Understanding COVID-19 case and mortality data in Germany (2020-2023)

Motivation

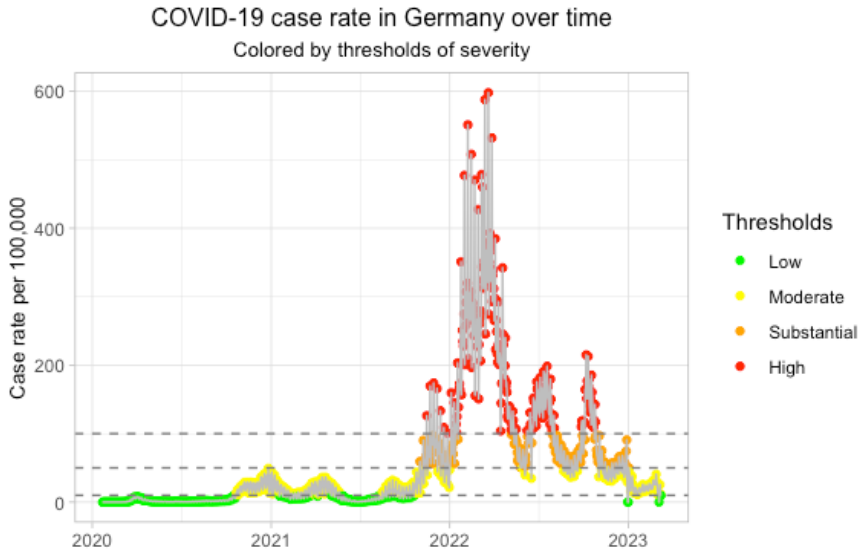
The COVID-19 pandemic had an enormous impact on many areas of public and private life, causing health problems as well as socio-economic challenges. It also led to an unprecedented abundance in data on the spread of infectious diseases. In this project, we visualize time series data on the disease between 2020 and 2023 for Germany, shedding light on the development of the pandemic in order to help informing evidence-based decision making in the future. In particular, we are investigating four different research questions, which will be introduced throughout this report.

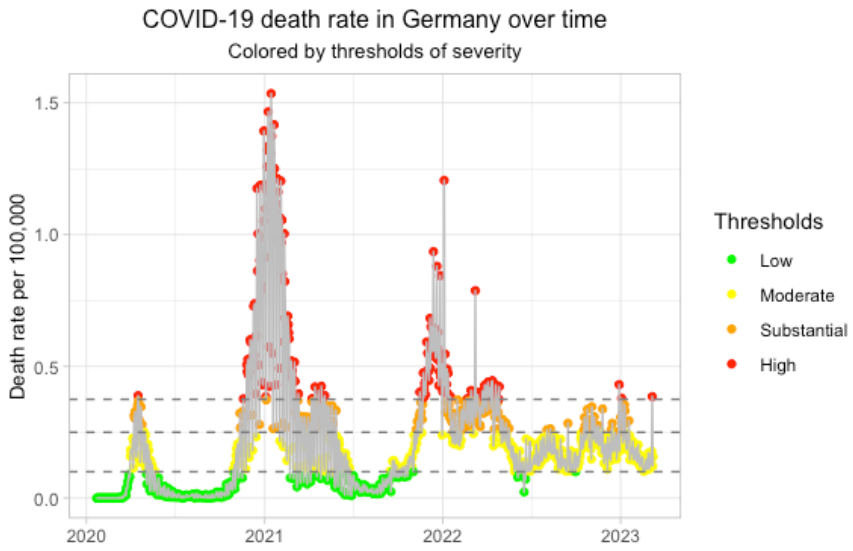
Data

We retrieve a comprehensive and well-curated global COVID-19 global data set from the Johns Hopkins University.[i] This data set comprises cumulative COVID-19 infection and death data for most days of the pandemic (2020-2023) and a large set of countries. Focusing on the time series for Germany, we clean the data by transforming cumulative into daily data and imputing some missing values (there is no data reporting on the weekends) with running averages. This way, we obtain a final data frame consisting of 17 variables with 1143 observations, corresponding to the days between January 2020 and March 2023. The variables include date, case and death numbers, calendaric data as well as some auxiliary quantities and categories.

