

# SIR Scaling

Chris H Wilson

4/2/2020

## Contents

Background	1
Setup	1
Applying the scale transition	3
Insights from the scale transition	4
Discussion and Conclusions	6
References	7

## 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, 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, we draw attention to the fact that mean-field assumptions will fail to hold, with potentially disastrous consequences for estimating the parameters of SIR-type (“compartment”) equations from aggregate data. Another implication is that pandemic progression is radically *non-ergodic*. Practically speaking, we arrive at three conclusions: 1. The crucial  $R_0$  parameter is not only time-varying but has significant spatial variations. These spatial variations are such that they lead to non-ergodic, power-law distributed case counts which frustrate mean-field models fit to aggregate data, **even if the underlying  $R_0$  distribution is not power-law distributed**. 2. Models should be as spatially explicit as possible, and fit to maximally disaggregated data. 3. Local authorities should not rely too much on aggregate or spatially-averaged statistics to guide decision-making, but rather adopt a strong precautionary approach.

## Setup

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

$$\frac{dS}{dt} = -R_0\nu I \frac{S}{N} \tag{1}$$

$$\frac{dI}{dt} = R_0\nu I \frac{S}{N} - \nu I \tag{2}$$

$$\frac{dR}{dt} = \nu I \quad (3)$$

These equations assume a well-mixed population (the mass-action principle). 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.

### Covid-19 in the United States

Date: 2020-04-04

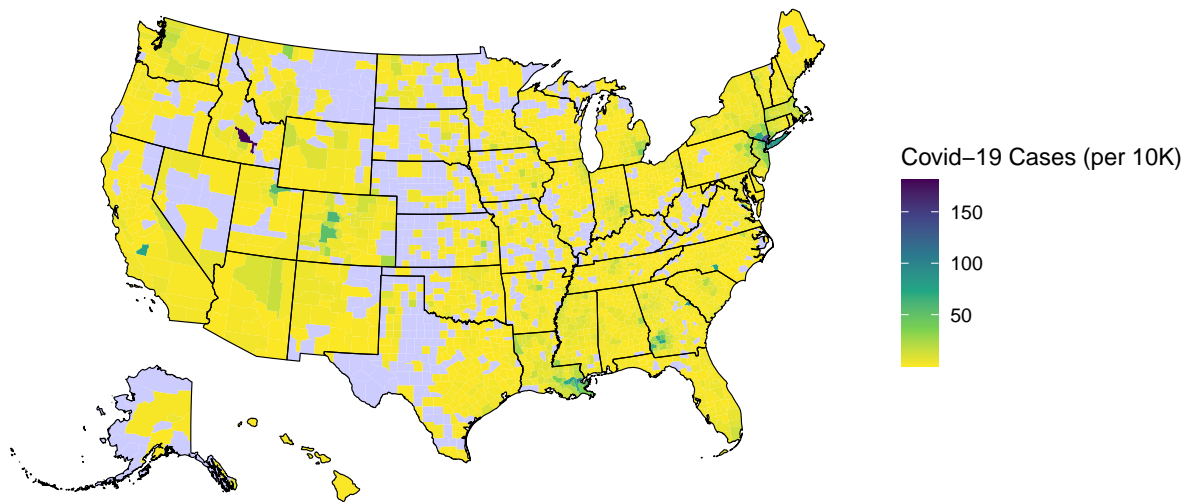


Figure 1: Map of confirmed Covid-19 cases in the US broken down by county. Data from the New York Time: <https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv>

