Squashing the sombrero

29 April 2020

Understanding the uncertainty in daily coronavirus deaths using the negative binomial distribution

COVID-19 is an epic human tragedy. While we cheer on our healthcare workers and observe social distancing, it is natural ask questions about how long the epidemic will take to get under control here the UK. To use Boris Johnson's analogy for the curve of peak deaths, *is the sombrero subsiding*?

This article aims to show how a basic model using Python and basic stats can help answer questions such as whether survival rates are trending up and how long it might take to unwind the lockdown. Our ultimate aim is to derive confidence intervals for the range of future daily deaths.

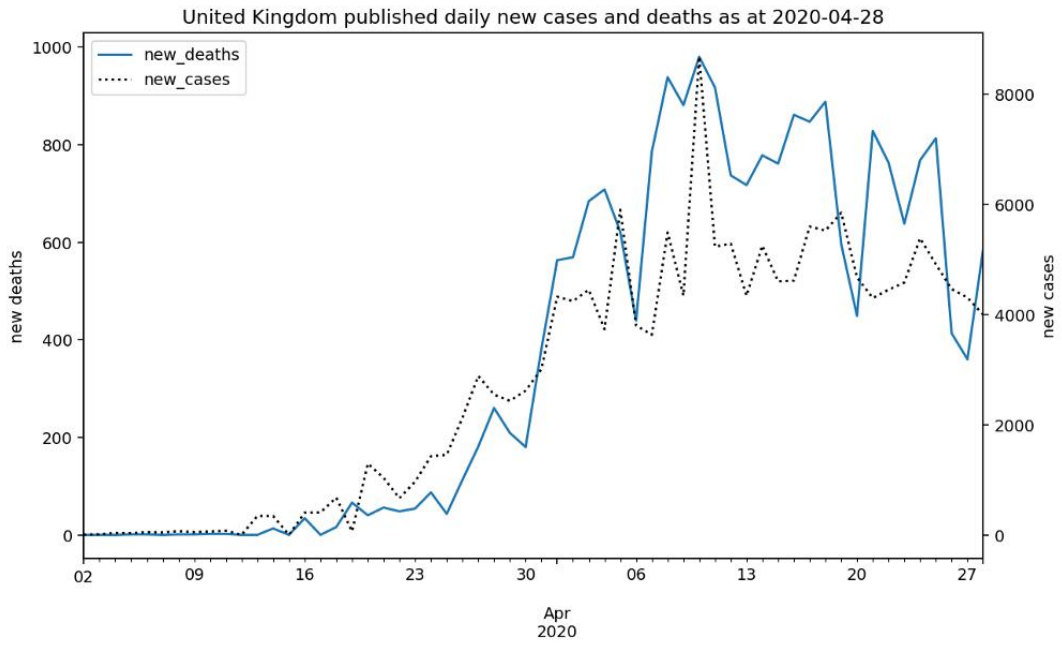
We use the Johns Hopkins database of new cases and deaths in each country and a negative binomial model to link deaths with new cases. This simple model has parameters that are easy to interpret. It also helps us to appreciate how hard it is for policymakers to see the trends through the noise.

Epidemiologists and other scientists in the field use far more advanced models, but they have to cope with the same underlying uncertainty in the growth rate of new infections,

survival rates and the lag between a new case and subsequent death for those unfortunate enough not to survive.

Highly volatile daily statistics…

Published daily deaths and new cases show a huge degree of variation, but trends are hiding in there somewhere. Here are the daily new cases and deaths to date for the UK:



The negative binomial distribution

We would like to estimate the number of deaths at each date $\boldsymbol{t}$ ($\boldsymbol{nd_t}$) assuming a survival rate ($\boldsymbol{s}$) and the number of new cases ($\boldsymbol{nc_i}$) reported at each date prior date $\boldsymbol{t}$. In statistical terms, we need:

$$\mathrm{E}\left[nd_t\right] = \sum_{i=1}^{t} nc_i\,(1-s)\,\boldsymbol{Pr}\left[\text{ dies at } t \;\middle|\; nc_i\right]$$

