SFI TRANSMISSION

COMPLEXITY SCIENCE FOR COVID-19

STRATEGIC INSIGHT: To forecast the spread of the novel coronavirus, we must attend to the quality and consistency of the data.

FROM: Luu Hoang Duc, Max Planck Institute for Mathematics in the Sciences; and Jürgen Jost, Max Planck Institute for Mathematics in the Sciences and the Santa Fe Institute

DATE: 30 March 2020 **NO:** 003.1

There is no shortage of data on the unfolding coronavirus epidemic. Countries around the world are publishing daily case counts, which should constitute a digital treasure trove for those of us who seek to understand and even forecast the spread of the epidemic. The problem with this massive quantity of data is its quality — datasets from different countries are not really compatible with each other, are often internally inconsistent, and in some cases could be politically manipulated.

So, what's a complexity scientist to do? In our research group in Leipzig, we believe we can establish general statistical regularities using simplifying assumptions and procedures that can compensate for data fluctuations. Below, we provide a few examples of problems that arise from inconsistent data, and solutions for making the most of it.

For each of the countries we survey, we distinguish different periods of pandemic development based on the respective growth rates for the number of infections recorded. In the beginning, the growth rate is typically extremely high but then weakens. In the final saturation phase, the growth rate has become so low that the development of the epidemic is essentially under control. Various countries are currently at different stages of development. In countries in which the growth rate is still very high, as is currently the case in Germany, it must be expected that a saturation phase will only occur after much higher case numbers.

How to handle data that are difficult to compare, unreliable, and inconsistent

There are now many data points from many countries on the spread of the coronavirus epidemic, updated at least once a day. But, as mentioned above, the data from different countries are difficult to compare.

Here are some of the problems: Test density and methodology vary greatly; not all virus carriers also show symptoms; not all infected people are identified; hospitals do not necessarily report releases to the authorities; those who have recovered at home will not always report; and the death toll is unclear, because it is difficult to distinguish between people who die *from* corona versus *with* corona. The epidemic is over when the number of active cases is zero, calculated as the difference between infected and recovered or deceased persons, but this calculation may be inconsistent or incorrect.

So, how can we deal statistically with such a data situation? Is it still possible to gain general insights into the course of the epidemic and perhaps even to make predictions about how long it could take for individual countries to bring the epidemic under control?

In short, we have to use simplifying assumptions and simple, robust procedures that can compensate for data fluctuations. Here, we assume that the ratio of reported cases to actual cases will remain reasonably constant, i.e., typically the test methodology and coverage will not change. Then the respective rates of increase will also be similar, and we can draw conclusions about the actual cases from the increase in reported cases.

We draw a simple regression line through the logarithmic rates of increase. Extrapolating this line yields a prognosis — very rough, of course — at which numbers the epidemic can probably be controlled and how long it will possibly last. The actual development will naturally depend on the measures taken to contain the epidemic and their implementation and compliance by the population.

We use data provided by WHO[1] and WORLDOMETER[2] and evaluate them for countries with over 1,000 reported infections. Here, we show here the data for Italy.[3]

The first two graphs on page 003.4 show the numbers of infected, deceased, and recovered persons and the active cases over time. The next two graphs show the number of daily new infections and new deaths. The final two graphs compare the growth rate of new infections with their linear regression (blue line). Large deviations from this straight line may indicate problems or systematic changes in data collection. The flatter the blue line is, the slower the epidemic weakens.

Heterogeneous contact networks

We also see important consequences for scientific models of the spread of epidemics.

Diseases are transmitted through contacts, and therefore many propagation models try to capture the network of social contacts. Typically, a fairly homogeneous network

[1] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/

[2] https://www.worldometers.info/coronavirus/

[3] See https://www.mis.mpg.de/covid19 for full data.

structure is used as a basis for the model to remain manageable, but the spreading of the coronavirus epidemic points to very heterogeneous underlying network structures. In South Korea, for instance, the virus apparently spread very quickly and strongly within a particular sect, which was favored by intensive contacts within the sect, but could then be confined because contacts with the rest of the population were apparently much thinner. The death rate in Italy is comparatively high, probably because the contacts between generations are more intense, allowing the virus to quickly reach the elderly, whereas in Germany and the Scandinavian countries it was probably first spread by ski tourists returning from their winter vacation in the Alps. In the Scandinavian countries there also seem to be two different waves of spread, unlike in the rest of the world.

