

PREPRINT REVIEW

Title: Estimating dates of origin and end of COVID-19 epidemics

Authors: Thomas Bénéteau, Baptiste Elie, Mircea T. Sofonea, Samuel Alizon

Comments to the authors

I have reviewed your preprint about estimating dates of origin and end of Covid-19 epidemics, and am offering some comments. My background is in computation and theoretical ecology, also with background and publication history in the sister science of theoretical epidemiology.

Abstract

First, your statement about the intensity of interventions, I believe, is right on track, and outbreaks of serious diseases that have been arrested before they became pandemics can result from careful intense intervention throughout the period of the disease. Also, the necessity of considering stochasticity at the beginning and at the end, when numbers are small, and considering super-spreading events is crucial in a disease model.

A perhaps minor point, I'm not sure the term "memory effects" will be understood by general readers, and should be explained in a few words in the abstract where it first appears, I think. In fact, I was not sure of what it meant when I read the abstract, having it conjure thoughts of memoryless distributions like the exponential, and other processes that are completely specified by only a state variable, such as a one-dimensional dynamical system.

Introduction

Another perhaps minor point, when you say in the introduction, "Most models," that opens it up to unnecessary argument. In fact, I immediately thought that by far most of my work on epidemiological topics has been on large-scale individual-based models, all of which were intrinsically stochastic, with the stochasticity emerging for small numbers and behaving deterministically for large numbers. So if you would just say, "Many models," it might be more accurate, and would avoid pointless objections.

A positive point, I think your discussion of the importance of the distribution of individual R_0 values is crucial and right on the point, especially for something like Covid where some large groups tend not even to believe in the existence of the disease, or in the value of any measures to contain it.

The discrete stochastic model

One question I have is, who is your audience? Are you aiming this at experienced mathematical modellers, or would you like to reach more general audiences, including students of epidemiology? If more general, then I would suggest a sentence explaining why the Poisson distribution is relevant here. It is the distribution of counts, so it applies here, but many aren't aware of the subtleties of different distributions, and a sentence or two could have a positive effect by helping to communicate that.

PREPRINT REVIEW

Related to audience, when I first saw t introduced, I saw it with the notation $t \in \mathbb{N}$, and I thought two things, (a) that t is a natural number, as indicated by the notation, and (b) hmmm, why use this mathematical notation? I think about what a great now-late mathematician told me when I was younger, “Do not put in symbols that which adequately can be explained in natural language.” All of my mathematically trained students and colleagues know the meaning of those symbols, but I think few of my biologically trained students do. You do shortly thereafter say that t is a number of days, but then shortly after that, in Equation 1, the first summation iterates i from 0 to t . That leads to confusion, because *zero is not a natural number*. And I agree it is convenient and common to use $t = 0$ as the starting time. But then you must not say $t \in \mathbb{N}$, right? Won't that generate confusion, or make it look like you are careless with meanings?

Regarding Equation 1, which is the starting point for the modelling discussions, let's see how well that can be understood by the general epidemiological modeller. First, Y_t is given as the modelled incidence on day t —which is to say the number of new infections on day t . (Though incidence can also mean the *proportion* of the population newly infected on day t , or other time period, so it might be good to clarify that.)

The calculation then sums over all days of the disease period, starting at time 0 and ending at the present day t ($\sum_{i=0}^t$). For each day within that, it sums over all new cases for that day, adding up the number of infections caused by those new cases for that day ($\sum_{k=1}^{Y_i} F_{k,i}$). Wait, something seems wrong! You can't calculate based just on new cases that day. It must be calculated based on all individuals who are infectious on that day, mustn't it? Do you mean to say that Y_i is *prevalence* rather than *incidence*?

Anyway, $F_{k,i}$ is defined as the *force of infection* for individual k first infected at time i , and that is multiplied by the number of individuals infected (multiplied in effect by summing over all such individuals). But force of infection is not to be multiplied by the number of infectious individuals, rather by the number of susceptible individuals who have been exposed, right? Are you referring to the infectivity (β) instead? That is the parameter that is to be multiplied by the number of infected individuals, properly prorated by the number of susceptible individuals as the infection expands through the population. I see that in Equation 2, where all individuals behave identically, you seem to switch to the term infectivity rather than force of infection.

In summary of this part, and throughout your whole manuscript, I would recommend that you carefully write out what you mean by each parameter, rather than just applying terminology like incidence, force of infection, infectivity, and so forth, and also make sure that your terminology matches its common useage. It doesn't seem to here, and that will cause confusion and doubt in readers, as it does in me. If there is confusion in the literature on some terminology, then state that and explain how you are defining the terminology for your manuscript. Otherwise it will become difficult for your readers to discern what your equations mean, or they will give up and conclude that your methods are not verifiable or comprehensible.

PREPRINT REVIEW

Computations

A principle that can be followed to help insure correctness is not to trust any mathematics that has not been verified numerically, nor to trust any numerical procedure that doesn't have a mathematically representation in its simpler forms. Of course, this cannot always be observed, because there are some procedures that do not have simple mathematical underpinnings.

Related to that principle, I see in your supplemental material that you provide differential equations further extending some of your work. Since supplemental materials are not particularly limited in size, it could also be useful to provide the actual source code, which would allow others directly to replicate and extend your work. That of course requires careful documentation of the source code, but such documentation also improves reliability of the results.

Earlier in the Covid pandemic, the world learned of a model used for consequential purposes that consisted of thousands of source line of undocumented C code, which was apparently not available at that time for review. Reliance on such models could backfire on the idea of modelling itself, and you could help combat that by putting well-documented source code in your supplementary materials.

Summary

I think you have important material to discuss concerning the stochastic nature of the onset and demise of disease outbreaks, including the current pandemic, and also about the effects of different rates of infections among different individuals or subpopulations.

However, I suggest that you very carefully review your use of terminology, and explain and verify how it fits into the mathematical forms, with simplified examples as part of the explanations, to help move your preprint to the next stage.

Best wishes!