A Web Data Dashboard for Covid-19 Analysis

Certification Project Data Scientist

by P. Eckert, Julia Poliak, Christoph Seng

04.12.2020 Bonn and Darmstadt

Agenda

1. Summary

The objective of this project was to compare the evolution of the COVID pandemic all over the world.

As none of the team members is an expert in terms of medicine, pandemics or epidemiology, it was decided to create one or more dashboards, that on the one hand give an overview over the pandemic itself but on the other hand provide an idea what kind of reports would be possible to provide to expert stakeholders.

The given data is very inhomogeneous from country to country in terms of frequency, content and reliability. So intense efforts were invested, to filter, rearrange and newly assemble the complex information structure.

First insights based on public discussions and arguments were generated and visualized. Two countries were compared explicitly, and in addition, we were able to implement a forecast for assorted countries based on some time series algorithms, considering long and short term seasonalization. Afterwards the key results were composed to some lucid dashboards provided in Tableau Online.

If the resulting prototype arouses the interest of some stakeholders, it would be quite easy to expand the data investigation and compose individual dashboards based on a given CI. As all used data is available online and updated frequently, the data processes could be migrated from csv files to a data-base solution with low effort.

2. Data Understanding and Preprocessing

We had 6 different internet sources each containing several csv-files to choose from.

Title	Content	Website
World	daily updated set of confirmed data, e.g. cases, deaths, regions	ourworldindata.org [1]
Europe	daily updated unharmonized sets of confirmed data, e.g. cases, regions (not all countries available)	github.com/covid19-eu-data [2]
USA	daily updated set of Johns-Hopkins-University for United States, e.g. cases, regions	github.com/CSSEGISandData [3]
Japan	daily updated set of confirmed data, e.g. cases, regions	kaggle.com/lisphilar [4]
Brazil	daily updated set of confirmed data, e.g. cases, regions	kaggle.com/unanimad [5]
India	daily updated set of confirmed data, e.g. cases, regions	kaggle.com/sudalairajkumar [6]

Table 1: Data Sets and Sources

Comparing the different data sets we found out that the different data files doesn´t quite fit together.
The granularity of location data was different. In *World* file the data is only on country level available, whereas the other data sets provide local administrative units (*lau)* or *nuts (in french*) with nuts_1, nuts_3 or even further, which makes a difference for the comparison of the contries, e.g. Japan provides “prefecture” and “regions”,Germany provides “federal states”, France provide departements. Also the data range about Covid19 infections and contamination differs, e.g. Germany provides the cumulative number of covid-19 cases, cases per 100 000 inhabitants and deaths per day, whereas France provides cumulative cases and no death rates. Some countries like Japan provide in addition information about free beds in hospitals and the number of tested and recovered patients.
In the final analysis we focussed only on new cases and cummulative cases per second level of location depth - nuts_1.

A suitable tool to illustrate the datasets and their columns is the so-called ER model, ER stands for Entity Relationship. Chart 1 shows the whole information set. This we used to find out, which file contains the necessary information and which data type is provided. That helped to make the merge of the files.