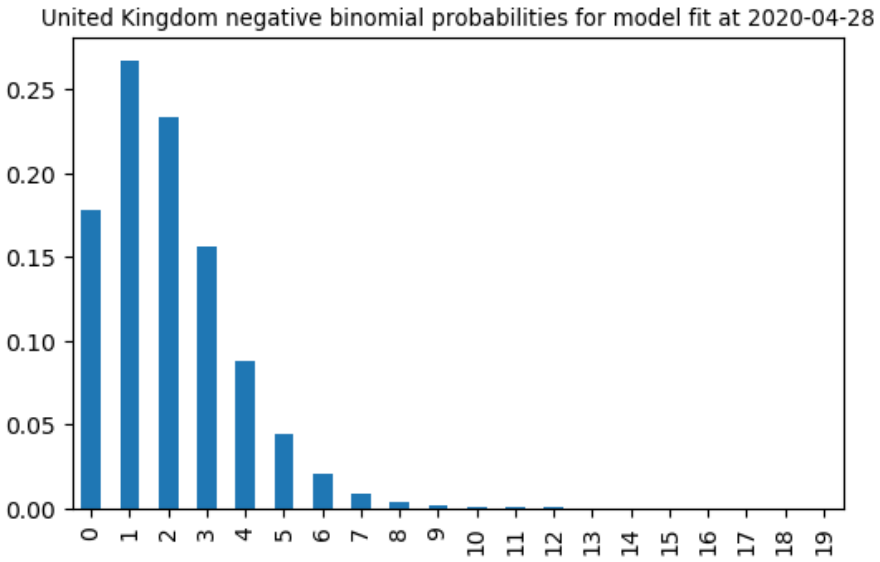
The negative binomial distribution is useful to describe these probabilities.

We assume that the lag between a positive test result (*i.e.* a new case) and death due to COVID-19 follows a negative binomial

distribution with parameters ***n*** and ***p***. This can be interpreted as the probability there will ***k*** failures until the ***n***-th success for ***k+n*** independent and identically distributed trials, each with probability of success ***p***. In the above expression ***k*** equals the lag, ***t-i***, so that:

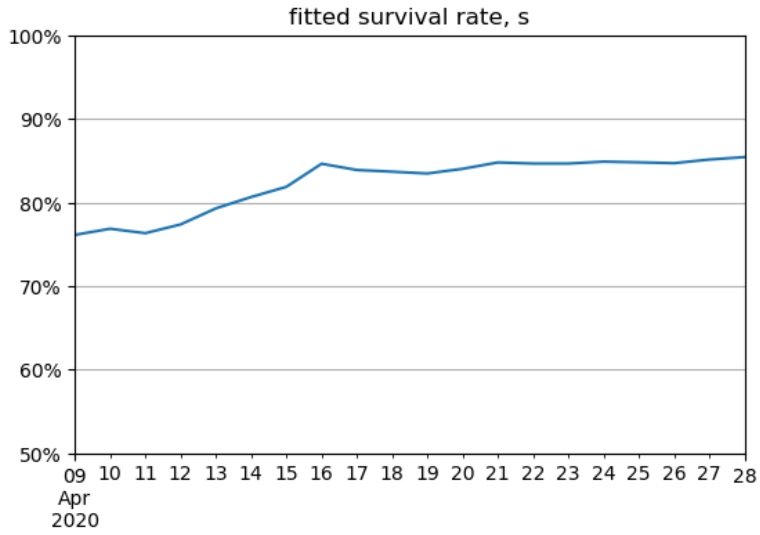
$$Pr\left[\text{ dies at } t \;\middle|\; nc_i\right] = \binom{t-i+n-1}{n-1} p^n (1-p)^{t-i}$$

Using the covid19.py code here, we can fit the survival rate, ***s***, and negative binomial model parameters, ***n*** and ***p***, at each date after the UK reached 100 known cases. The latest fit implies an 85% chance a new case will survive and for those unfortunate enough not to survive we expect a mean time to death of 2 days, but a fairly long tail:



While this mean time to death seems low, it is higher in other countries with more extensive testing. Furthermore UK published deaths are not mapped to date of death and new cases are not mapped to the specimen date. Finally, until everyone is tested, the deaths are also significantly understated as depicted in this recent article in The Economist.

