Analyzing Covid-19 Data with SIRD Models

Peter Turchin
Complexity Science Hub Vienna and University of Connecticut
March 23, 2020

Summary: The goal of this exercise is to estimate the effects of various measures implemented by national governments to slow down and reverse the spread of the Covid-19 epidemic. A direct approach to answering this question is to calculate a temporally varying measure of the rate of change in the number of infected (for example, the basic reproductive number, R_0) and then correlate its observed changes with the control measures implemented by governments. However, one problem with this approach is that it assumes that the true number of infected, I(t), is known (or that the observed number of infected is a constant proportion of the true number). However, we know that the coefficient of detection changes throughout an epidemic, starting at a low level and subsequently increasing. Thus, both control measures and changes in the coefficient of detection affect the dynamics of the directly calculated R_0.

The idea here is to build a process model for the Covid-19 epidemic, based on the SIR framework, and add to it an observation model that translates the true number of infected into the number of registered cases. Analysis of data with such a mechanistic model has advantages over purely phenomenological approaches. By building in the model what is known, we can use sparse data most effectively to estimate what is unknown. We can use this approach to investigate a variety of hypotheses that have been proposed for Covid-19. The question is whether the parameters of both the process and observation models can be simultaneously estimated using available data. If yes, then this approach will allow us to filter out the effects of changing detection rate, and to estimate the impact of an intervention.

In the following sections I describe the modeling approach, discuss its limitations, and illustrate the approach by applying it to the data from South Korea, Hubei (China), and Italy.

Modeling Framework

The starting point is a discrete-time version of a standard epidemiological SIRD model.

Variables:

t	time in days
S(t)	the number of susceptibles
I(t)	the number of infected
R(t)	the number of recovered
D(t)	the number of deaths
N	= S + I + R + D, the total population (a constant)

Equations (suppressing time-dependence on the RHS, thus, S(t) → S):

$S(t+1) = S - bSI/N$

$I(t+1) = I + bSI/N - gI - dI$

$R(t+1) = R + gI$

$D(t+1) = D + dI$

Parameters

b	transmission coefficient between I and S
g	recovery rate of I
d	death rate of I

The exponential growth rate during the initial phase of the epidemic, when $S \approx N$, is $r_0 = b - (g + d)$. Note that this is the (per capita) growth rate per day; thus, its units are day^{-1}. It is different from the basic reproductive number, R_0, which is dimensionless. Epidemic grows when $r_0 > 0$ and dies out otherwise.

Modeling Intervention Impact

I will assume that the effect of an intervention on a model parameter can be captured with a logistic curve. Specifically, I will model an intervention that reduces the transmission coefficient as follows:

$b(t) = b_0\{1 - 1/(1 + \exp[-\theta(t - t_b)])\}$

where b_0 is the initial level of b, t_b is the time of intervention, and θ regulates the steepness of the curve (how rapidly the transition happens). This formulation assumes that b(t) eventually approaches 0, but this assumption can be relaxed by adding another positive constant to the logistic function:

$b(t) = b_{min} + (b_0 - b_{min})\{1 - 1/(1 + \exp[-\theta(t - t_b)])\}$

Multiple interventions can be modeled by "stacking" two or more logistic curves, which results in a staircase-like pattern of change.

Other hypotheses worth investigating include the effect of overwhelmed medical system on the death rate, and, conversely, expanding the capacity to treat patients, which is expected to reduce d.

Observation Model

Because many Covid-19 infectives are asymptomatic or experience mild flu-like symptoms, we know that the official data underestimates I. Additionally, and especially during the early phase of the Covid-19 epidemic, some of the deaths due to it might have been recorded as due to other kinds of pneumonia. For this reason, we need to model the observation process. Let I*(t) be the observed number of infected. Then, and suppressing the time notation,

$I^* = qI$

where q is the probability that an infected individual would be recorded as such. Similarly, the number of recorded recovered patients is $R^* = qR$, because the $(1 - q)$ fraction of them is never detected. I will assume that the recorded number of deaths is close to the actual number, so that $D^* = D$.

The probability of detection is expected to vary with time, starting at a low level before the authority and the public are aware of the epidemic, and growing to a high number depending on how massively asymptomaitc individuals are tested.

Limitations and Potential Extensions of the Basic Model

There are a number of ways in which the basic model can be made more realistic. However, realism comes at the expense of the need to get more detailed data.

Country and Geographical Effects

1. Most obviously, all parameters should be made country-specific.

2. More interestingly, we could estimate the effect of ambient temperature on Cov-d19 transmission. It is currently hypothesized that the optimal temperature is around 10 degrees C. This will become important as we move into summer.

3. The basic model lacks migration terms; these can be added.

Age Effects

We already know that the death rate varies dramatically with the age of the patient. There are also possible effects of age on transmission parameter (the very young and the very old tend to come in contact with fewer people). We can estimate such effects assuming that we can get age-structured data.