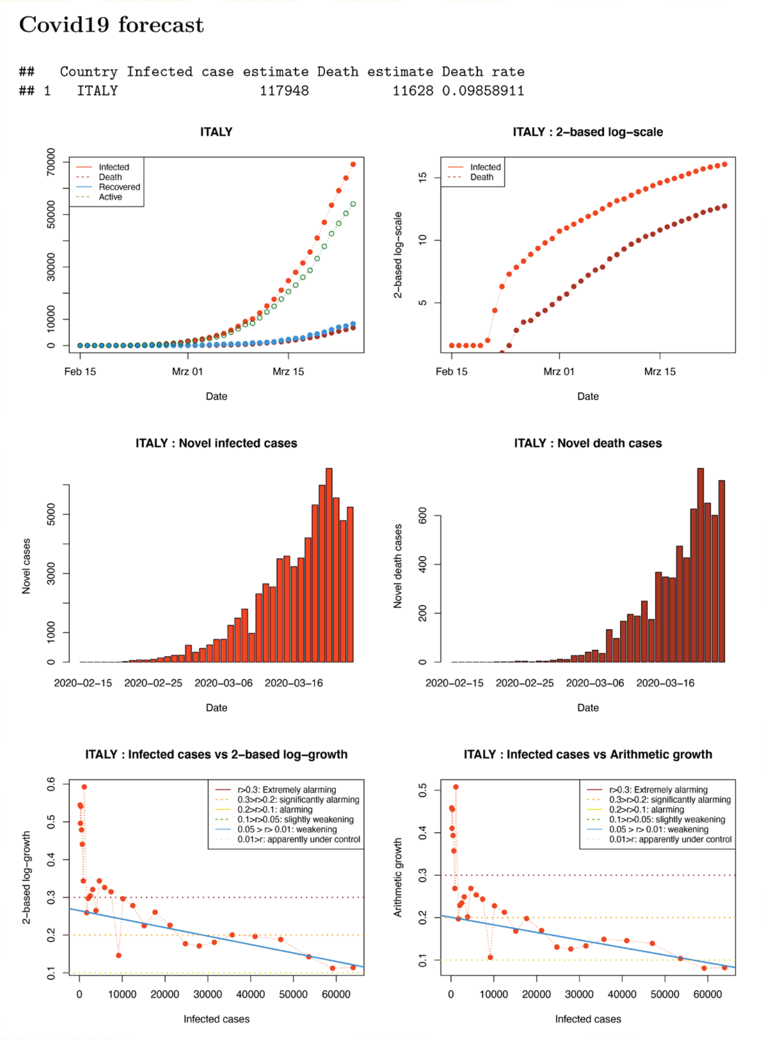
Conclusions

Reported data about the COVID-19 epidemic are obviously incomplete, vary greatly between countries, are possibly politically manipulated, and are typically internally inconsistent. If we want to draw any reasonable statistical conclusions at all from such data, we need methods that can identify some robust trends. We, therefore, looked at the dynamics of the growth rate and see regularities there that are captured by a simple linear regression. This leads to prognoses, which, of course, are rough and tentative and will be affected by political measures taken and the compliance by the populations in the various countries.

We also see the scientific challenge for models of epidemic transmission in networks of social contacts that may be very heterogeneous — for instance, subgroups with few outside contacts, or in contrast inter-generational contacts that can quickly carry infections into high-risk groups.

We hope that a deeper understanding of these and other problems will allow us to better cope with such epidemics in the future.

Read more posts in the Transmission series, dedicated to sharing SFI insights on the coronavirus pandemic: santafe.edu/covid19



Covid19 forecast

```
##   Country Infected case estimate Death estimate Death rate
## 1   ITALY                 117948          11628 0.09858911
```