As imperfect as the data may be, the fitted UK survival rate has been thankfully trending up in the past 20 days:

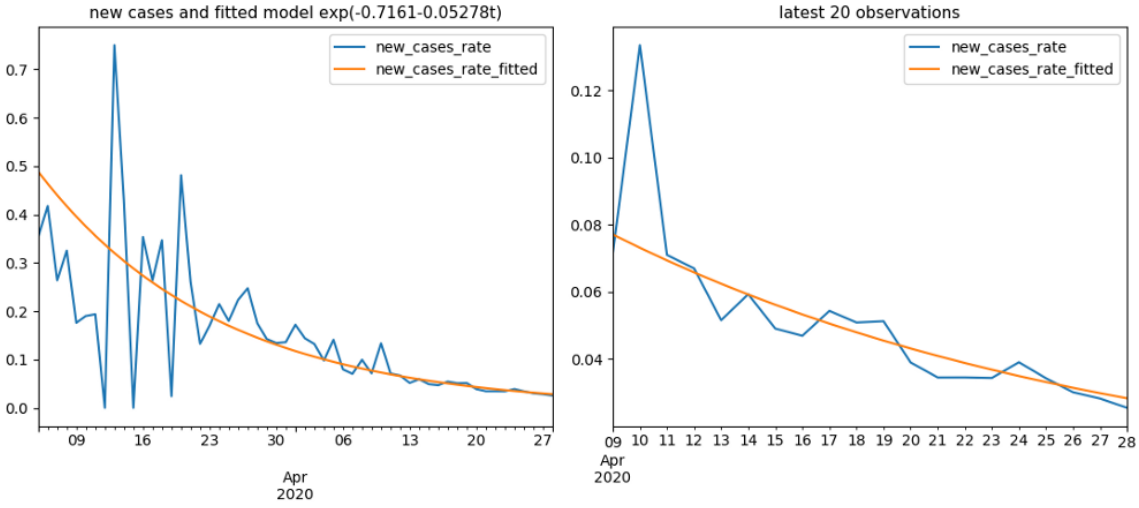


The fitted model can project daily deaths from current cases, but what about future new cases and consequent deaths?

We first need to check if we can forecast new cases with any confidence.

