trabandt (2020), "The Macroeconomics of Epidemics." NBER Working Paper 26882, National Bureau of Economic Research.

Fernández-Villaverde, Jesús and Charles I. Jones (2020), "Estimating and Simulating a SIRD Model of COVID-19 for Many Countries, States, and Cities." NBER Working Paper 27128, National Bureau of Economic Research.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin (2013), *Bayesian Data Analysis.* CRC press.

Gibson, John (2020), "Government Mandated Lockdowns Do Not Reduce Covid-19 Deaths: Implications for Evaluating the Stringent New Zealand Response." Working paper.

Harvey, Andrew and Paul Kattuman (2020), "Time Series Models Based on Growth Curves with Applications to Forecasting Coronavirus." *Covid Economics, Vetted and Real-Time Papers.*

Ho, Paul, Thomas A. Lubik, and Christian Matthes (2020), "How to Go Viral: A COVID-19 Model with Endogenously Time-Varying Parameters." Federal Reserve Bank of Richmond Working Paper 20-10.

Koroloev, Ivan (2020), "Identification and Estimation of the SEIRD Epidemic Model for COVID-19." Working paper, Binghamton.

Li, Shaoran and Oliver Linton (2020), "When Will the COVID-19 Pandemic Peak?" Cambridge Working Papers in Economics 2025.

Liu, Laura, Hyungsik Roger Moon, and Frank Schorfheide (2020), "Panel Forecasts of Country-Level COVID-19 Infections." NBER Working Paper 27248, National Bureau of Economic Research.