## Covid-19 mortality in the United States

Date: 2020-04-04

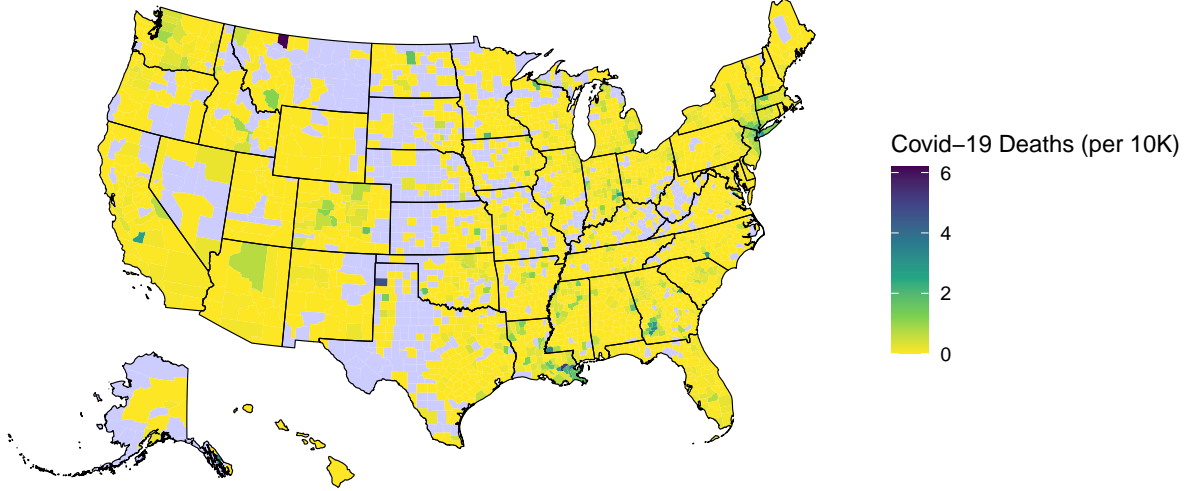


Figure 2: Map of confirmed Covid-19 deaths in the US broken down by county. Data from the New York Time: <https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv>

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:

$$\frac{dS_x}{dt} = -R_{0_x} \nu_x I_x \frac{S_x}{N_x} \quad (4)$$

$$\frac{dI_x}{dt} = R_{0_x} \nu_x I_x \frac{S_x}{N_x} - \nu_x I_x \quad (5)$$

$$\frac{dR_x}{dt} = \nu_x I_x \quad (6)$$

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

$$\frac{dI_x}{dt} = \nu_x (R_{0_x} - 1) I_x \quad (7)$$

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:

$$\mathbb{E}_{\cap}[\frac{dI_x}{dt}] = \mathbb{E}_{\cap}[\nu_x(R_{0x} - 1)I_x] \quad (8)$$

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  $\nu$ , and so our county units vary principally in  $R_0$  (and I collapse  $R_{0-1}$  to  $\psi$  for simplicity) and the current number of infecteds  $I$ . Given their representation as random variables in the patch model, the aggregate model is simply the product of the random variables:

$$\overline{\frac{dI_x}{dt}} = \overline{\nu\psi_x I_x} = \nu\overline{\psi I} + cov(\psi, I) \quad (9)$$

In short, equation 9 simply says we have to correct the mean-field model ( $\nu\overline{\psi I}$ ) with the covariance in  $R_0$  and  $I$ . Qualitatively, this differs from simple exponential growth, and can be expressed more generically as:

$$\frac{dx}{dt} = \lambda x + f_t(\lambda, x) \quad (10)$$

which is manifestly a non-autonomous ODE. The problem here is that no simple functional form exists for the non-autonomous term, which is effectively an emergent property of how the pandemic develops over space and time.

So far so good. If we want to rescue good parameter estimates for  $R_0$  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  $I$  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  $\alpha$  exponent. Rather than national-level, I show this in Fig. 3 specifically for the case of New York (Fig 3a has all county trajectories, 3b has a histogram).