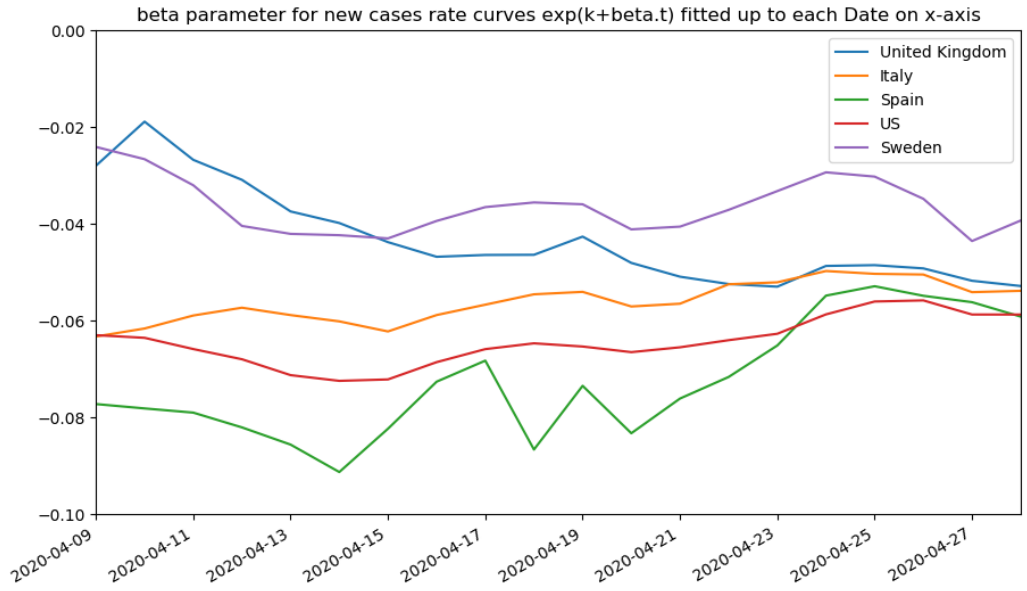
Effect of the lockdown

We hope to see the downward trend in new cases, as measures like the lockdown take effect. An exponential curve of the form ***exp(k+βt)*** gives a reasonable fit for many the countries in the dataset used. Here it is for the UK:



Here *t* is defined as the number of days since the first date where there are at least 100 new cases a day. A negative parameter *β*

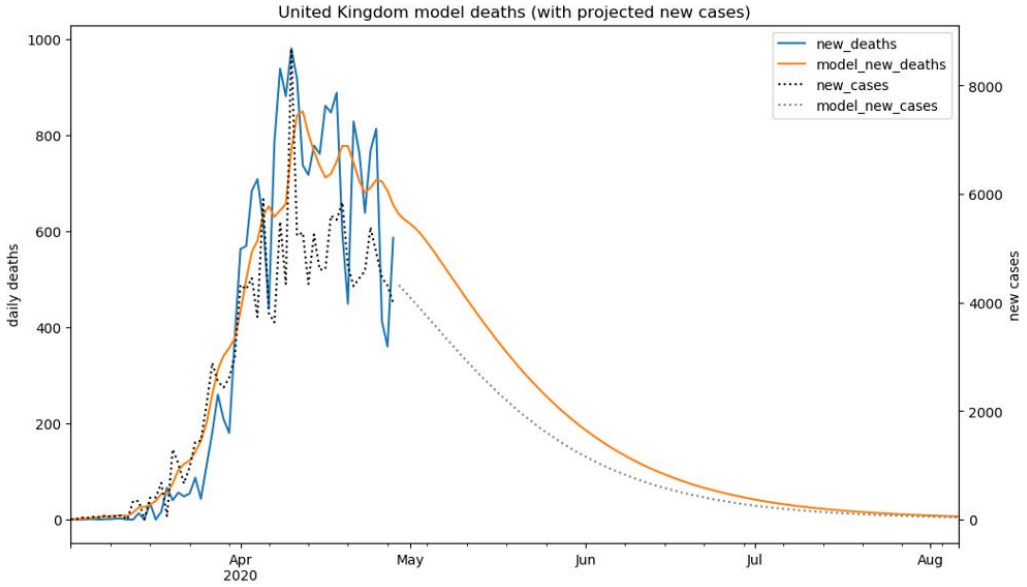
implies the *rate of growth of new cases is declining with time*. This is the case for many countries:



These fitted rates suggest new cases are declining in relative and more importantly in absolute terms.

