Predicting Case Fatality Rate for COVID-19 from outside air temperature

Benedikt Hopf
Matrikelnummer 4243889
benedikt.hopf@student.uni-tuebingen.de - github repository

Abstract

In this project I use German COVID-19 data from the RKI (Robert Koch Institut) together with German Weather Data, provided by the Deutscher Wetterdienst (German Weather Service), to predict the COVID-19 case fatality rate using outside air temperature and time since outbreak. I use a simple custom model to recreate the actual numbers and show, that COVID-19 in Germany might have been much worse, had the breakout been a couple of months earlier.

1 Introduction

The pandemic caused by SARS-CoV-2 and the accompanying disease COVID-19 is clearly one of if not the largest worldwide crisis since the second world war. The virus originated in Wuhan, China in late 2019 and then quickly spread over the entire world, with the first infection in Germany recorded by the RKI on January 2nd 2020. The first wave then followed in March 2020 [1].

Obviously the development of a pandemic depends on many factors, but a recurring theme seems to be, that cases and case fatality rates are lower in summer, most likely due to higher temperatures. Therefore, people spend more time outside where aerosol transmission is harder, and as such fewer people get infected. And the ones that do, do not get quite as sick, as the dose of virus they received is smaller. Pujadas et al. [2] showed, that viral load in the blood is an indicator for the risk of death.

So in this work we will see, how well case fatality rate can be predicted from outside temperature and the time passed since the first outbreak, since this also seems to lower case fatality rate. We will not look at the number of infections, since this is highly dependent on political measures like lockdowns and as such very hard to model.

2 About the data

For this task two sources of data have been used: For one this is the official German COVID-19 [1] data as provided by the Robert Koch Institute, which is the German authority that is officially responsible for tracking the pandemic, backed by the German Bundesministerium für Gesundheit (federal Ministry for health). The data is supplied as a `csv`-file. The temperature data [3] for Germany has been retrieved from the Deutscher Wetterdienst (German weather service), which is the official authority for weather in Germany, backed by the German Bundesministerium für Digitales und Verkehr (federal Ministry of digital matters and health). The data is supplied as several `zip`-files containing `csv`-files, that have been combined into one dataset.

Looking at the two datasets (case fatality rate and air temperature) next to each other it is immediately obvious, that there seems to be a negative correlation between them. High temperatures go with lower case fatality rates and vice versa. A possible explanation for this has already been given in section 1. Another thing to note, is that the case fatality rate is generally decreasing, which is likely due to medical advances and increasing immunity in the population.

3 Modelling case fatality rate

This section will discuss the model choice and provide a plot, that demonstrate model performance.

3.1 The model

As the model for this problem, a multiplicative model of an exponential and a sigmoid as been chosen:

$$a \cdot e^{b\cdot(t-\Delta)} \cdot \sigma\left(c \cdot (\vartheta - d)\right) \tag{1}$$

where

- t is the time since the first case in days.
- ϑ is the outside air temperature in $^\circ C$.
- a is a multiplicative scaling factor in $\frac{deaths}{cases}$.
- b is the rate of decay of the case fatality rate due to time, given in $\frac{1}{days}$.
- c is a multiplicative factor for the influence of temperature, given in $\frac{1}{^\circ C}$.
- d is an offset for temperature, given in $^\circ C$.
- σ, refers to the logistic sigmoid $\sigma(x) = \frac{1}{1+e^{-x}}$.
- Δ is the day of the peak of the case fatality rate during the first wave.

Note, that Δ is mathematically unnecessary as it can be completely absorbed in a, and only serves to shift the data to a fixed location and thus make the parameters more interpretable. t and ϑ are the inputs, a, b, c, d are learned and Δ is fixed.

This model has several advantages, compared to a simple linear model like

$$a \cdot t + b \cdot \vartheta + c. \tag{2}$$

These advantages are in the range of the inputs and the result. Case fatality rate is a metric that cannot become smaller than 0, which a linear model could. Also, the reason to include the temperature is, that it is assumed to be a measure of how much time people spend outside. This is a metric, that can attain values in $[0, 1]$ and neither more nor less. When it gets warm in summer at maximum all people can be outside at all times and never more. The same holds on the other side in winter. Time is included, as it can measure medical progress, which makes sense to be quite quick in the beginning and then becoming slower over time, as there are less undiscovered easy improvements. So an exponential fits that shape. Finally, a multiplicative model makes more sense than an additive on, since the assumption is, that good enough medical advancements can eventually counteract temperature variations completely and high temperatures make, that medical help is not as necessary, which cannot be represented by an additive model. In a multiplicative model one of the factors becoming (close to) 0, results in the entire output being (close to) 0.