Micro-scale ABM

We can also build a detailed agent-based model, which could allow us to capture all kinds of interesting effects, such as spatial proximity, belonging to large well-mixed groups (e.g., attending a big class at the university), etc. This could be a worthwhile direction to go in, assuming that we can get detailed data to empirically base such an ABM.

Process and Measurement Noise

The procedure for fitting models to data (see below) assumes that stochasticity only affects the relationship between "true" data and measured data. In other words, it assumes that Covid-19 epidemics develop entirely deterministically. We should explore the effects of this assumption by adding dynamic (process) noise to the model.

Methods: Fitting the SIRD Model to Data

The CSSEGISandData repository provides national data on three time-series of Covid-19 indicators: daily counts of confirmed cases (C), recovered cases (R), and deaths (D). I used these data to calculate additional time-series:

$I(t) = C(t) – [R(t) + D(t)]$	The number of Infected (aka Active Cases)
$delC(t) = C(t+1) –C(t)$	New cases
$delD(t) = D(t+1) –D(t)$	Daily deaths

All six series (C, R, D, I, delC, delD) provide useful insights into the dynamics of a Covid-19 epidemic. As an initial venture into fitting the (process+observation) SIRD model I attempted to minimize a measure of fit for all six data series. The measure of fit I used is known as the coefficient of prediction

$$predR2 = 1 – \Sigma(Y – X)^2/ \Sigma[X – mean(X)]^2$$

where Y(t) are model predictions (Y stands for C, R, etc), X(t) are data, and mean(X), rather obviously, is the mean of X. This measure, known as the coefficient of prediction, is similar to the coefficient of determination in regression. It also reaches maximum = 1 for perfect prediction. Unlike regression R2, however, predR2 can be negative, when the model predicts data worse than the data mean.

Maximizing prediction R2 is equivalent to minimizing the sum of squared deviations between the predicted and observed. This is a "quick and dirty" approach; a more proper one would be to use either likelihood or fully Bayesian methodology.

The objective function that I used in exploring how model parameters affect the fit to data was simply an average of predR2 for all six data series. My parameter estimation strategy involved two steps. First, I systematically varied model parameters on a regular grid, covering the range of possible values, and looked for a parameter combination that maximized mean predR2. Second, after identifying a general region of parameter space, I fine-tuned the estimates using the nonlinear optimization function in R, optim().

Results

South Korea

I downloaded data from the CSSEGISandData repository on March 21. The period covered by data extends from Jan. 22 (t = 1) to Mar. 20 (t = 59). See Figure 1.