A slowly subsiding sombrero

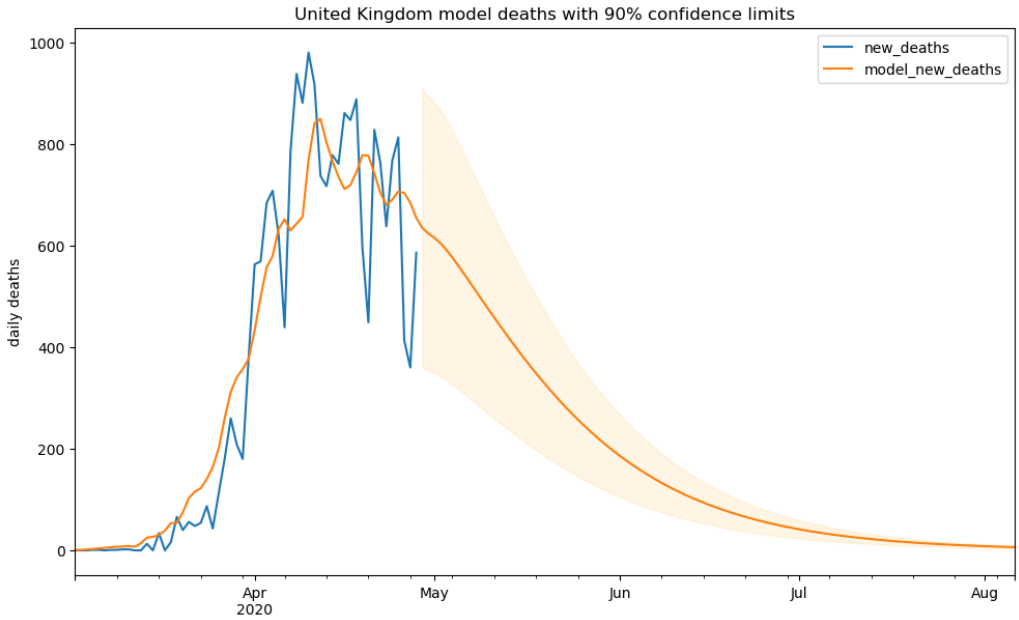
Here is the projection of UK daily new cases and deaths based on the latest fits for the rate of new cases growth, survival rate for new cases and negative binomial distribution for the lag between the date of a new case and death:



UK cumulative deaths by 2020-08-06 of 38736, 55 % of 21678 deaths to date

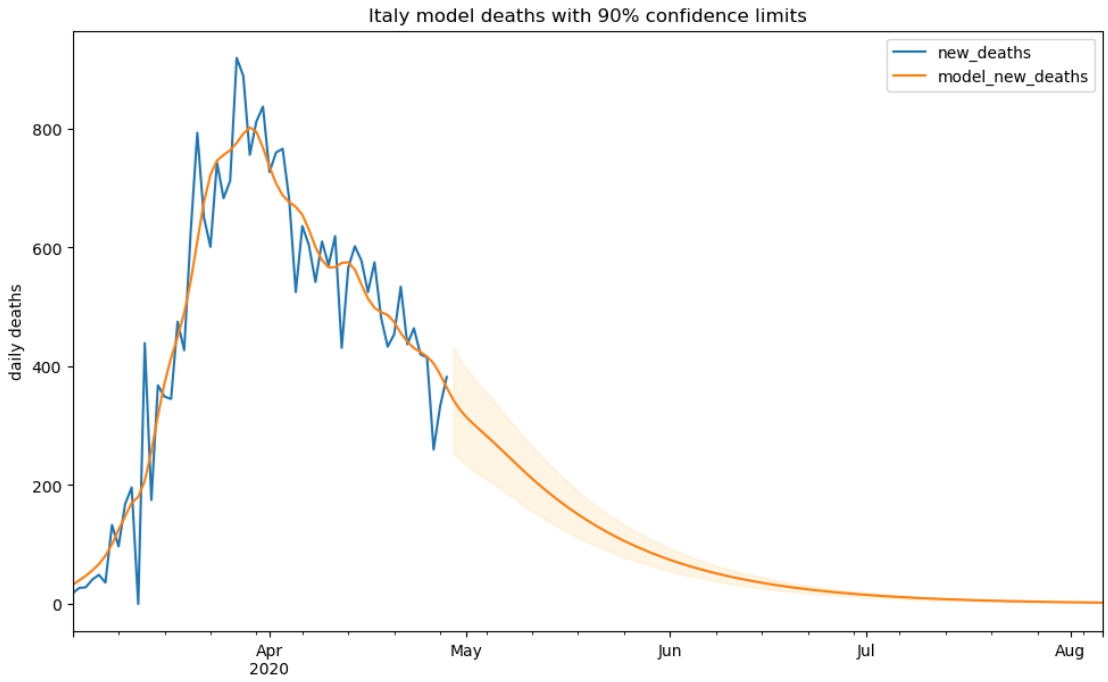
The peak appears to be behind us, though we are likely to be only halfway through the death toll of this terrible pandemic—if current conditions are maintained. This projection shows the decline in daily deaths will be far more gradual than the increase before the peak.