Q1: General patterns and public policy

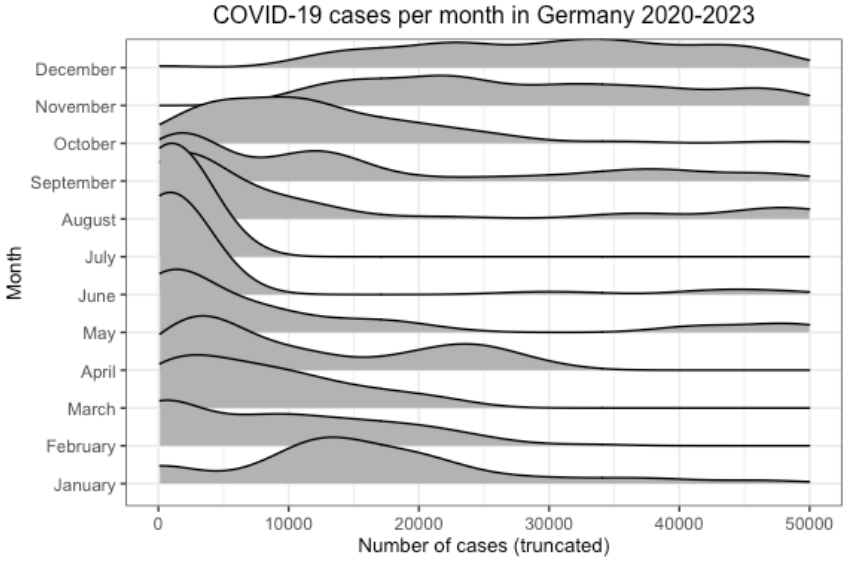
The first question we are interested in is what general (seasonal, yearly) patterns we can identify in the data, i.e., when peaks and troughs occur. Also, we want to elaborate on the effect of public health measures on case and death numbers. In order to do so, we firstly construct a thresholded time series scatter plot, visualizing COVID-19 case and death rates over time.





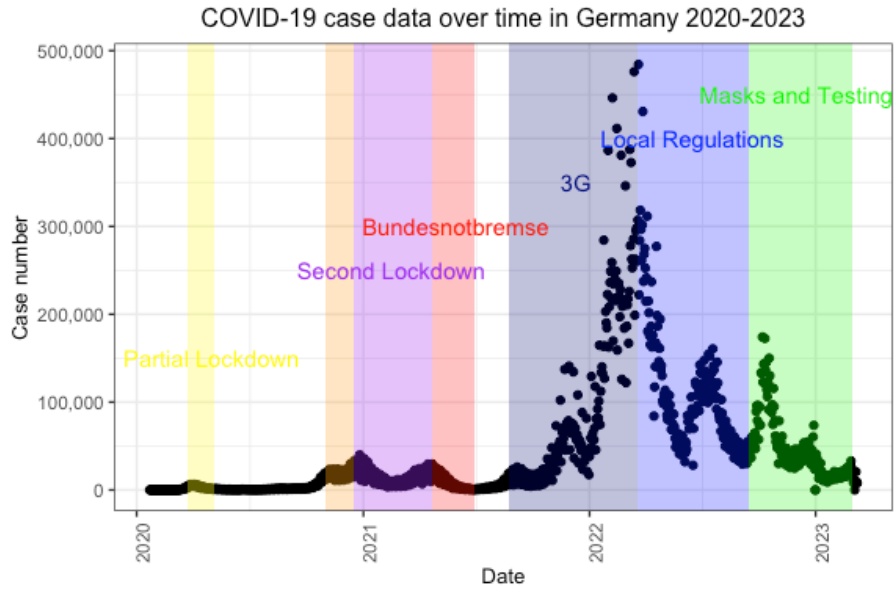
In both plots, we can see the occurrence of multiple peaks (waves of the pandemic), indicating the cyclical nature of the disease's spread. Interestingly, the strongest peak in case rate occurred in early 2022, while the strongest peak in death rate occurred in early 2021, with later peaks decreasing in scale. One potential explanation for this is that appropriate preparation of hospitals was able to reduce mortality even in the face of drastic infection rates at later stages of the pandemic. This relationship will be investigated further in the next section.