3.2 Training and basic evaluation

The model was then trained using `scipy.optimize`'s `minimize` routine and the *Nelder-Mead* method with quadratic loss. The resulting fit can be seen in Figure 1. One can see, that the prediction oscillates strongly, since also the temperatures are oscillating a lot. For this reason another line has been plotted, that is a 14-day average of the prediction. In general the prediction follows the actual case fatality rate pretty well.

Since the model is designed in such a way, that it cannot capture the beginning of a pandemic (as case fatality rate must be monotonic as a function of time), the model has only been trained starting from the peak of the first wave and should only be evaluated from there. The time before that is shown in the plot, but marked as before the beginning.

4 Scenario simulation

Now, that we have seen, that the model works decently well on data, that we know the right answer to, we can start with the actual interesting part of predicting, what would have happened, had the pandemic arrived in Germany at a different time.

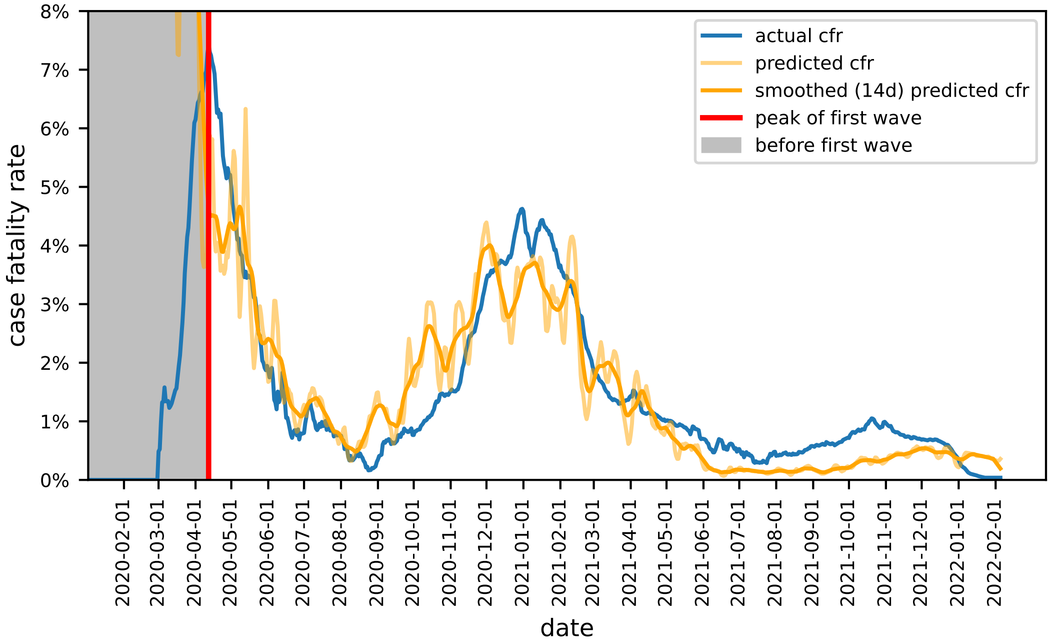

Figure 1: Actual case fatality rate and predicted case fatality rate plotted against time. The smoothed line has been created by averaging over 14 days.

A scenario like this has been calculated for up to one year before the actual outbreak and is shown in figure 2. One can see, that most curves peak much higher, which makes sense, since the actual pandemic was relatively quickly slowed down by the coming summer, which helped a lot at a time, where the pandemic was still very young. A more qualitative statement about that can be seen in figure 3.

5 Caveats

Obviously a pandemic is a very complex matter and cannot be described sufficiently well by just time since outbreak and outside air temperature. There are many other factors, that play a role in that, such as the wearing of facemasks, vaccination (which is somewhat accounted for in the decay over time) and the arising of different variants, such as the recent Omicron variant. As such these predicted scenarios might be far off. Also, the case fatality rate should not be confused with the number of deaths, since a higher number of deaths would lead to a stronger political reaction, which in turn would reduce the number of new infections.

On the other hand the model's predictions on past data are quite reasonable and there exists a convincing explanation for why things get predicted as they do as well as why the model is set up as it is (see sections 1 and 3).

6 Conclusion

Concluding one can say, that it is reasonably possible to predict case fatality rate from air temperature and time since outbreak, using a relatively simple and explainable model (section 3). However, these predictions are by no means perfect or guaranteed to be correct, as discussed in section 5, since there are a vast number of factors playing into a pandemic, such as political decisions.