Next we add a 90% confidence interval for daily deaths. These bounds are based on the model errors since model deaths exceeded 100 days:

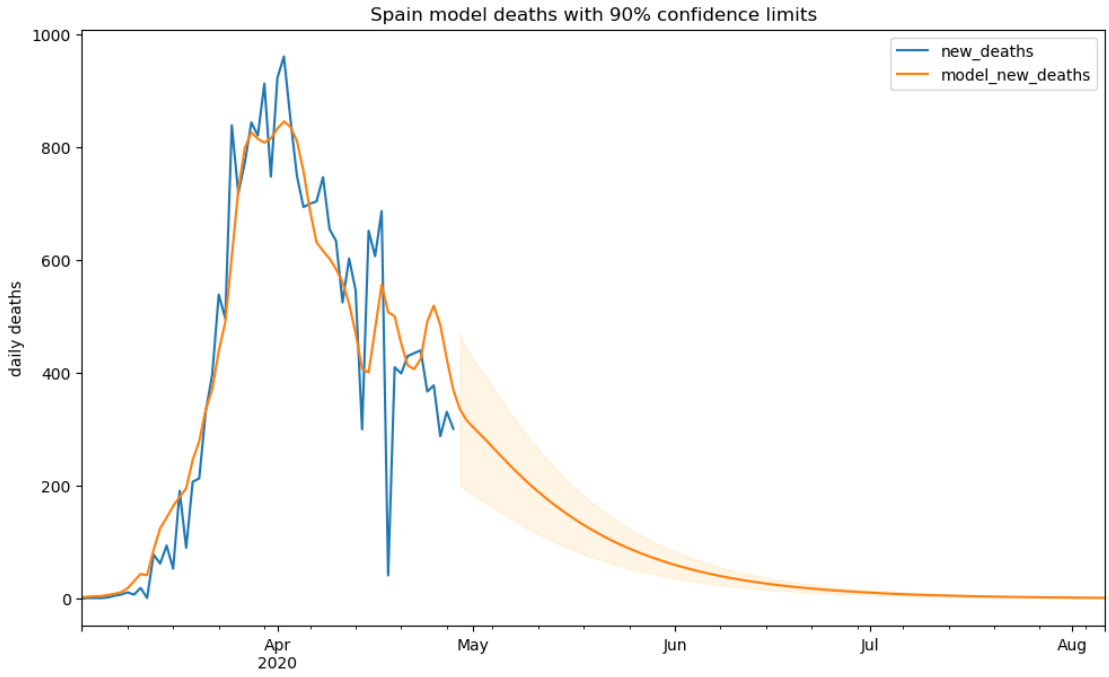


This shows the dilemma occupying the UK government. As the lockdown unwinds and conditions change, it is hard to be reasonably confident that intensive care resources will not be overwhelmed.

Still, there are grounds for hope. Here are the 100-day projections for Italy and Spain:

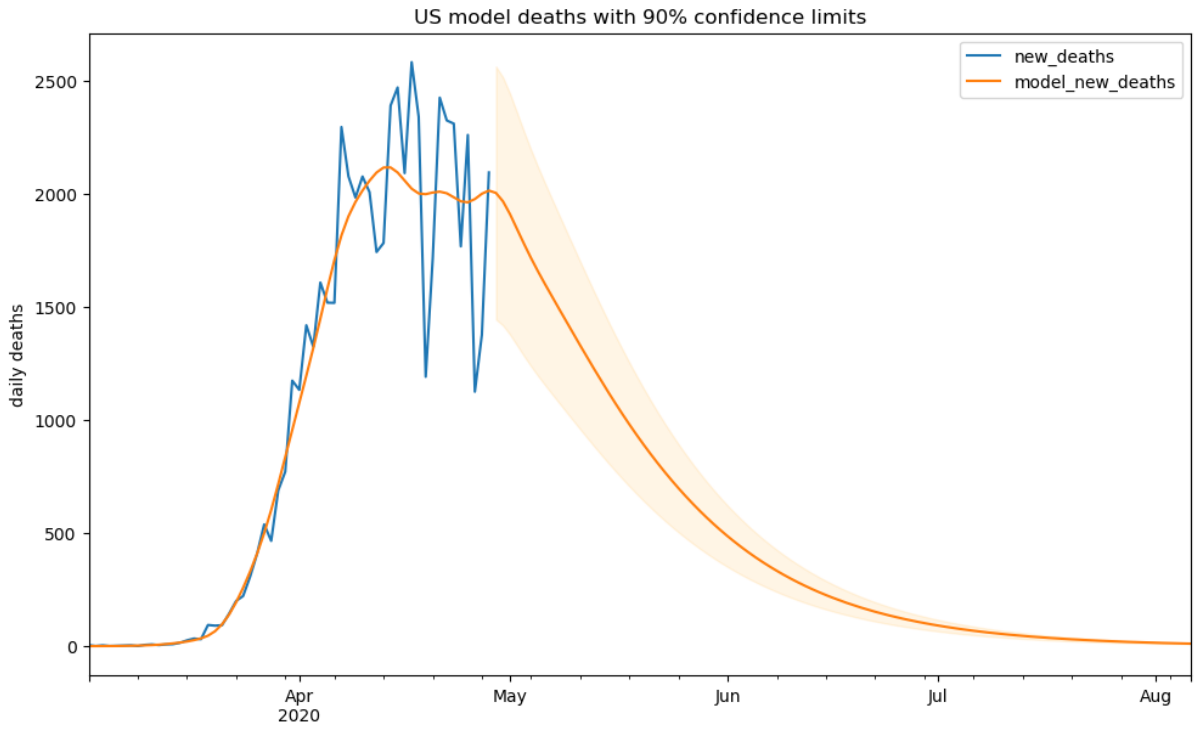


Italy: cumulative deaths by 2020-08-06 of 34954, 78 % of 27359 deaths to date

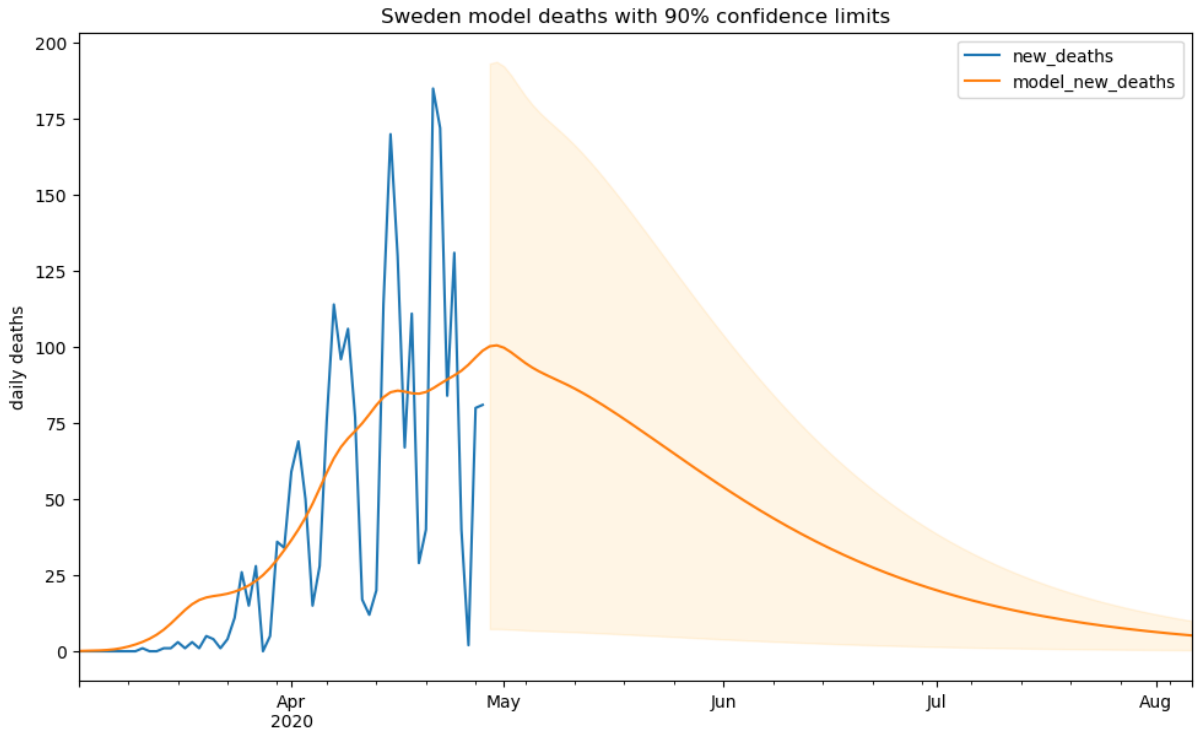


Spain: cumulative deaths by 2020-08-06 of 30468, 78 % of 23822 deaths to date

Here are various other countries in the Johns Hopkins dataset:



US: cumulative deaths by 2020-08-06 of 105928, 55 % of 58355 deaths to date



Sweden: cumulative deaths by 2020-08-06 of 6460,36 % of 2355 deaths to date

SUMMARY

Cases and ***deaths*** data each have their own pros and cons. Cases lead deaths, but are incomplete sample of the population. Deaths

lag cases, but are a more objective and complete picture even with the caveats above.

By using a negative binomial model to combine cases and deaths, we hope to eliminate some of both of these sources of noise and get a better read on the underlying trends. The model does explain some of the variation in daily deaths, but uncertainty in survival rates and new cases growth still make it difficult to discern trends with a great degree of confidence.

The charts above are saved to the github repo here as new data emerges—or you may prefer to generate them yourself.

ACKNOWLEDGEMENT

Greg Solomon and his early insights into the significance of the virus were the original impetus for this model.

APPENDIX

Here are some further notes on the models and assumptions used.

Models for the rate of growth of new cases

