All models are wrong, but some are useful

Lukáš Poláček

2020-04-10

Radoslav Harman, my former statistics professor (and also one of my favorite people), is working on a stochastic model of COVID-19 in Slovakia. Below is a picture from my simulation of the model. Red/orange line is the actual number of new positive cases each day – the real data. Blue dots represent outcomes of the simulation, the darker the more frequent. The picture is made of out 200 runs for one particular setting of free parameters b_0 , α and *prefix* (their meaning and exact values aren't important for now).



Figure 1: Simulation of daily new COVID-19 cases

On March 21, 41 people tested positive out of 440. The previous day (March 20), 13 out of 367 tested positive. With only only 20% more tests on March 21 we got 3 times more (41 versus 13) positive results!

The model and my simulation predicted about 10 to 30 positive tests from March 20 to March 22. The simulation is clearly off around this time – it overshoots on March 20 and 22 and undershoots on March 21. How could we fix it? My suspicion is that the population tested on March 21 was qualitatively different, for example it could have contained a big group of people returning from abroad. Actually, this is exactly what happened on April 5, when a group of 35 people returning from Austria tested positive. April 5 is the second to last day in the graph and for this day the model underestimated the number of positive new cases by a huge margin.

The result of the model could be improved by providing information about where each tested person

traveled and who they interacted with in the last few days (and this is exactly the type of information that needs to be collected for life to get back to normal).

All models are wrong, but some are useful. – George Box

The model is wrong about daily new cases, but actually being right about daily new cases is not its goal. It's supposed to describe the behavior of the epidemic in the long term. The following graph is similar to the previous one, but contains cumulative/total cases instead. Also, it contains two more days. Again, the red line is the reality while the simulation is blue.



Total COVID-19 cases for $b_0 = 190$, $\alpha = 1.25$, *prefix* = 4

Figure 2: Simulation of total COVID-19 cases

It looks like this configuration of parameters b_0 , α and prefix fits the cumulative data fairly well, since the red line stays within the simulated blue zone (except for the last two days when we know that the tested population was somewhat special). As Rado said himself, right now the model admits two scenarios fairly different from each other: one where the disease is spreading slowly since mid-February and one with the disease spreading since the end of February, but much more rapidly.

Is such a model useful? Sure, it could be more precise, but at least it gives a range and that's a start. I'd say that it's *fairly useful*. And as always, we need more and better data.

Exponential-growth models no longer useful

Exponential growth has been mentioned a lot regarding the spread of COVID-19. However, in countries implementing measures such as social distancing (most of Europe) or contact tracing and quarantine (Asia), models with exponential growth are so far from reality that they are no longer useful. And compared to the model above, they cannot be improved with more and better data.

Exponential functions are straight lines on semi-log graphs, so the total number of cases should form a straight line on such a graph. Below is a picture of total cases in Italy on a semi-log graph together with the exponential function $200 \cdot 1.145^t$, where t is the number of days since the 200th case.



Figure 3: Total confirmed cases in Italy on a semi-log graph.

The number of cases is not a straight line at all! Clearly, we haven't applied enough logarithms, so let's add one more log to the mix and plot it on a log-log graph. Note that now also the horizontal axis is log-scale.



Figure 4: Total confirmed cases in Italy on a log-log graph.

Now it looks much more like a straight line! For about 25 out of the 50 days plotted, it fits the function $2.5 \cdot t^3$ which is the straight dashed line on the graph. That's a polynomial and not an exponential function. Is it just a coincidence?

According to the paper Fractal kinetics of COVID-19 pandemic by Ziff and Ziff published in February 2020, the growth of active cases in China was much slower than exponential – roughly $0.0854 \cdot t^{3.09} \cdot e^{-t/8.9}$. Notice the exponent of 3.09 being almost the same as 3 for Italy, though the formula plotted above for Italy didn't have exponential decay. There is another important difference – the two graphs for Italy contain *total confirmed cases* while Ziff and Ziff quantify *active cases*. In particular, active cases can decrease while total confirmed cases never decrease, since it's the number of people who tested positive.

A few weeks ago, Slovak mathematicians Katarína Boďová and Richard Kollár looked at multiple countries using the same principles presented by Ziff and Ziff, and made a generalized formula for the number of active cases N(t) on day t (t = 1 is chosen as the first day with 200 active cases).

$$N(t) = \frac{A}{T_G} \cdot \left(\frac{t}{T_G}\right)^{6.23} \cdot e^{-t/T_G}$$

- T_G is a country-specific parameter the average speed in days in which "the sick are removed from the system" (to stop spreading the disease).
- A is a scaling constant.
- 6.23 is not just some constant that fits the data but a root of an important equation according to the authors. We'll have to wait for the manuscript to know more details.

On 2020-03-30, they made the following predictions for active cases in 7 countries.

Country	T_G	A	Max. cases	Max. date
USA	10.2	72329	$1\ 241\ 389$	2020-05-08
Spain	6.4	3665	99 978	2020-04-12
Italy	7.8	4417	99 459	2020-04-12
Germany	6.7	3773	98 038	2020-04-14
United Kingdom	7.2	2719	66 082	2020-04-21
France	6.5	1961	53060	2020-04-12
Iran	8.7	2569	51 773	2020-04-21

Scoring the predictions

On 2020-04-10, 11 days after the predictions, I've looked at the data to see how did the predictions do. However, I've disqualified France, since they first withheld information for multiple days and then reported all of it on 2020-04-04, making it unfair to anyone predicting trends. Slate Star Codex questions Iran's data, but I haven't investigated that, so I keep my rating for now.