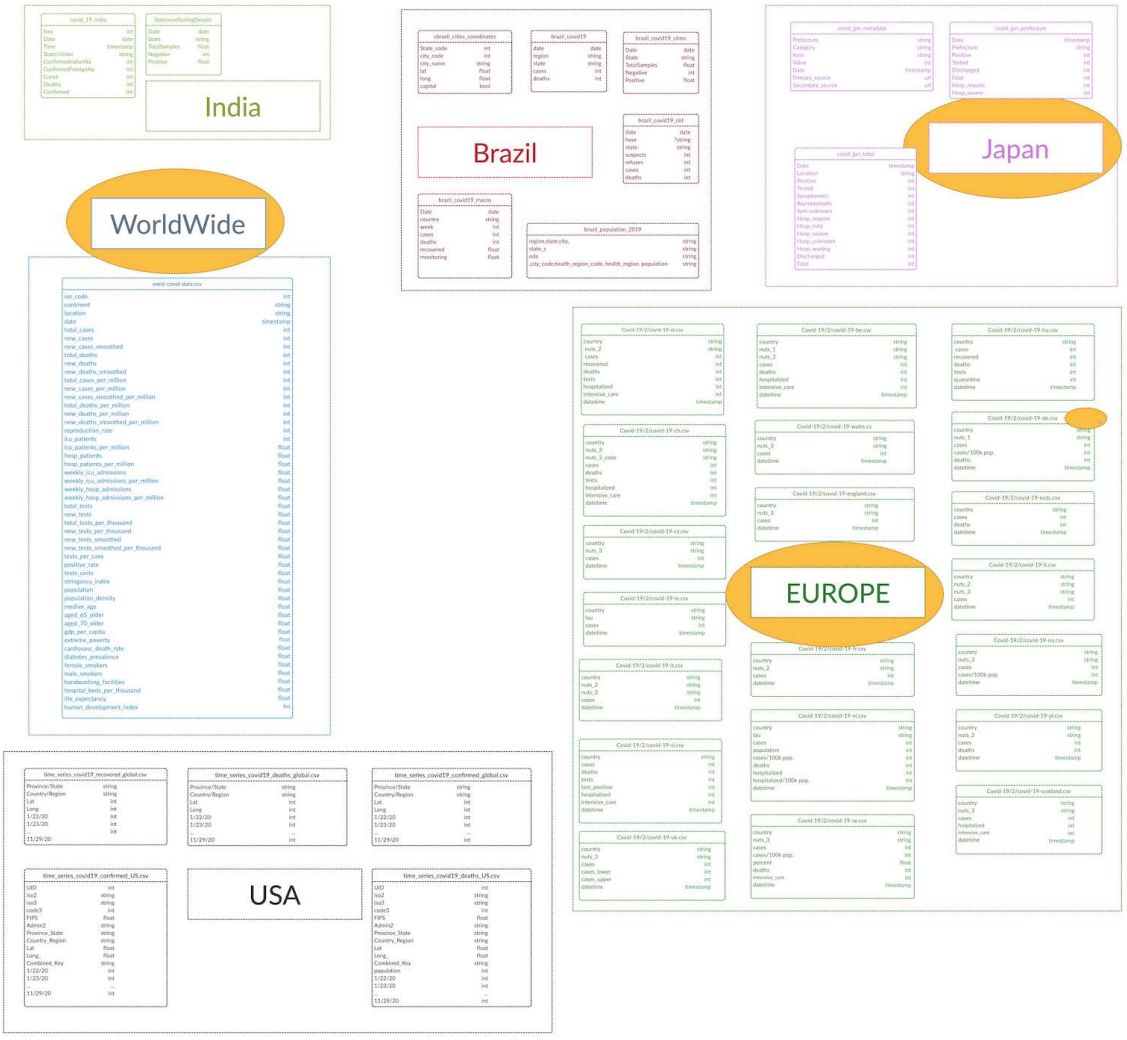

Figure 1: Complete ER Model

Within this report we focussed on World, Europe, and Japan datasets, which is marked in yellow. Chart 2 shows Germany with all its columns and data types as an extract of the ER model:

Figure 2: Extract of ER Model

To be able to compare countries we dropped several information columns due to the fact that only a limited number of countries provide such information. For some countries, like France, we calculated the daily changes out of the cumulative data. Finally, we joined several country sets into one file to enable direct comparison via data analytic techniques on more detailed level.
For comparison of all countries we took the dataset World, for regional comparisons we chose some datasets and joined them together.
We eliminated double information in the dataset World, e.g. sum of all countries, named world.

3. Data Analysis and first results

We collected the first occurrence of Covid19 for each country to see when the virus had affected the different countries. Figure 1 shows that at the beginning of 2020 only two areas are affected (Hong Kong and China), two weeks later additional three countries, whereas the next two weeks the virus entered 10 and 11 countries, respectively. After some weeks new countries joined. In the last week of February more than 30 additional countries started to deal with the virus. Five weeks later almost every country had its first infection case. Interestingly, in October some small Oceanian islands registered their first virus case.