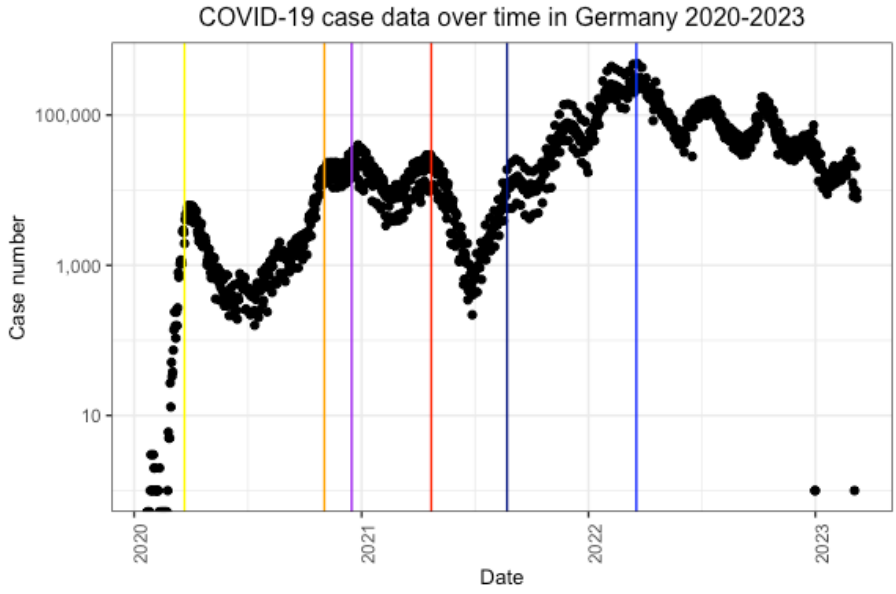
The periodic nature of the time series motivates considering the distribution of case numbers per month, visualized in a ridgeline plot.



It is apparent that there exists a stark discrepancy in monthly case patterns, with winter months generally experiencing a considerably higher amount of case numbers. (A similar pattern is evident in deaths.) From an immunological perspective, this seems reasonable since winter months often correspond to weakened immune systems due to the climatic conditions in Germany.

One central aspect of the analysis of disease data is to inform public health actionables with it. Hence, we visualize the relation between COVID-19 cases and public health measures in Germany. In order to get better traction on the effect of the measures, we plot the y-axis in logarithmic scale.



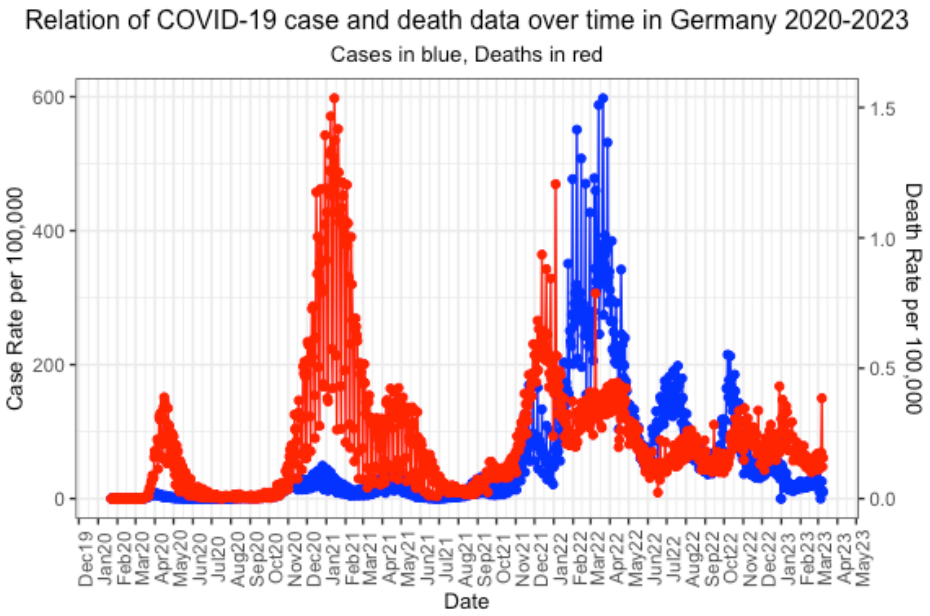


Here, it becomes apparent that after each implementation of health policy measures, the case rate declined, which indicates that the measures were successful in limiting the spread of the pandemic. Interestingly, in most cases the effects of the measures seem to fade after some time, which might be due to a tendency of the population to adhere to the regulations less strictly over time.

