Insights from scale transition into compartmental epidemiological models

Chris H. Wilson

4/4/2020

Table of Contents

# Background

The world is currently in the grip of a pandemic caused by the novel coronavirus SARS-CoV2. At last count, the official tally worldwide is north of 1 million confirmed cases and ~53K deaths (<https://experience.arcgis.com/experience/96dd742462124fa0b38ddedb9b25e429>). A great deal of attention has focused on a variety of efforts to forecast the progression of the disease, often relying on epidemiological theory (e.g.), statistical curve fitting (e.g.), or in the best case scenario, principled reconciliation of data and theory (e.g.). The fundamental theory underlying our understanding of infectious disease dynamics is encapsulated in the classical SIR equations, which I reviewed here for easy reference (<https://rpubs.com/chwilson101/587211>). Although the numerous assumptions of the SIR class of models are widely acknowledged, certain implications appear to be less widely appreciated. In this note, I draw attention to the fact that mean-field assumptions will fail to hold, with disastrous consequences for estimating the parameters of SIR-type equations from aggregate data. Another implication is that pandemic progression is radically *non-ergodic*. Practically speaking, this implies that local authorities should not rely on aggregate or spatially-averaged statistics to guide decision-making.

# Setup

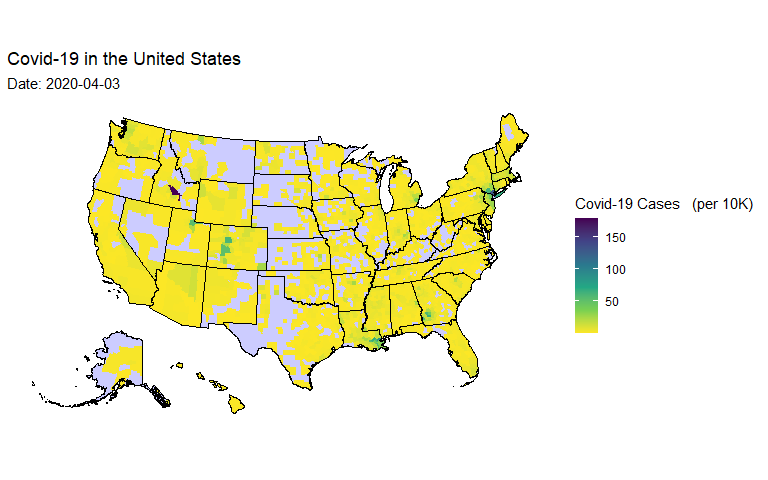
To recap, the SIR equations, as I derived in (<https://rpubs.com/chwilson101/587211>), are:

$$\tag {1} \frac {dS}{dt} = - R\_0 \nu I \frac{S}{N} $$

$$\tag {2} \frac {dI}{dt} = R\_0 \nu I \frac{S}{N} - \nu I $$

$$\tag {3} \frac {dR}{dt} = \nu I $$

These equations assume a well-mixed population (the mass-action princple). It is reasonable to question this assumption at large scales. For instance, a casual glance at a map of the US based on the newly released county level data from the NYT shows striking spatial heterogeneity (Fig.1). This heterogeneity is apparent at both national and state levels.



*Figure 1: Map of confirmed Covid-19 cases in the US broken down by county. Data from the New York Time:* [*https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv*](https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv)

However, it might be countered that we can still usefully apply the theory if only we are willing to interpret the parameters in a “mean field” fashion. Practically speaking, this is exactly what is assumed any time such a model is fit to aggregated data. We hope that the temporal trend (time series) and spatial variations hang together, amounting to an assumption of some form of ergodicity.