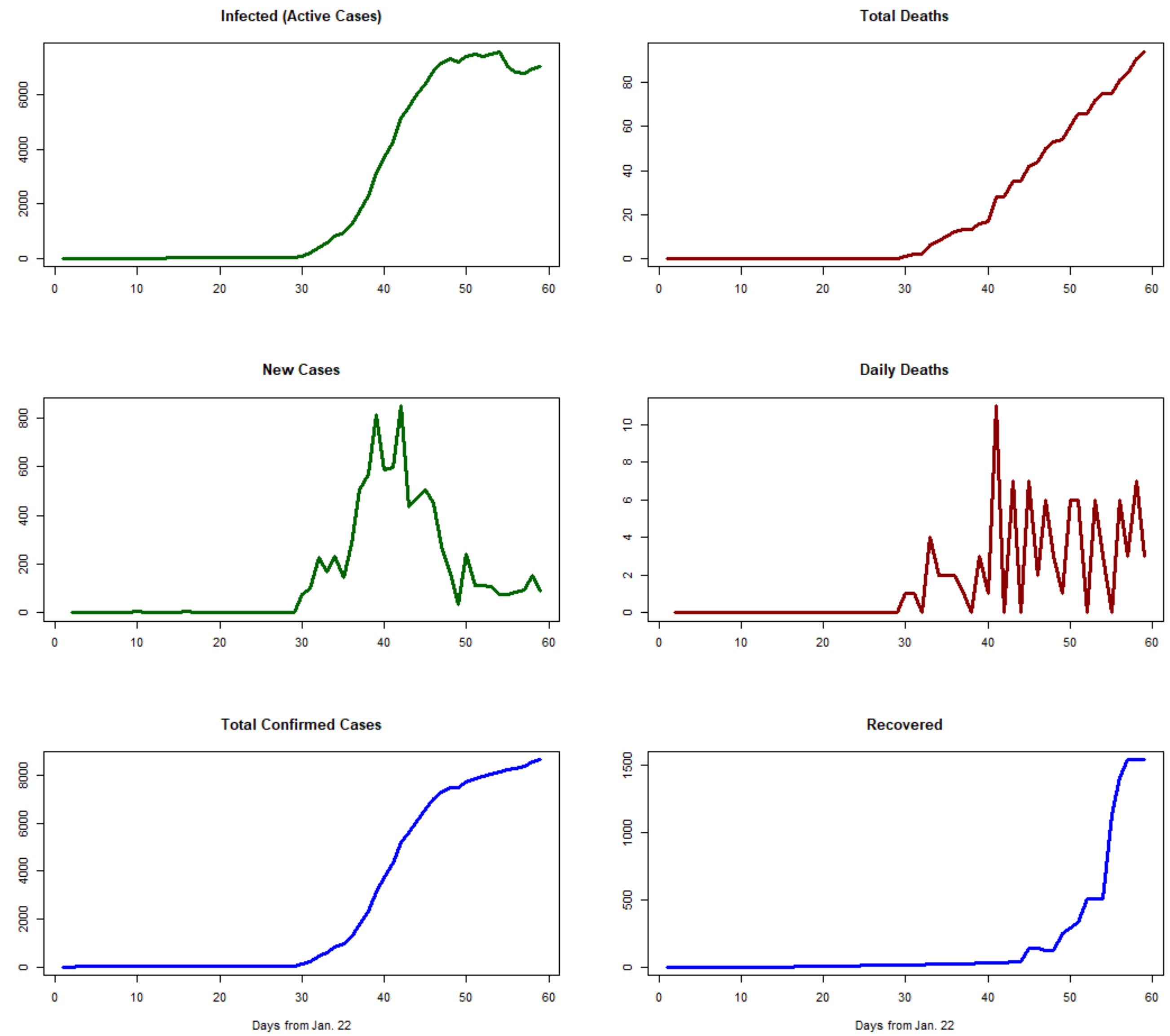

Figure 1. The data on the course of the Covid-19 epidemic in South Korea, Jan. 22 – Mar. 20, 2020. *Source:* CSSEGISandData repository.

The SIRD model fits these data quite well (Figure 2).

Prediction coefficients were generally high or very high, closely approaching 1 for several data series (Table 1). The worst predR2 was for Daily Deaths (del D), but that is as expected given high day-to-day fluctuations in this indicator. Another indicator, for which model trajectory deviates significantly from the data is recovered (R). However, this deviation may be due to a time-varying threshold used by the authorities to declare that a patient is recovered (this is suggested by the faster-than-exponential spike of R following day 55).