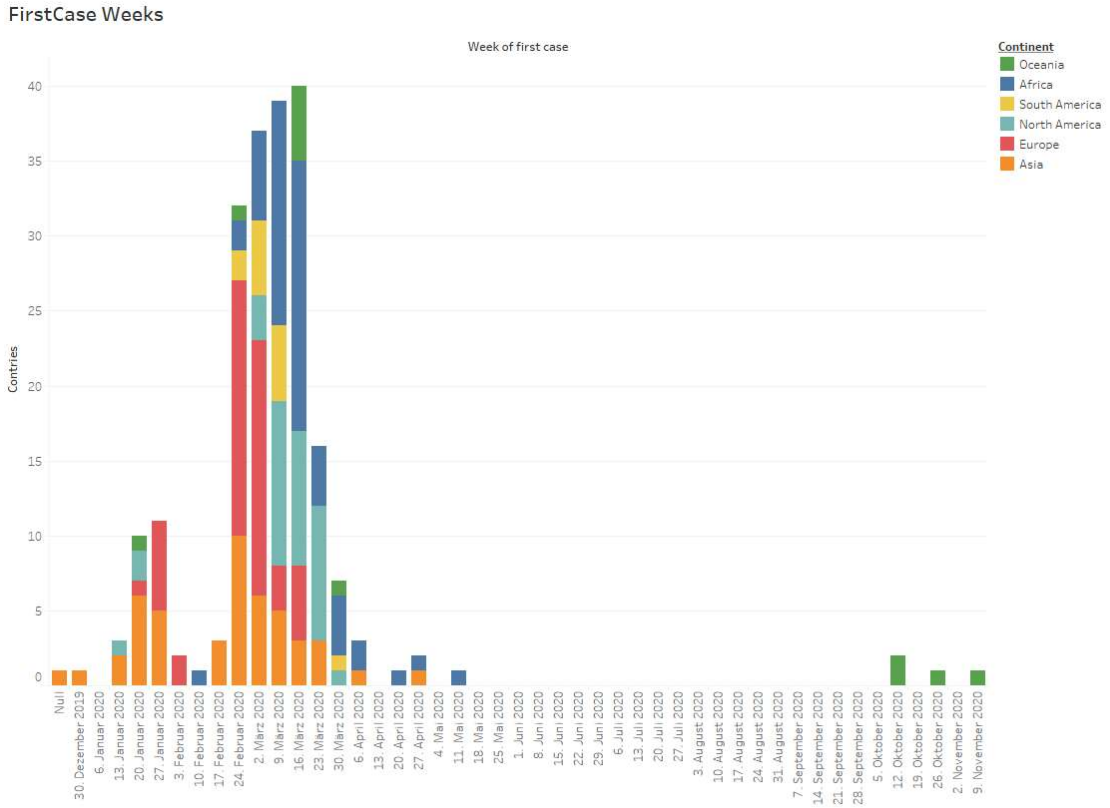

Figure 3: First Cases per Country

Figure 4 shows in more detail the wave of infection during the first quarter 2020. Each day additional countries registered cases of infection for the first time.

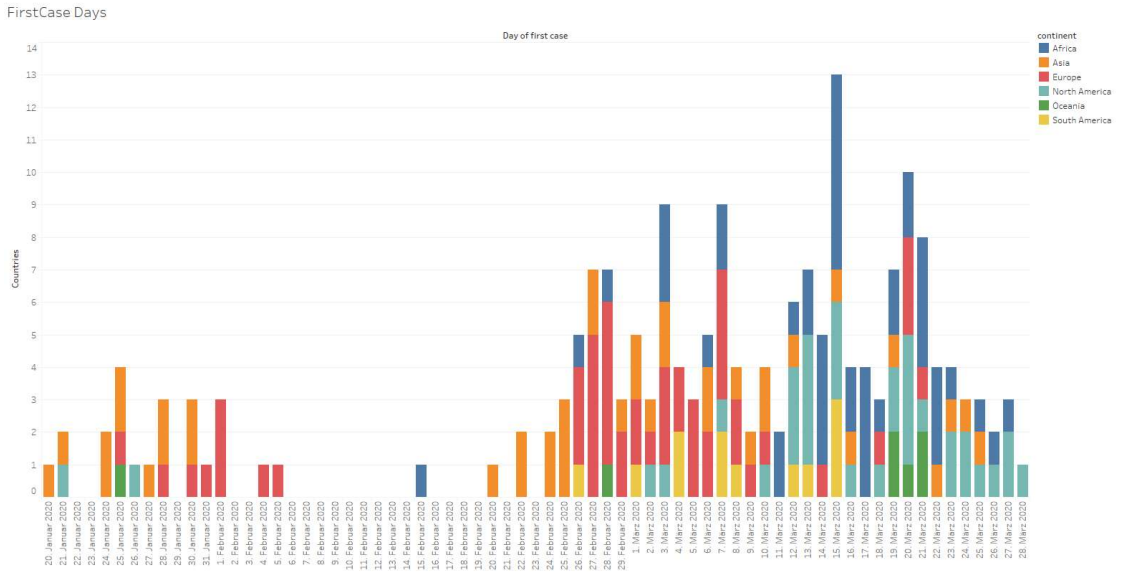

Figure 4: Number of Countries with first registered case of infection (Q1 2020)