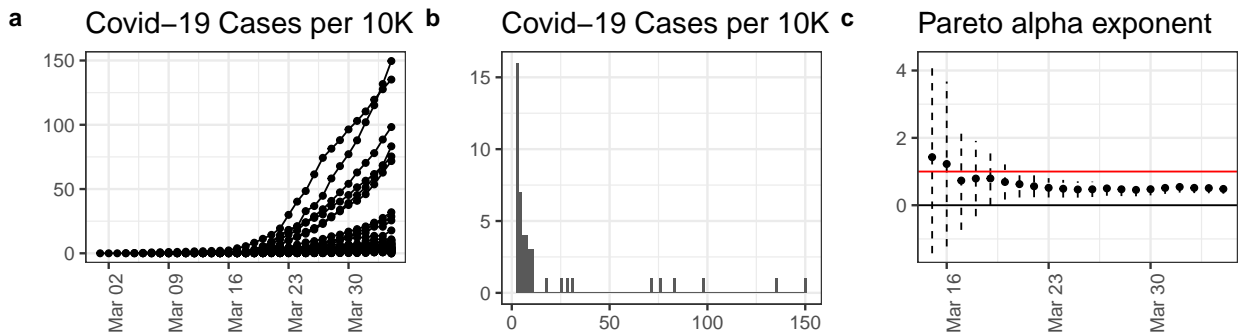


Figure 3: a) time series of confirmed Covid-19 cases from every county in New York State, b) histogram of the distribution of cases on 2020-04-04 c) Result of fitting distribution of Covid-19 cases on each date to a

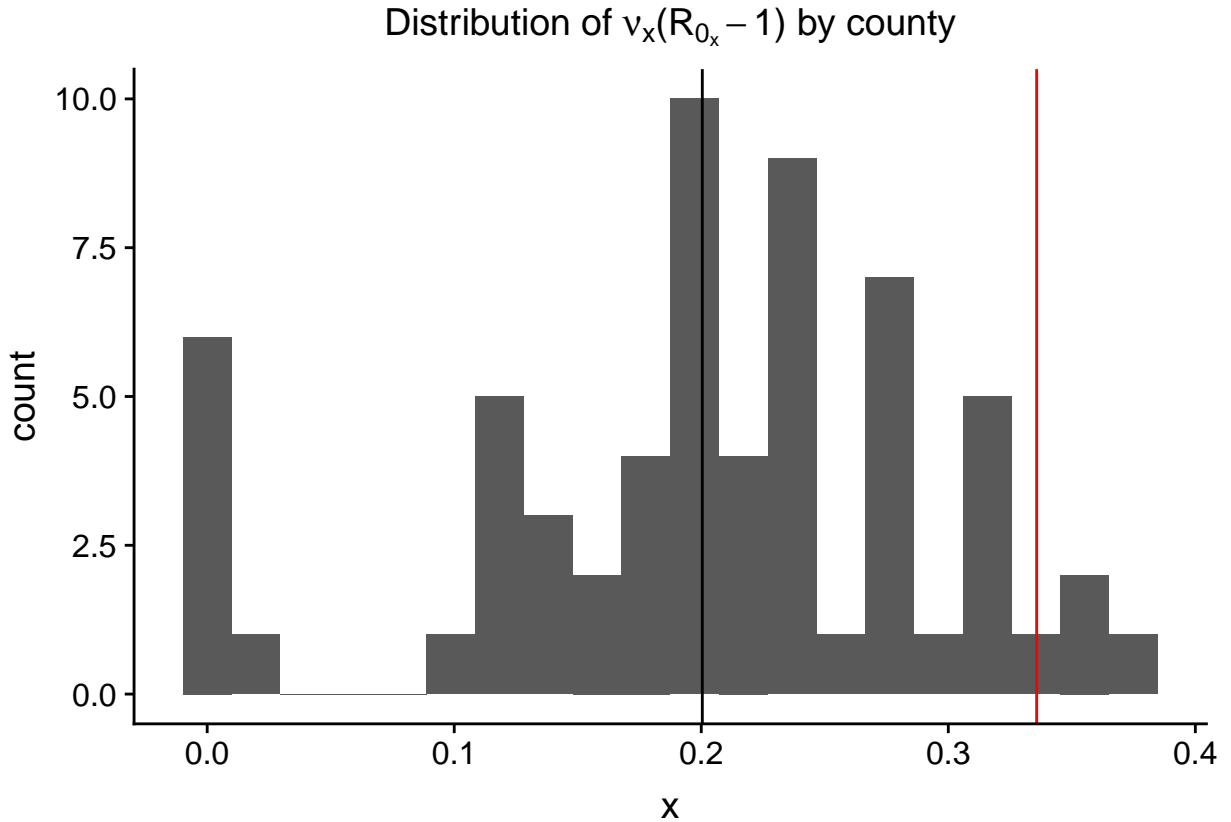
*Pareto distribution via simple maximum likelihood. Graph shows maximum likelihood estimate  $\pm 2$  SEs. Red line shows  $\alpha = 1$ , below which the first moment goes to infinity (is undefined)*