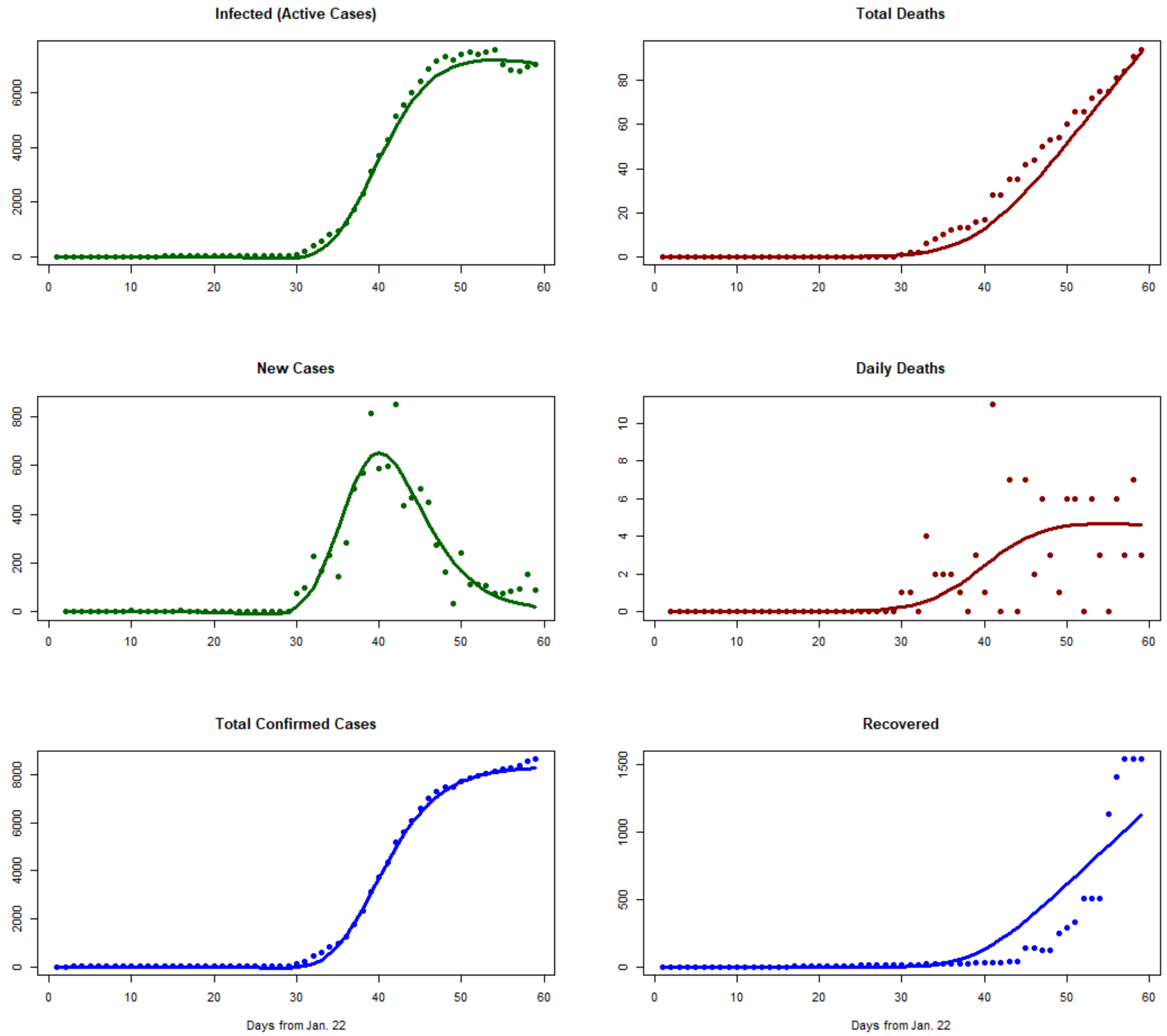

Figure 2. Comparison between model-predicted trajectories (curves) and data (points): South Korea, Jan. 22–Mar. 20.

```
   Var   predR2
     I    0.995
     D    0.968
  delC    0.898
  delD    0.412
     C    0.999
     R    0.789
  Mean    0.843
```

Table 1. Coefficients of prediction for the six South Korea data series.

Table 2 lists the estimated parameters.

Parameter	**Value**	**Explanation**
N	50000000	total population = 50 mln (S. Korea)
I0	0.033	the number of infected at t = 0
beta0	0.432	transmission rate at t = 0 (b_0)
theta	0.234	steepness parameter of the logistic curve
b_date	36.1	day at which the logistic curve for b(t) inflects
gamma	0.0078	recovery rate of infected
delta	0.00044	death rate of infected individuals
q1	0	initial detection rate of infected
q2	0.68	detection rate of infected at the end
theta_q	0.3	steepness parameter for the q logistic curve
q_date	33	day at which the logistic curve for q(t) inflects

Table 2. Estimated parameters.

Finally, we examine what the estimated parameters tell us about the effect of interventions in stemming the Covid-19 outbreak in South Korea. First note that the detection rate was 0 until day 30. What this suggests is that initially the epidemic was growing "below the radar screen" for several weeks. South Korean authorities started testing for Covid-19 in early February, and the scale of testing was massively expanded after Feb. 20, which closely corresponds to day 30 when model-predicted detection rate began increasing.

Second, the infection rate initially was very high, with the exponential rate of increase of I around 0.4 day^{-1} (in other words, daily increase rate of 40%). This parameter began declining after day 25 (mid-February), but reached low levels only close to day 50 (early March).

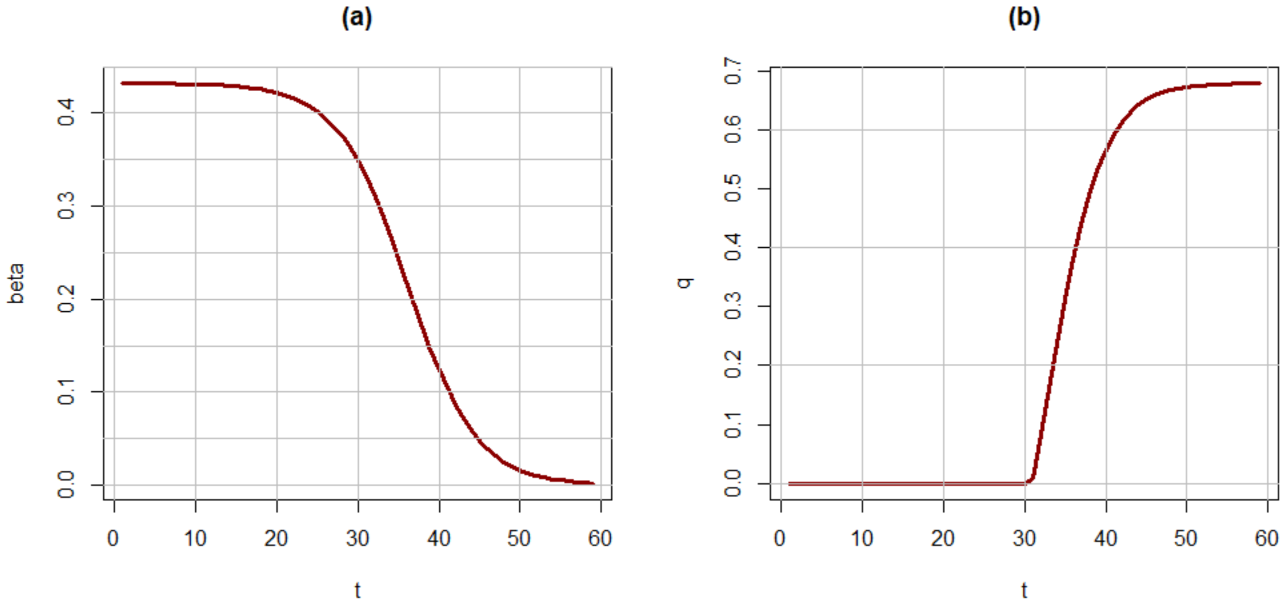

Figure 3. Estimated time-dependence in (a) the infection rate, b(t), and (b) the detection rate, q(t)