Overall, we found that the COVID-19 pandemic exhibited a cyclic nature characterized by different waves, where strong seasonal pattern can be identified. Additionally, we found a clear relation between the implementation of public health policy measures and a decrease in case numbers.

Q2: Relation between case and death rates

Building upon the availability of both case and death data, we next examine how case and death rates compare. Our visualization choice is a time series scatter plot with adjusted y-axis scales that allows for direct comparison of patterns and magnitudes even when the actual rates are scaled differently.

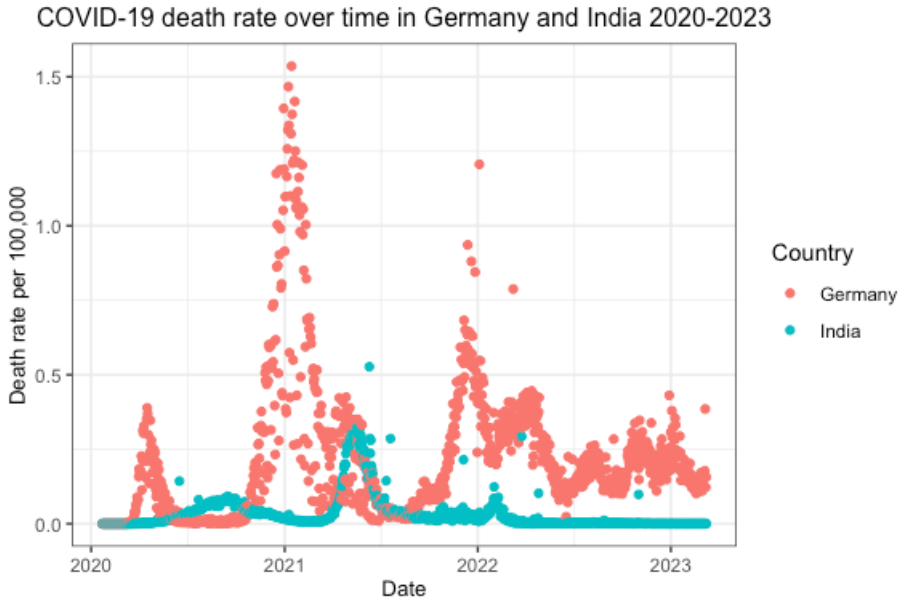


From this plot, it is evident that the cycles of increased case and death rates align, but that magnitudes differ greatly. That is, in the 2021 wave, Germany did not experience a high case rate compared to the 2022 wave, but the death rate surpassed that of the latter waves significantly. This suggests that the public health and (emergency care) medical system was simply not prepared for a large-scale pandemic like COVID-19, which led to a disproportionately high mortality in early stages of the spread. With time going on, epidemiological preparedness and societal readiness increasing, and society and hospitals adapting to the new situation, Germany was able to reduce the death rate, even when case rates immensely surmounted those of earlier wave.

Hence, we find that case and death rate behave quite similar over time in terms of spikes, as expected, but that the magnitudes of rates differ greatly, most likely due to external factors such as health care infrastructure.

Q3: Comparing COVID-19 globally

Transcending the view on one country, we strive to compare the pandemic in Germany with other countries as well. Here, we want to compare data in Germany and India, for which we use a time series scatter plot.