# Applying the scale transition

The scale transition theory from ecology (notably developed by Chesson XXXX,and XXXX) can provide theoretical insight into the breakdown of this assumption and its consequences. We apply it as follows: let us consider each disaggregated unit (in this case, a ‘county’) as our local ‘patch model’. We index the location in space with ‘x’, and rewrite our equations for any given patch as:

$$\tag {4} \frac {dS\_x}{dt} = - R\_{0\_x} \nu\_x I\_x \frac{S\_x}{N\_x} $$

$$\tag {5} \frac {dI\_x}{dt} = R\_{0\_x} \nu\_x I\_x \frac{S\_x}{N\_x} - \nu\_x I\_x $$

$$\tag {6} \frac {dR\_x}{dt} = \nu\_x I\_x $$

For simplicity, I will simply linearize equation 5 under assumption that our attention is on early phase ():

$$\tag {7} \frac {dI\_x}{dt} = \nu\_x (R\_{0\_x} - 1) I\_x $$

Following Chesson XXXX, if we assume there are a large number of spatial locations, we can estimate the area-wide dynamics by taking the expectation over space:

$$\tag {8} \mathop{\mathbb{E\_x}}[\frac {dI\_x}{dt}] = \mathop{\mathbb{E\_x}}[\nu\_x (R\_{0\_x} - 1) I\_x]$$

That is to say, for a sufficiently short interval of time, the RHS of equation 8 will represent the change in the number of infecteds in the population. In most applications of the scale transition, we are particularly concerned by the combination of non-linear equations and heterogeneity. In this case, our dynamics are first order, and at first glance this would suggest that our model should scale up just fine, even with lots of heterogeneity.

Wrung.

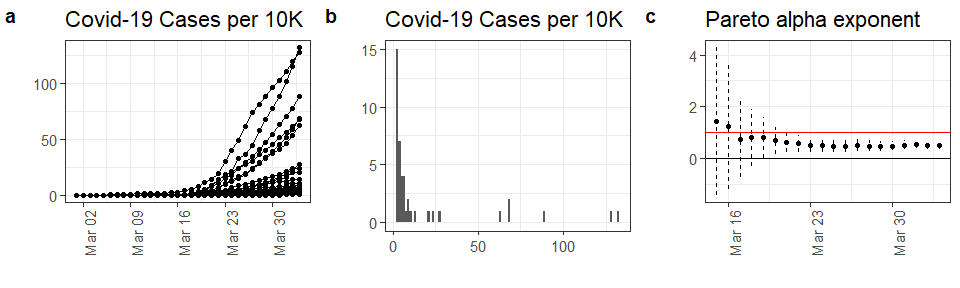
I will illustrate the breakdown of this hope with an even simpler model. Let us suppose for now that there is not meaningful variation in , and so our county units vary principally in (and I collapse -1 to for simplicity) and the current number of infecteds . Given their representation as random variables in the patch model, the aggregate model is simply the product of the random variables:

$$\tag {9} \overline{\frac {dI\_x}{dt}} = \overline{\nu \psi\_x I\_x} = \nu \overline{\psi}\overline{I} + cov(\psi,I)$$

In short, equation 9 simply says we have to correct the mean-field model () with the covariance in and . So far so good. If we want to rescue good parameter estimates for from aggregate time series data we just need to know the relevant covariance.

# Insights from the scale transition

But wait! Look again at Fig. 1. It turns out the distribution over is **extremely** fat-tailed. Even after normalization by population density within a county (itself a fat-tailed quantity), it remains fat-tailed and fairly well described by a Pareto distribution with low exponent. Rather than national-level, I show this in Fig. 2 specifically for the case of New York (Fig 2a has all county trajectories, 2b has a histogram).