Hubei (China)

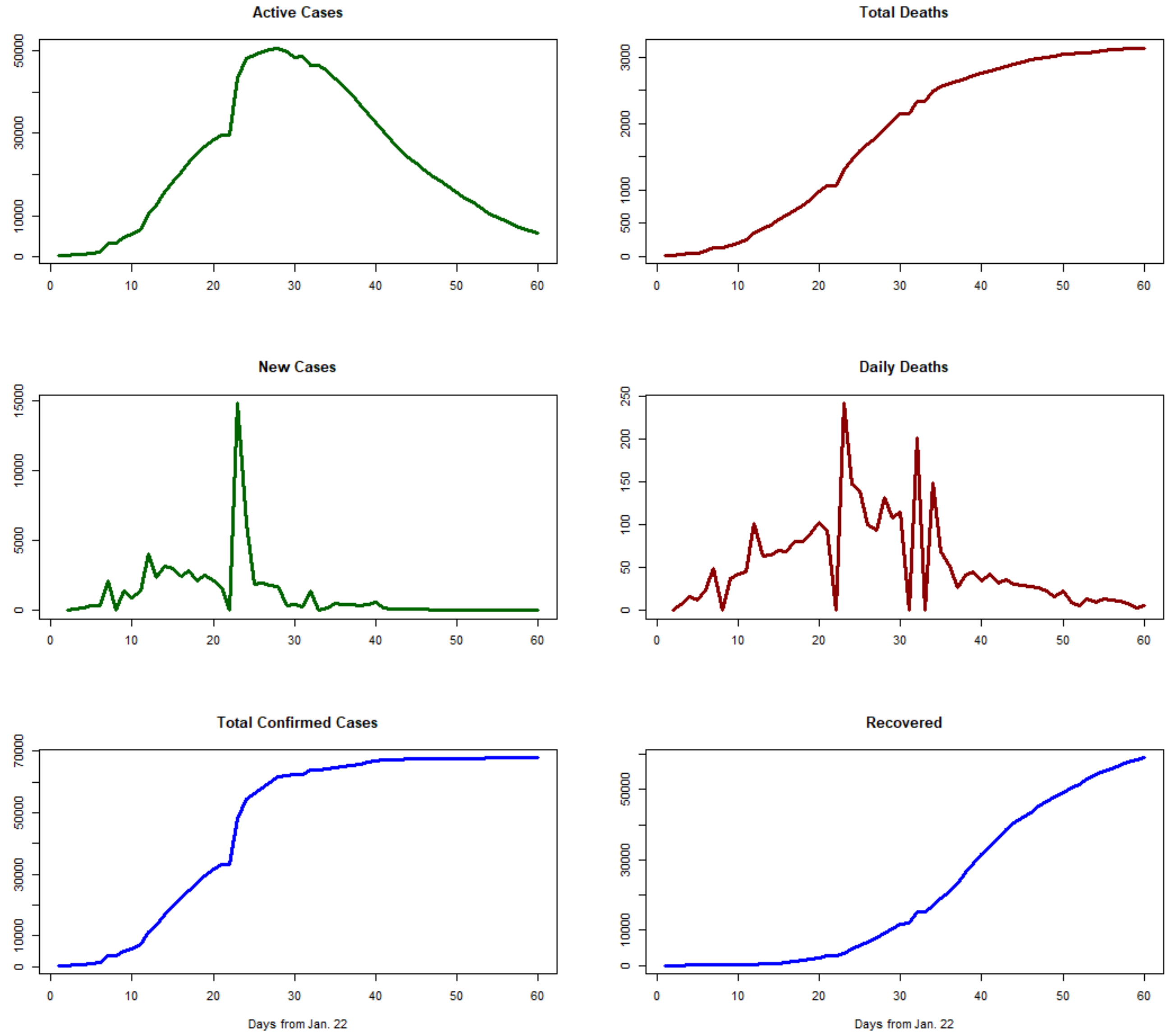

Figure 4. The data on the course of the Covid-19 epidemic in the Hubei Province (China), Jan. 22 – Mar. 20, 2020. *Source:* CSSEGISandData repository.

Analyzing Chinese data offers a substantially greater challenge then the case is for South Korea, because China's National Health Commission has repeatedly changed how it counts coronavirus cases. The biggest change was implemented on Feb. 12, which resulted in the huge spike of New Cases on Day 23 (see Figure 4). Unfortunately, in total, China has revised guidelines on recording confirmed cases numbers at least six times since Jan. 22. I have chosen to incorporate into the model only one change, which had the greatest effect on the reported cases, but this is clearly an area where additional effort could repay in improving the model's performance.