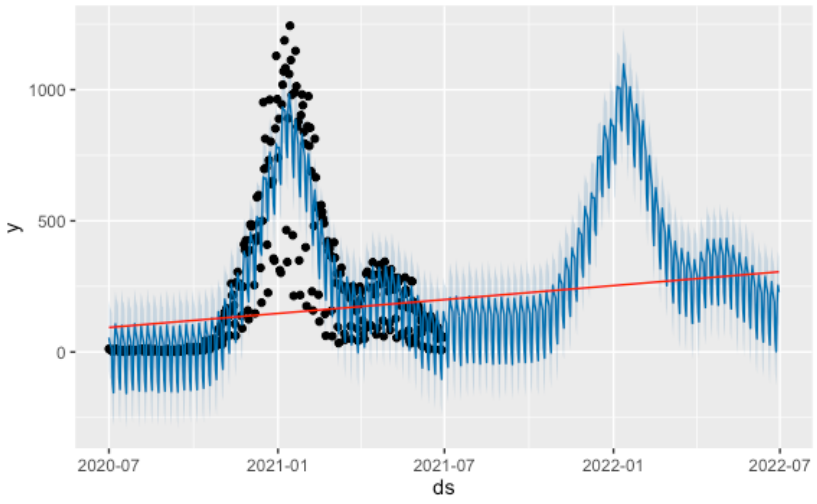


We can see that dependence of the pandemic on seasonality is evident in both countries (with cycles not aligning, which is due to different start dates and policies in each country), but again that magnitudes differ greatly. Interestingly, the death rate in India is considerably lower than in Germany over almost all of the three years. Now while this could lead us to claiming that India was more successful in mitigating drastic outcomes of an individual's infection through good medical infrastructure and urgent care, another reason for the apparent pattern in the above plot might be the data itself: Even though the Johns Hopkins University gathered the data centrally, the reporting mechanism of the raw data vary greatly by country. Hence, it seems likely that India experienced a massive underreporting of cases and deaths, whereas reporting in Germany was more comprehensive. Overall, however, the exact reasons for the trend remain opaque.

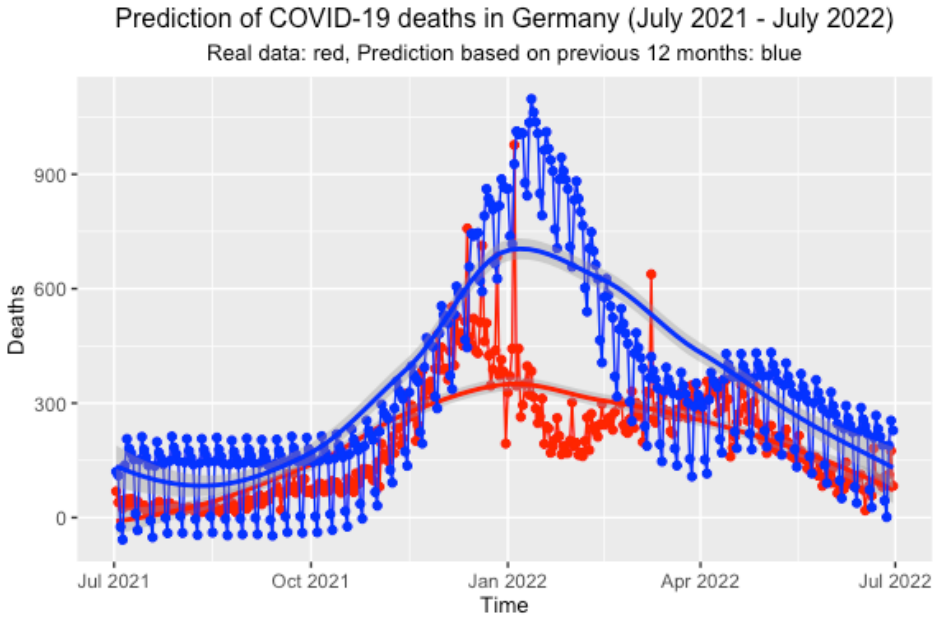
Q4: Predicting COVID-19 cases

Lastly, we want to venture into the wide field of epidemiological forecasting, and ask whether or not we can predict COVID-19 deaths based on historical data. In order to get traction on this question, we utilize the *prophet* library developed by Facebook[ii] for the purpose of improved time-series prediction.

Using case data from July 2020 to July 2021 as training data, the library outputs the following prediction for the 12-month period following the training data.