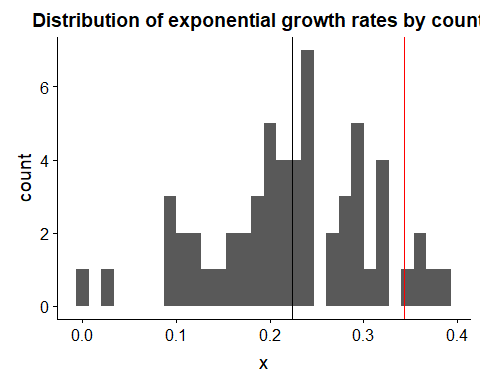


*Figure 2: a) time series of confirmed Covid-19 cases from every county in New York State, b) historgram of the distribution of cases on 2020-04-03 c) Result of fitting distribution of Covid-19 cases on each date to a Pareto distribution via simple maximum likelihood. Graph shows maximum likelihood estimate +/- 2 SEs. Red line shows alpha = 1, below which the first moment goes to infinity (is undefined)*

On 2020-04-03, the Pareto exponent among New York counties, estimated by maximum likelihood is 0.51, and has been declining since 2020-03-15.

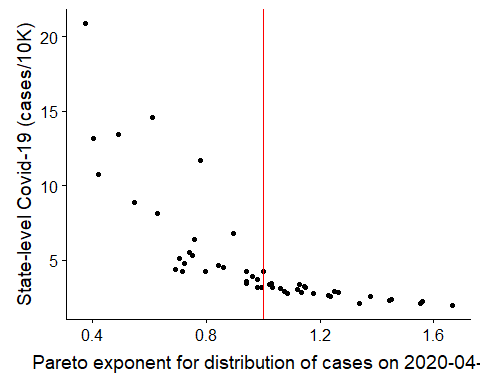
Fat-tailed distributions bring some important caveats to the table. First, for the standard Pareto distribution where the goes below 2, the variance becomes infinite! Practically speaking, this means it is undefined. Thus, any covariance term (such as the correction in 9) is indeterminate. When the parameter goes below 1, even the mean (first moment) no longer exists, thus rendering the rest of the terms in equation 9 problematic.

Unfortunately, most states are showing alpha < 2, and several alpha < 1. When we, for example, fit an exponential model iteratively to each county in New York and take the mean of the estimated growth parameters (reflecting spatial variations in ), it differs systematically from the estimate based on all counties summed together, just as our analysis suggests (Figure 3). In fact, we should expect a fairly high positive covariance/correlation between and , given the dynamics at play, and thus the aggregate estimate will be systematically higher than the mean of the disaggregated estimates.



*Figure 3: Distribution of exponential growth rates fitted to each county in New York separately. The black line shows the mean of these growth rates (the mean-field interpretation of SIR), whereas the red line shows the actually estimated growth rate from the state-level data fitted as an aggregate. This clearly demonstrates the breakdown of the mean-field assumption, exactly as predicted by the theoretical analysis here.*

What’s more, increasing fat-tailedness within a state is exponentially correlated with the state-level aggregated prevalence of Covid-19. This fine-grained spatial pattern is absolutely critical to accurately understanding and forecasting Covid-19 (Figure 4).



*Figure 4: As Pareto exponent declines, the aggregate state situation gets exponentially worse. The spatial dynamics of Covid are such that individual counties/cities are driving the entire story.*

# Conclusions

Formally, the equations (4,5,6) cannot necessarily be solved given the closure problem (minimally, we would need equations desribing how all the covariances evolve, and to demonstrate that the whole system obeyed conservation principles, etc.) However, our analysis of equation 9, in tandem with the data shown in figures 1 and 2, is sufficient to show that estimates of from aggregated data are systematically biased. Unfortunately, since the moments of the relevant distributions do not exist, we cannot rigorously account for this bias. Attempts to empirically estimate the covariance in equation 9 will require substantially more data than otherwise, and the empirical mean of will fail to represent the “true” mean in equation 9 when the distribution is sufficiently fat-tailed.

Our best chance at understanding and forecasting Covid-19 is to develop models at as granular a level as possible, and then rigorously and explicitly account for spatial interconnections and processes. This is a lot of work, and may be logistically challenging. Accordingly, for strategic planning purposes, SIR-type models should be used to educate the intuition about what is possible, but decisions should be made with a rather strong precautionary approach.