Figure 5 shows that the development of new cases per day differs for the countries in several ways. We see different distances between the maxima comparing two countries (e.g. US - Italy), the number of maxima is different (US, India), the difference of the heights of the maxima differs (US, France)

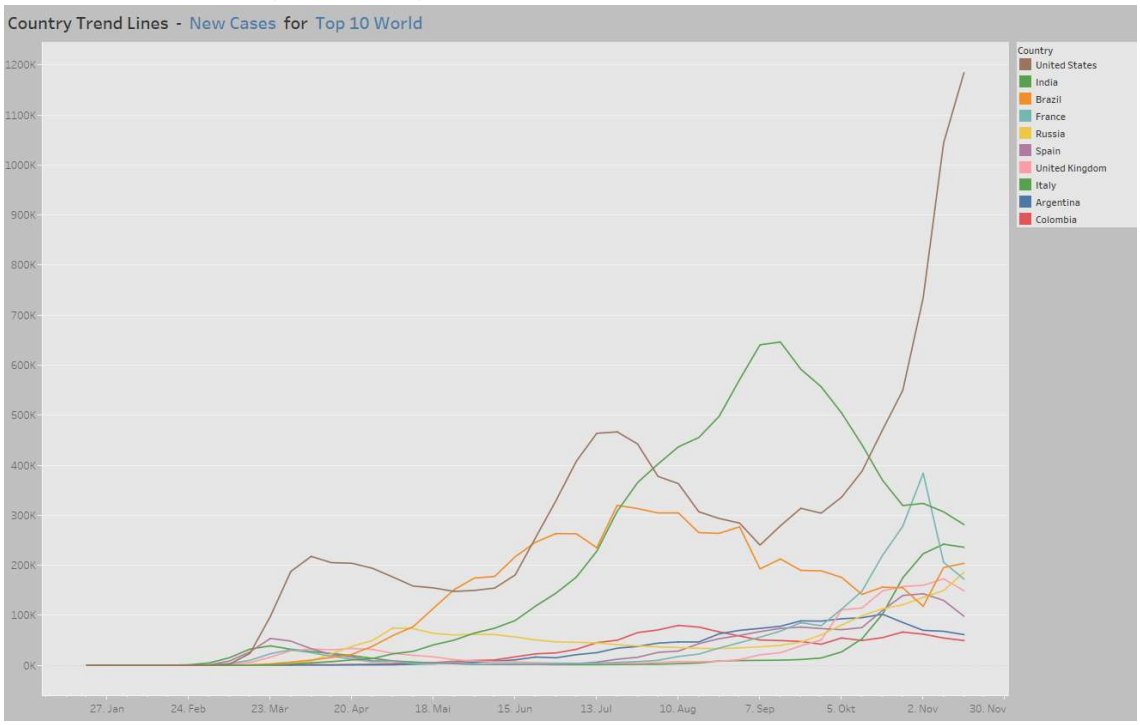

Figure 5: Development new cases per country

A different comparison was done with Germany and Japan, since these countries have a similar demographic curve and comparable economic strength. But the results showed different pictures using the same scale for number of infections per region/federal state.

A second look on Japan with a higher granularity scale shows similarities in the spread of virus in areas of higher population.