Keeping in mind these caveats, the model was then used to predict, the case fatality rates, if the pandemic had had its outbreak at an earlier time. Doing so it is quite obvious, that we have been somewhat lucky, in terms of the time of outbreak, since most scenarios would have resulted in higher maximum case fatality rates and almost all of them in higher average case fatality rates.

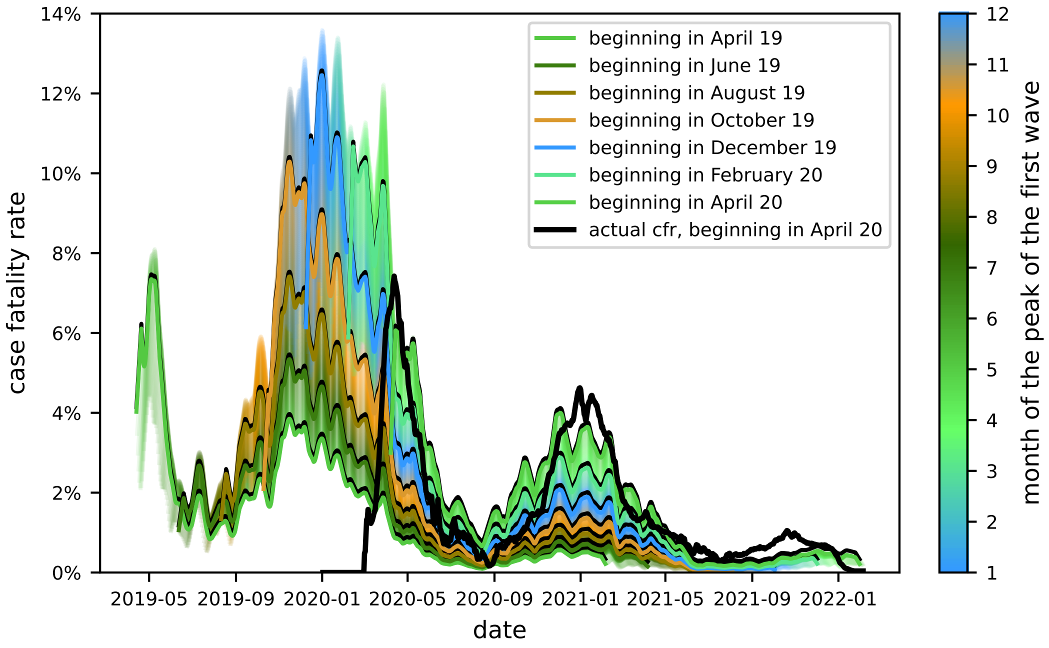

Figure 2: Predicted scenarios for peak to one year before April 2020. The actual case fatality rate is shown in black. The predictions are color-coded by the season, their beginning falls into. The background is made up of scenarios for every individual day.

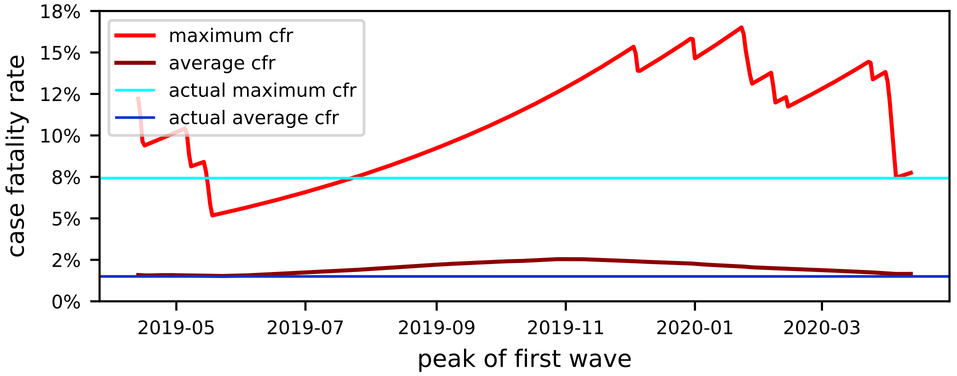

Figure 3: Maximum and mean case fatality rate as a function of the time of outbreak. The actual values of these quantities are shown as horizontal lines.

References

[1] Robert Koch Institut. Robert koch institut covid-19 data. download, Jan 2022.

[2] Elisabet Pujadas, Fayzan Chaudhry, Russell McBride, Felix Richter, Shan Zhao, Ania Wajnberg, Girish Nadkarni, Benjamin S Glicksberg, Jane Houldsworth, Carlos Cordon-Cardo, and et al. Sars-cov-2 viral load predicts covid-19 mortality. website, Aug 2020.

[3] Deutscher Wetterdienst. Deutscher wetterdienst daten. download, Jan 2022.