On 2020-04-04, the Pareto exponent among New York counties, estimated by maximum likelihood is 0.48, 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  $\alpha$  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  $\alpha$  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  $R_0$ ), it differs systematically from the estimate based on all counties summed together, just as our analysis suggests (Figure 3).

Note that the exponential growth parameter fit to each county separately is identical to estimating  $\nu_x(R_{0_x} - 1)$  from equation 7, while fitting an exponential model to the state-aggregate data the exponential growth parameter is equivalent to fitting the mean-field portion of equation 9  $\nu\psi\bar{I}$ , while ignoring the effectively non-autonomous term  $cov_t(\psi, I)$ . In fact, we should expect a fairly high positive covariance/correlation between  $R_0$  and  $I$ , given the dynamics at play, and thus the aggregate estimate will be systematically higher than the mean of the disaggregated estimates. The non-autonomous term leads to higher growth increments than the mean-field.



*Figure 4: 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).

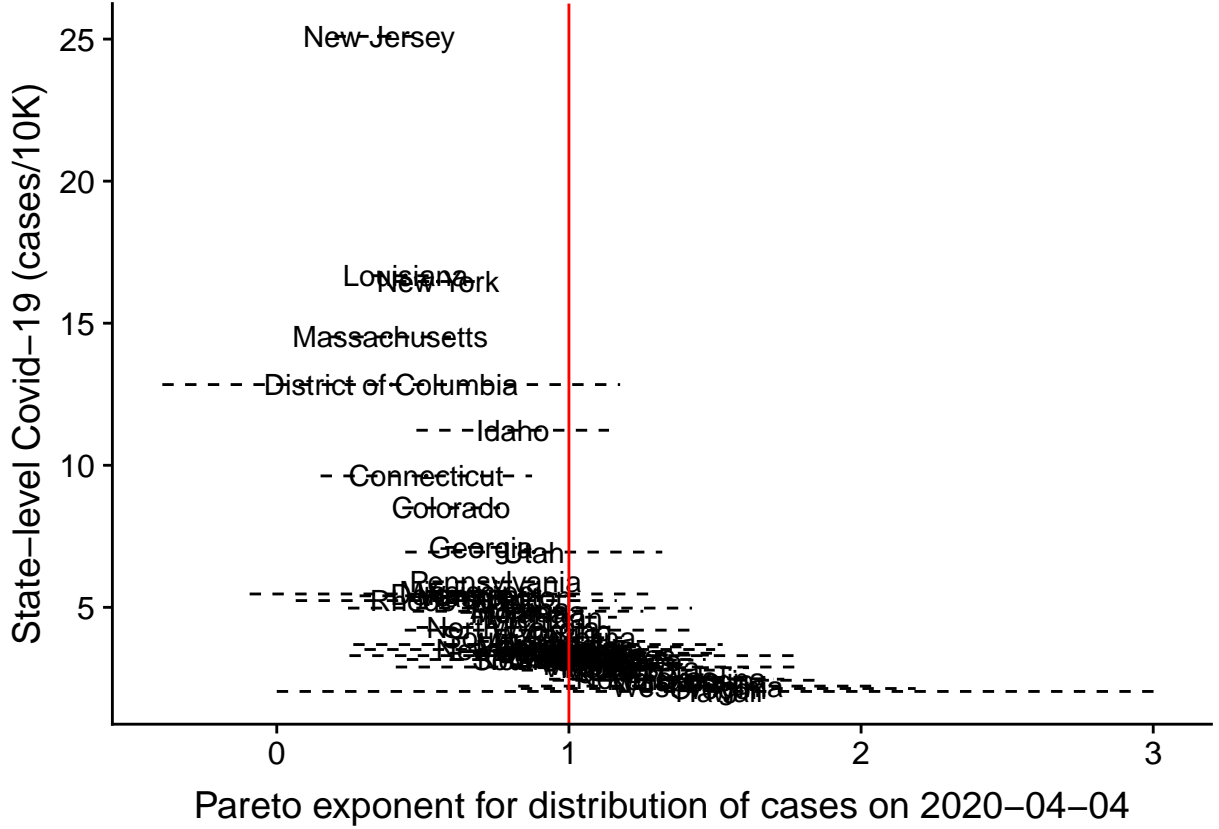


Figure 5: As Pareto exponent declines, the aggregate state situation gets exponentially worse. The spatial dynamics of Covid are such that individual counties/cities are driving the entire story.

## Discussion and Conclusions

Formally, the equations (4,5,6) cannot necessarily be solved via spatial expectations given the lack of functional form for the non-autonomous covariance terms and the closure problem they induce. Minimally, we would need equations describing how all the covariances evolve while still fulfilling conservation principles. However, our analysis of equation 9 (and 10), in tandem with the distribution of county-level data, is sufficient to show that estimates of  $R_0$  from aggregated data are certainly *systematically biased*. Unfortunately, since the moments of the relevant distributions often do not exist, attempting to correct for this bias with empirical estimates is problematic, since the empirical estimates themselves will be deeply biased (Cirillo and Taleb). Where  $\alpha < 2$  (effectively everywhere, Fig. 5), the second moment for  $I$  is infinite and thus the non-autonomous covariance term is mathematically undefined. Where  $\alpha < 1$  (several states, Fig. 5), the first moment is infinite and thus the whole equation is thrown into doubt.