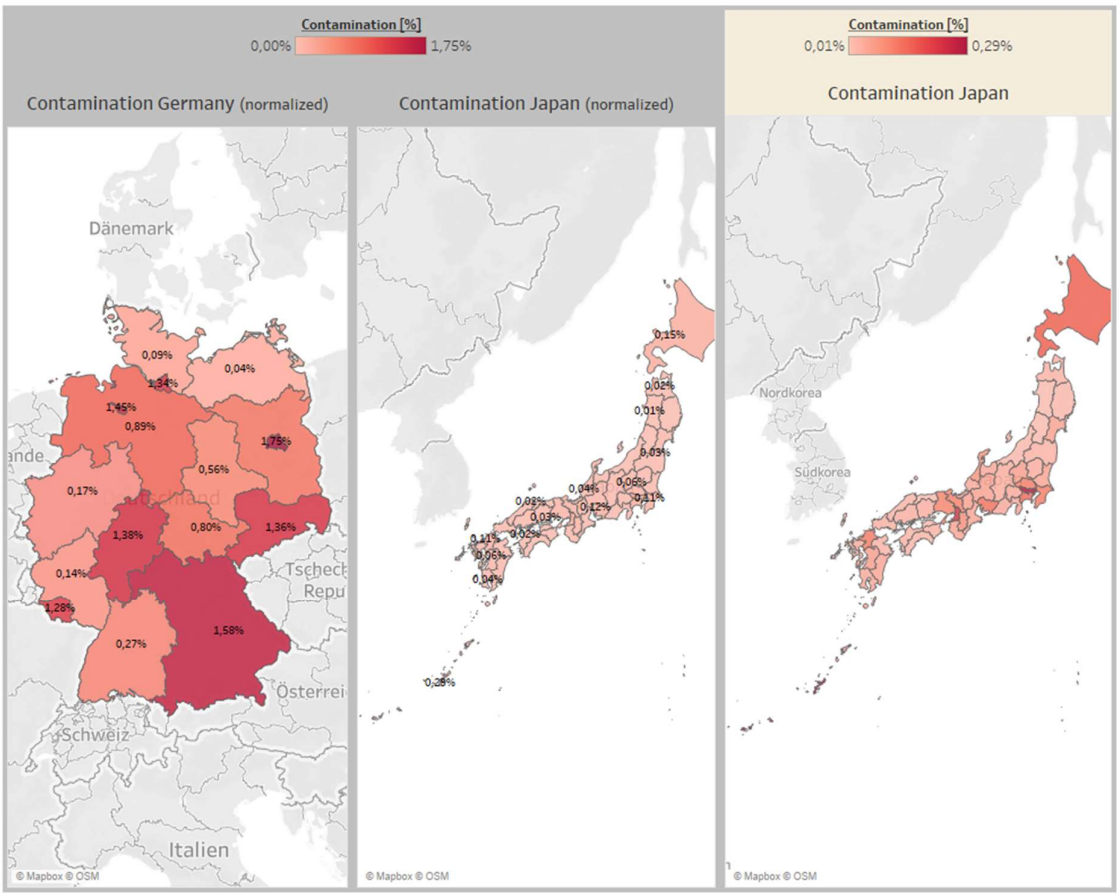

Figure 6: Comparison Germany and Japan

4. Prediction of future development

Within our project we have many datasets providing the number of new cases of Covid19 for the different countries within the world. All datasets do have one row per day, the beginning of counting the cases differs from country to country depending on when the first occurrence was detected.
Predicting the development of cases in each of the countries is of high interest. As all countries do have different developments within their cases each prediction has to be made for each country. We could choose different models for each country or we choose a model which allows to adjust the model parameter according to the different behaviour.
The model should be able to handle increasing and decreasing of new cases in two different ways.
We chose the *SARIMA* model (Seasonal Auto Regressive Integrated Moving Average), which can handle seasonal effects. The *SARIMA* model will analyze by statistical methods the trends and the seasonal effects a pandemic like Covid19 consists of. *SARIMA* consists of two parts, the first, *ARIMA*, tries to figure out how a stationary time series can be described by functions and parameters, the later integrates the seasonal effect.

Usage of *SARIMA* to predict development of a time series:

1) Define Model
    a) in total 7 parameters to be chosen
    b) choose error function to evaluate the model accuracy
    c) per grid search calculate error for different combinations of parameters
    d) select parameter set which gives the lowest error
2) Fit Model
    a) run the model with the chosen parameters with the training data, the actual new cases
    b) returns a set of figures which can be compared with the actual data
3) Predict future development
    a) run the model with the number of time steps desired.
    b) returns the predictions

The outcome of the model is a set of parameters for each country investigated as well as a prediction of the development of the new cases within the next days.

country	p	q	d	P	Q	D	s	actual: new cases 29.11.2020	predicted: new cases 7.12.2020
Germany	6	2	1	1	1	2	7	14611	8275
USA	0	2	1	1	1	4	24	154893	228418
Sweden	5	3	1	1	0	2	25	5464	6114

Table 2: Parameter set and prediction for some countries

The model will predict how the actual time series will proceed. The Figure 7 shows the development of new cases within Germany:

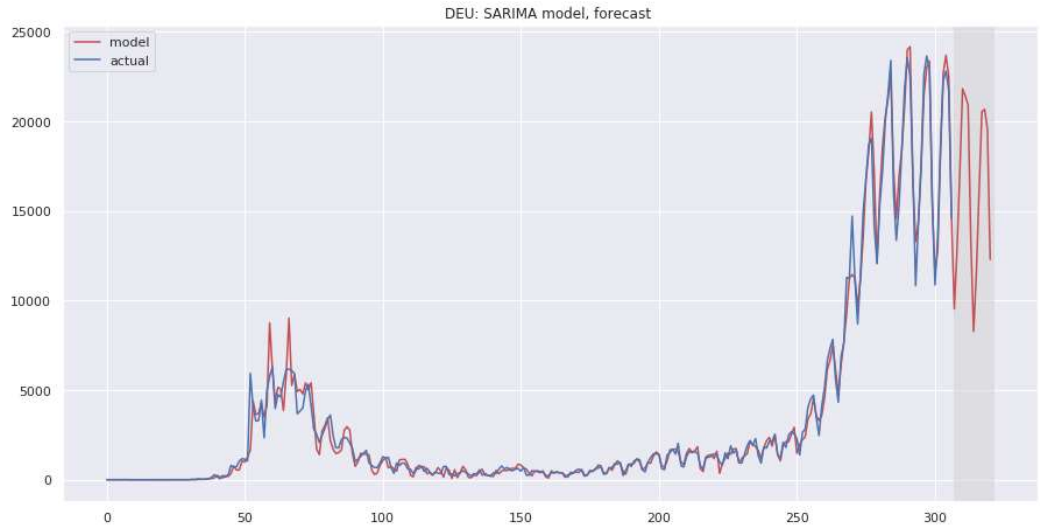

Figure 7: Development and 14 days prediction for Germany

The prediction (red line) follows the 7-days motion (blue line) within the numbers as well as the declining after reaching a high point some weeks before. This gives us some hope that the lockdown started Nov 1st impacted the new cases development strongly.

For Sweden, as shown in Figure 8, the decreasing new cases is not yet predicted. Sweden didn't go for a lockdown in November, so the pandemic will go on and the infections will increase at least within the next two weeks.

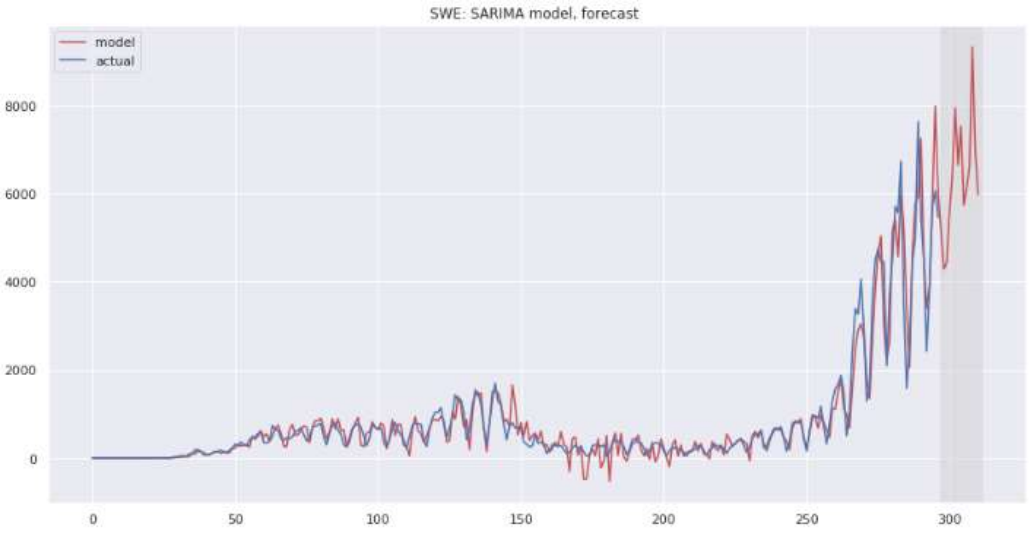

Figure 8: Development and 14 days prediction for Sweden

5. Web based Dashboard

The visualization of the data has been conducted by using the dashboard software *Tableau*. With *Tableau* we composed several dashboards composed of graphs and figures which allow the user to select the interesting countries, relevant metrics and specific time stamps to start a deep analysis.