Public policy interventions also involved a series measures. I will focus on two most important ones. The first one is the imposition of *cordon sanitaire* around Wuhan (the capital of Hubei) on Jan. 23 which was followed by suspension of public transport and a ban on all vehicular traffic. Second, on February 2 the government implemented the policy of centralized quarantine and treatment of all confirmed and suspected cases. I will model these interventions by stacking two logistic curves, one on top of another, as described above.

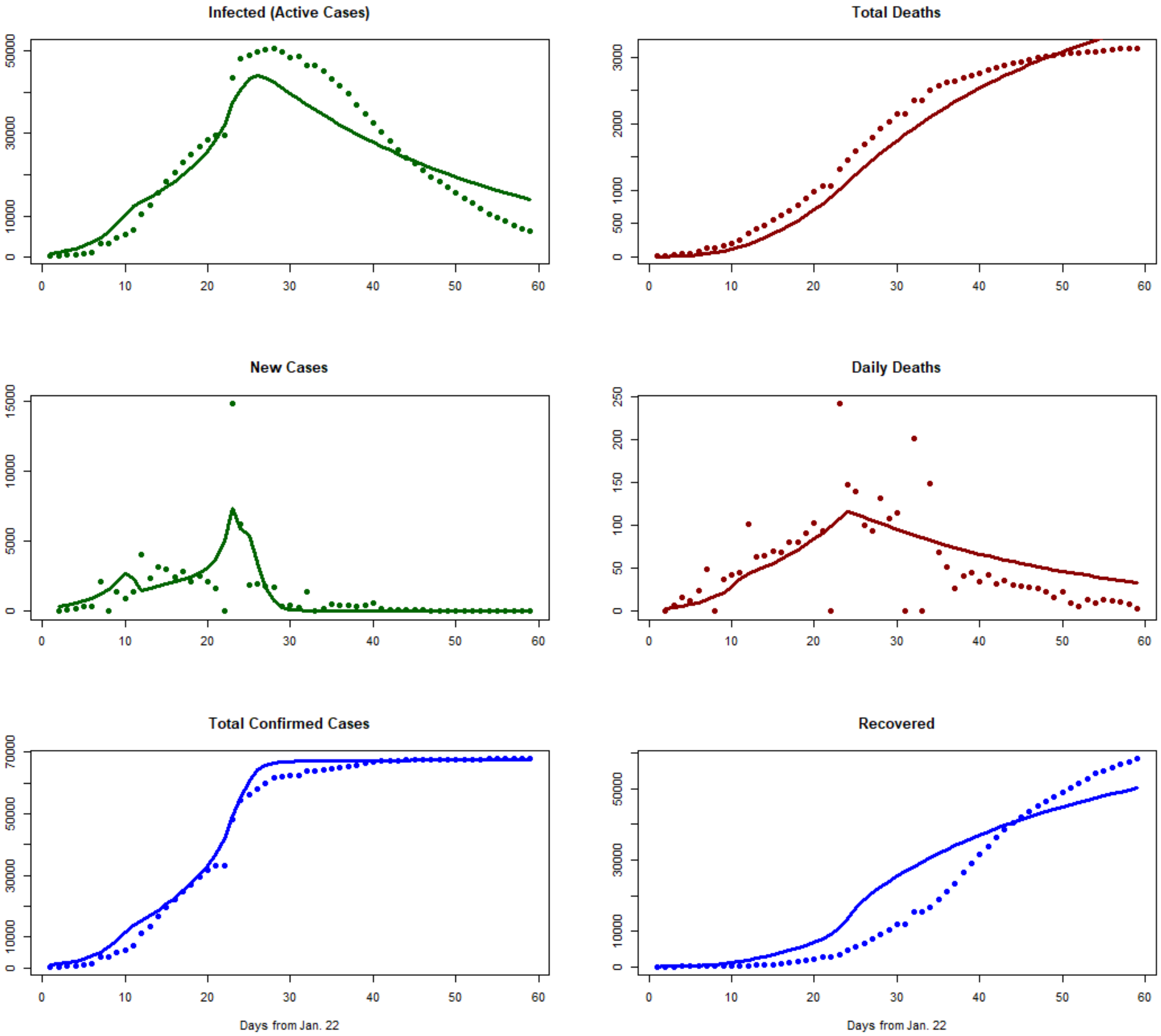

Figure 5. Comparison between model-predicted trajectories (curves) and data (points): Hubei, Jan. 22–Mar. 20.

```
Var  predR2
    I   0.883
    D   0.958
 delC   0.550
 delD   0.411
    C   0.988
    R   0.878
 Mean   0.778
```

Table 4. Coefficients of prediction for the six Hubei data series.

Table 5 lists the estimated parameters.