Policymakers will initially concentrate on this source of uncertainty in the short-term, since it is easier to influence than the survival rate. Ultimately we expect that the growth in new cases with tend to zero, so it is no surprise we see a negative growth rate β in the model $exp(k+\beta t)$. There is a test function in the code that confirms that if new cases had remained constant in absolute terms after reaching 100 cases, the fitted β would be presently be of the order of -2.5%. Since the fits for the countries in the chart are far lower than this, we assume above that the growth of new cases is declining in absolute terms with time.

It is useful to exponentially weight the data with a halflife in days. A feasible range is between 2 and 10 days. The fitted models use the median β within this range.

The confidence intervals shown are based on a projection of new cases at this median beta. The true confidence intervals are almost certainly wider than this central projection.

Summoning up the central limit theorem

If the lag ***k*** between infection and death follows a negative binomial distribution , the number of deaths at each future date ***t*** from cases ***k*** days before will follow a binomial distribution. For deaths in the order of magnitude of this pandemic, we can use the central limit theorem to derive confidence intervals for these daily deaths. Such bounds will be v e r y w i d e and won't include the model and parameter uncertainty that our bounds based on residuals include, not to mention the uncertainty in the new cases growth rate.

The central limit theorem will be harder to justify when aggregating daily deaths to find cumulative deaths of pandemic, because deaths are less likely to be independent. I therefore decided not to pursue this approach.