Figure 9 shows the generated dashboard showing development of infection across the world:

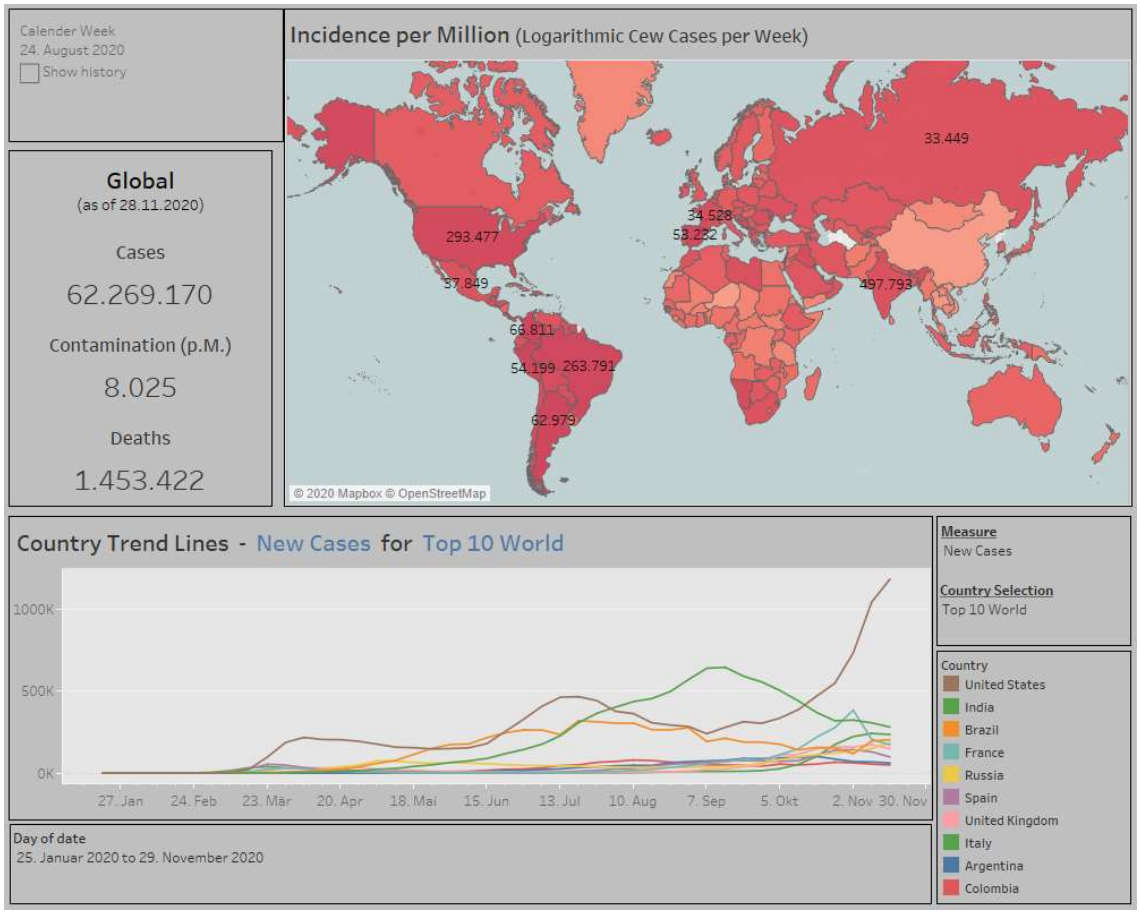

Figure 9: Dashboard Entire World

The four provided dashboards only represent a subset of possible tools, that could be provided to possible stakeholders.

All graphical elements can easily be adapted to the stakeholder's predefined CI.

For the development of this prototype we decided to use static, local CSV-files as data sources. If the dashboard would be needed on a regular basis, it can easily be connected to a data base, which would be updated by scheduling the data preprocessing jobs.

6. Sources

[1] https://ourworldindata.org/coronavirus-source-data
[2] https://github.com/covid19-eu-zh/covid19-eu-data/tree/master/dataset
[3] https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_1 9_time_series
[4] https://www.kaggle.com/lisphilar/covid19-dataset-in-japan?select=covid_jpn_total.csv
[5] https://www.kaggle.com/unanimad/corona-virus-brazil?select=brazil_covid19.csv
[6] https://www.kaggle.com/sudalairajkumar/covid19-in-india?select=covid_19_india.csv

for dashboard measures:
https://www.bundesregierung.de/breg-de/aktuelles/bund-laender-beschluss-1811744
https://www.bundesgesundheitsministerium.de/coronavirus/chronik-coronavirus.html