Country	Subjective rating
Italy	****
United Kingdom	****
Spain	****
Germany	****
USA	****
Iran	**

Italy

The prediction (dashed line) is spot-on. Note that the green zone marks the data available at the day of prediction, so until March 29.



Figure 5: Active cases in Italy until 2020-04-09 together with the predicted trend

Spain

The original prediction (upper dashed line) was a bit pessimistic, but another curve with a slightly lower T_G of 6.2 (versus the original 6.4) fits the data really well so far.



Figure 6: Active cases in Spain until 2020-04-09 together with the predicted trend

Germany

Germany, like Spain, at first looked like having lower T_G of 6.3, but recently their patients are recovering really fast and the number of new cases is steady. It might mean that their T_G got even lower but it could also mean that there is something wrong about the model.



Figure 7: Active cases in Germany until 2020-04-09 together with the predicted trend

USA

USA had the most pessimistic prediction and it's good that the number of cases is smaller than predicted. I scored it highly ($\star \star \star \star$), because keeping the predicted T_G but lowering the value of A by 15% is in line with the trend. However, USA is still before the first inflection point, so it's too early to make any confident judgements and I'm not confident about my exact score either.



Figure 8: Active cases in USA until 2020-04-09 together with the predicted trend

For the sake of brevity I skipped the two remaining countries, but you can check their predictions on a web dashboard. The data is updated daily.

Discussion

The classic exponential-growth models have a key assumption that infected and uninfected people are *randomly mixing*: every day, you go to the train station or grocery store where you happily exchange germs with other random people. This assumption is not true now that most countries implemented strict measures such as *social distancing* or *contact tracing* with quarantine.

You might have heard of the term six degrees of separation, that any two people in the world are connected to each other via at most 6 social connections. In a highly connected world, germs need also very short human-to-human transmission chains before infecting a high proportion of the population. The average length of transmission chains is inversely proportional to the parameter R_0 (which you probably heard of).

When strict measures are implemented, the random mixing of infected with uninfected crucial for exponential growth is almost non-existent. For example with social distancing, the average length of human-to-human transmission chains needed to infect high proportion of the population is now orders of magnitude bigger. It seems like the value of R_0 is decreasing rapidly with time, since you are meeting the same people over and over instead of random strangers. The few social contacts are most likely the ones who infected you, so there's almost no one new that you can infect. Similarly for contact tracing and quarantine – it's really hard to meet an infected person when these are quickly quarantined.

The formula $N(t) = (A/T_G) \cdot (t/T_G)^{6.23} \cdot e^{-t/T_G}$ has two free parameters A and T_G . One objection to the finding might be that with two parameters you can create many different functions, making fitting arbitrary curves easy. However, a simple analysis shows that A and T_G only scale the graph of the function vertically and horizontally. The observation is left as an exercise to the reader. As a hint, here's a picture of three functions $t^{6.23}/e^t$, $(t/2)^{6.23}/e^{t/2}$ and $2t^{6.23}/e^t$.



Further reading

- Richard Kollár on Facebook, he regularly posts updates about their research.
- Fractal kinetics of COVID-19 pandemic by Ziff and Ziff
- Polynomial growth in age-dependent branching processes with diverging reproductive number by Alexei Vazquez.
- Martin Niepel's notes (in Slovak only)
- Martin Niepel's lecture (in Slovak only)
- Richard Kollár on national TV (in Slovak only)

Back to the first model

The model of Radoslav Harman is an explanatory model: it tries to explain the past. It estimates when the infection arrived to Slovakia and how fast it's spreading since then. Also, it's estimating how good the testing selection process is, whether we test too many people with common cold or flu instead of COVID-19 patients.

These three estimation goals are expressed as the following parameters of the model.

- b_0 is a measure of overall quality of the testing selection. That is, the greater b_0 , the better the efficiency of the system in restricting non-COVID-19 individual from testing. The smaller b_0 , the more non-COVID-19 individuals get tested.
- α or γ_2 describe the growth of infected cases. The model works both for polynomial growth (parameter α) and for exponential growth (parameter γ_2).
 - α is the exponent in the polynomial growth function t^{α} . It's very hard to estimate T_G for Slovakia, since we have few dead and recovered, so there is no support for exponential decay yet.
 - $-\gamma_2$ is the rate of exponential growth after March 12, when Slovakia implemented restrictive policies. Judging by the mobility report released by Google, March 12 seems to be the right choice.
- t_{max} is the total length of the simulation. (My code instead uses $prefix_length$ as a parameter and calculates $t_{max} = prefix_length + days$, where days is the number of days for which we have data.)

For each combination of parameters, we run multiple simulations. Each configuration of parameters receives a score based on how close it is to reality.

The code is on github **O**, including visualizations and dashboards running on a web server. If you'd like to help in any way, feel free to get in touch.

Picture time!

Let me end with some pictures from the simulations. They confirm Rado's claim that we cannot distinguish between start of the infection in mid-February versus end of February with faster spread. You can also view the visualizations online: polynomial and exponential growth.

The axes of the heat maps on the pictures show b_0 and prefix length. The closer the color on the heat map is to yellow, the better a particular configuration matches reality. Notice that yellow color spans through the pictures, so we have plenty of fairly good configurations. There are two pictures for polynomial growth of $t^{1.24}$ and $t^{1.28}$, and one for exponential growth of 1.04^t . So far, the scores of these three setups aren't that different from each other, but we expect it to change with time.





Figure 9: Heat map of errors for different b_0 and prefix length, with growth of $t^{1.24}$



Figure 10: Heat map of errors for different b_0 and prefix length, with growth of $t^{1.28}$



Figure 11: Heat map of errors for different b_0 and prefix length, with growth of 1.04^t