Parameter	Value	Explanation
N	60000000	total population = 60 mln
I0	2287	the number of infected at t = 0
beta1	0.22	transmission rate at t = 0
beta2	0.12	transmission rate after the first intervention
theta	7	steepness parameter of the logistic curve
b_date1	10	day at which the logistic curve for b(t) makes the first transition
b_date2	22.7	day at which the logistic curve for b(t) makes the second transition
gamma	0.0337	recovery rate of infected
delta	0.0015	death rate of infected individuals
q1	0.42	initial detection rate of infected
q2	0.65	detection rate of infected at the end
theta_q	1	steepness parameter for the q logistic curve
q_date	24	day at which the logistic curve for q(t) inflects

Table 5. Estimated parameters for Hubei.

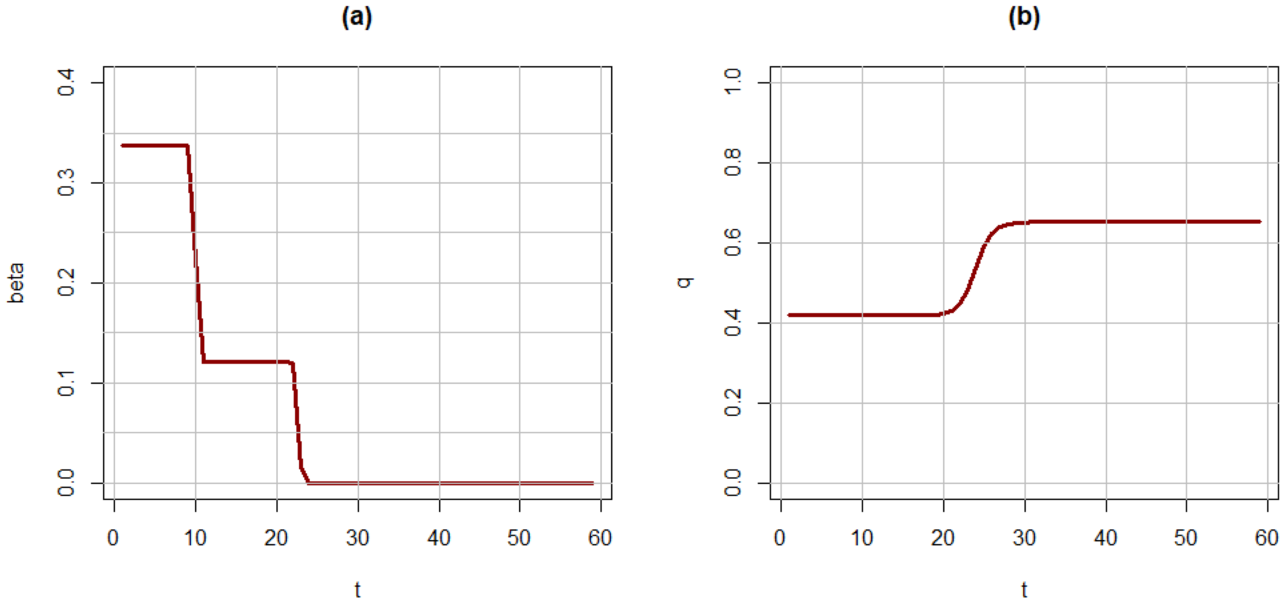

Figure 6. Estimated time-dependence in (a) the infection rate, b(t), and (b) the detection rate, q(t)

As expected, the degree of fit is poorer than for South Korea. Nevertheless, the model captures well the timing of all critical transition points, such as the peaks in Active Cases and Daily Deaths), and yields a decent approximation to the levels at which these transitions occur. Furthermore, where the model fails, it does so in interesting and instructive ways. In particular, underprediction of the Active Cases level at which the curve peaks is probably due to a failure to model some change in how Active Cases are reported. As Figure 6b shows, the detection rate is estimated to increase on Day 24 in a fairly mild way, whereas the observed jump in the data is much larger.

A very interesting comparison is for Daily Deaths (and Total Deaths). The model underestimates actual mortality during the period 25–35 (late January) and overestimates mortality for the period following Day 35 (most of February). The most likely explanation is that the medical facilities in Hubei were severely overwhelmed, which resulted in elevated levels of mortality, as many patients couldn't be properly treated. After Feb. 2 the Chinese authorities were able to treat all severely affected patients, and the mortality rate decreased. This decline was very substantial, eyeballing the curve, death rates were reduced by 50% or more. Modeling temporal dependence in the death rate, delta, with a logistic curve should yield a more precise quantitative estimate of this effect.