Our theoretical analysis not only pinpoints the difficulty in deriving accurate mean-field parameter estimates/interpretations for  $R_0$ , it is also directionally correct about the implication of the resulting bias for aggregate-scale pandemic behavior. Equation 9 clearly suggests that aggregate growth rate should be higher than mean-field (since first principles suggest a positive covariance between  $R_0$  and  $I$ ), a result demonstrated in Figure 4. Moreover, we also see that the severity of the state-wide aggregate situation is exponentially related to the underlying fat-tailedness of the spatial distribution of cases across counties in Figure 5.

Altogether, we argue our best chance at understanding and forecasting pandemic progression is to develop models at as granular a spatial level as possible, and then rigorously and explicitly account for spatial interconnections and processes. One recommended approach is the use of movement data to explain the high degree of spatial autocorrelation evident in Figures 1 and 2, where we observe a seemingly stochastic distribution of hotspots at the national level coupled with local (county-county) diffusion.

The work required to develop and fit granular models is large and should not be underestimated. More highly parameterized models also induce their own problems with uncertainty and specificity, and ensuring models are sufficiently robust to a wide range of error is critically important in these life-and-death applications. Accordingly, for strategic planning purposes, we recommend that 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. Local officials who question the integrity or depth of their county-level data, should consider smoothing their estimates based on spatially-weighted local neighborhoods. Coarse-grained state-level and national-level statistics should not be used to guide local-level decision-making, except with an extremely coarse grain of salt.

## References