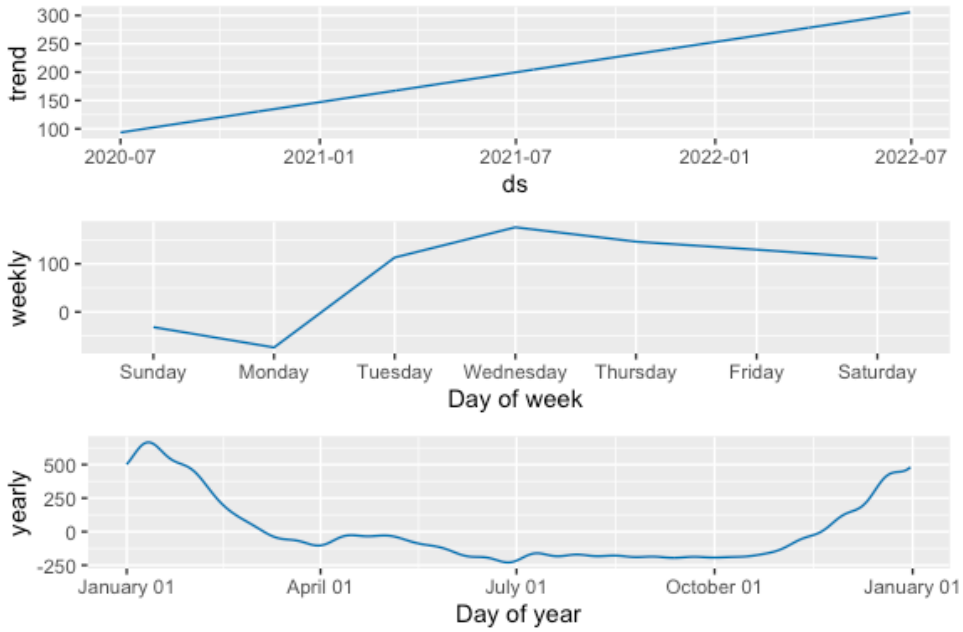


We see that the prediction replicates previous patterns, forecasting a wave in the winter months of 2022. Now we can compare the prediction to the actual data.



Here we see that forecasting is indeed possible, since the trend has been identified correctly. Interestingly, the spike in deaths in winter 2022 was overpredicted, which might result from new public health measures and improved readiness of hospitals that helped to prevent high mortality.

However, good information on seasonality and external factors (such as policy measures) is required for accurate predictions, and exact magnitudes are extremely difficult to predict. Also, the quality of the prediction depends heavily on the training data we choose.



On top of the above plots, the prophet library provides us with a breakdown of how different seasonalities (yearly, weekly) influence the prediction of death rates. The yearly pattern visible above verifies the observation we made earlier about increased deaths in the winter months. The weekday pattern potentially arises from increased social activities over the weekend, which then materialize in the data throughout the week due to a lag in reporting time. (We have to be careful with inferences based on

weekdays, since we have smoothed out the weekends already, and since there is most likely overreporting at the beginning of the week due to the surplus of weekend cases.)

Overall, we can state the forecasting is possible up to a certain degree accuracy, but we should remain aware that such predictions are always dependent on the data and a variety of external factors.

Conclusion

Throughout this project, we have seen that data science visualizations can be helpful to understand the COVID-19 pandemic. In particular, we examined the cyclic nature of the spread, seasonal patterns as well as differences between case and death rates, and between Germany and India. Importantly, we saw that timely public health measures mitigated the negative impact of the pandemic, especially in terms of death numbers.

In terms of limitations, we should remark that our analysis is somewhat one-dimensional, since we only focus on case and death rates, and not on other societal impacts on a socio-economic level that public health measures such as lockdowns can trigger. Also, we have to be aware that the reasons for visible patterns might not always be obvious or simple. As we hinted at in the section on country comparison, for example, different data reporting mechanisms can introduce bias into our visualizations. For time series predictions, we should note that the data was comprehensive for Germany, but only covered a relatively small time frame of three years due to the recent nature of the pandemic. For such short time windows, the predictions can potentially depend on random variance in training data too much (overfitting).

Moving forward, possible extensions of the projects include analyzing more than the two given countries to identify continental or global patterns, and to utilize other forms of visualization such as heatmaps or choropleths. Lastly, for predictive methods, we could consider more variables to increase accuracy.

[i] COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University: https://github.com/CSSEGISandData/COVID-19
[ii] Taylor, Sean & Letham, Benjamin. 2017. Forecasting at scale. 10.7287/peerj.preprints.3190v2.