Italy

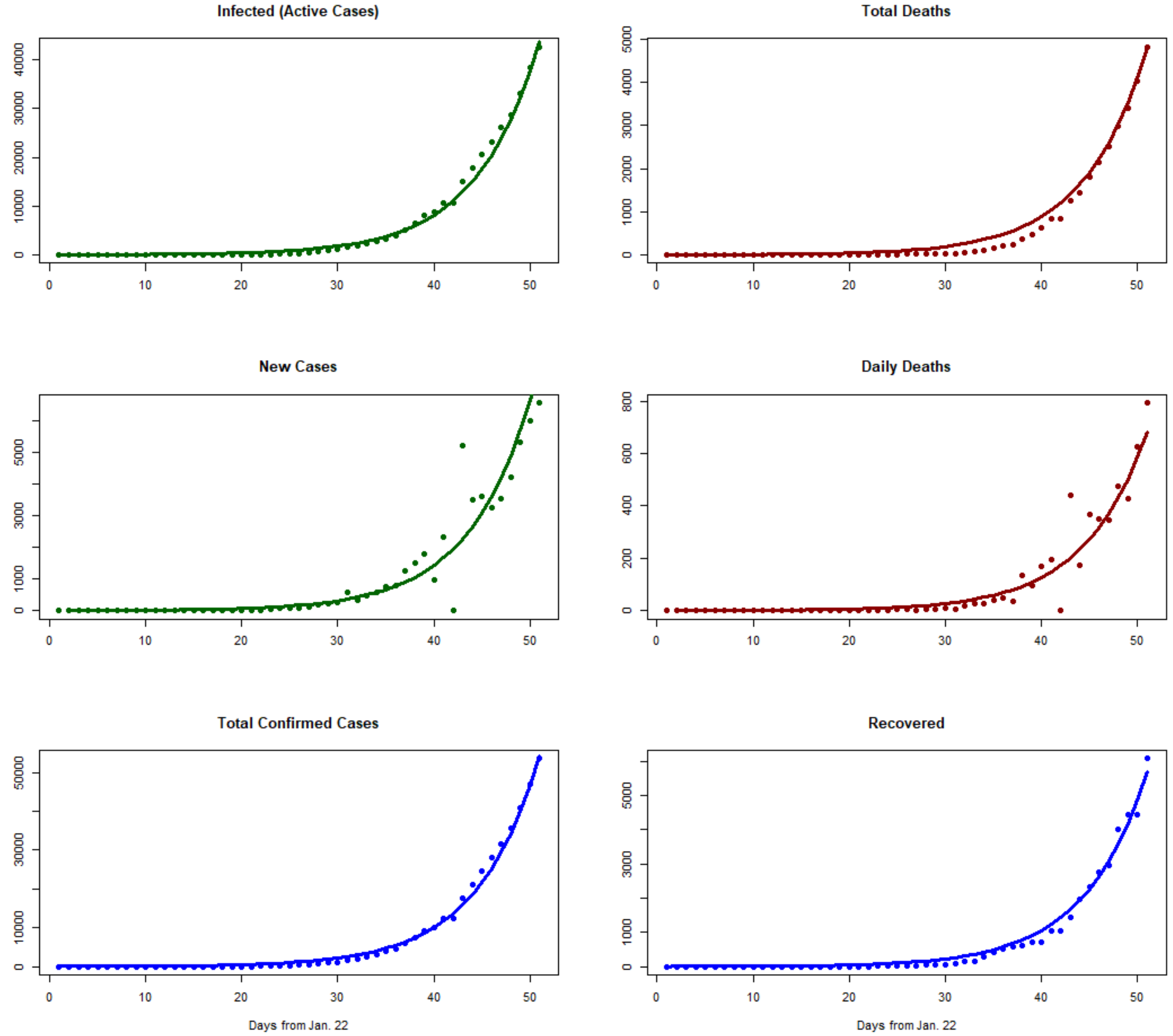

Figure 7. Comparison between model-predicted trajectories (curves) and data (points): Italy, Feb. 1–Mar. 20.

Results for Italy can be summarized in three sentences.

First, only the basic SIRD model is needed. There is no sign of any change in its coefficients.

Second, the model fits exceedingly, even frighteningly, well. The mean prediction R2 is 0.96.

Third, it means that the situation in Italy is extremely dire. There are no signs whatsoever that the measures implemented so far had slowed down the epidemic. However, it should be noted that there could be a time lag of 5–7 days before a successful intervention's effect would show up in these data.

Conclusion

I have tested the approach with three case studies, varying in the complexity of selecting the appropriate combination of mechanisms. The simplest is Italy, where (so far) the basic SIRD model with time-invariant parameters provides a very good fit. The intermediate case is South Korea, in which there was a single massive and highly successful intervention. The most complex case is Hubei (China), in which the authorities repeatedly changed the methodology of counting Active Cases and implemented a series of interventions before managing to press the transmission coefficient close to zero.

Clearly more work is needed. For complex cases like China we need a better understanding of how various interventions could result in changes of beta, delta, and q.

We need a better statistical model, ideally a Bayesian approach, that would yield not only point estimates of parameters but also quantify uncertainty associated with these estimates.

This approach can be applied to all countries for which data are available. However, analyzing each country requires substantial fine-tuning of the model, based on what is known about the timing and potential impact of various interventions. Such information should be built in as hypotheses about possible changes of parameters rather than estimated from data (following the maxim of don't estimate what is known).

One important limitation of this methodology is that generally it shouldn't be expected to yield accurate estimates of the overall level of the detection rate (q), only changes in q. This is a limitation imposed on us by the nature of data, rather than methodology. Random sampling of population without respect to symptoms should yield direct estimates of the proportion of asymptomatic cases, which can then be fed into the (process+observation) model.