A Web Data Dashboard   
for   
Covid-19 Analysis

Certification Project Data Scientist



by P. Eckert, Julia Poliak, Christoph Seng

4.12.2020 Bonn and Darmstadt

Agenda

[1. Data Understanding and Preprocessing 3](#_Toc57980433)

[2. Data Analysis and first results 5](#_Toc57980434)

[3. Prediction of future development 7](#_Toc57980435)

[4. Web based Dashboard 9](#_Toc57980436)

[5. Conclusion 10](#_Toc57980437)

[6. Sources 11](#_Toc57980438)

# Data Understanding and Preprocessing

This project investigates the data of infections of Covid19 across the whole world, its development over time and regions. We received data from different data sources and preprocessed it to make comparisons, deeper data analytics, predictions to the near future. Finally, we built a dashboard to provide access to anybody for making its own research.

We used data from 6 different internet sources each containing several csv-files (table 1):

|  |  |  |
| --- | --- | --- |
| **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 John-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 not all sets contain the same granularity of location depth. In World data set only data on country level available, whereas the other data sets provide local administrative units (lau) on different levels, eg. Japan provides “prefecture” and “regions”, Germany provides “federal states”. The data about Covid19 infections differs, e.g. Germany provides the number of new cases 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.

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.

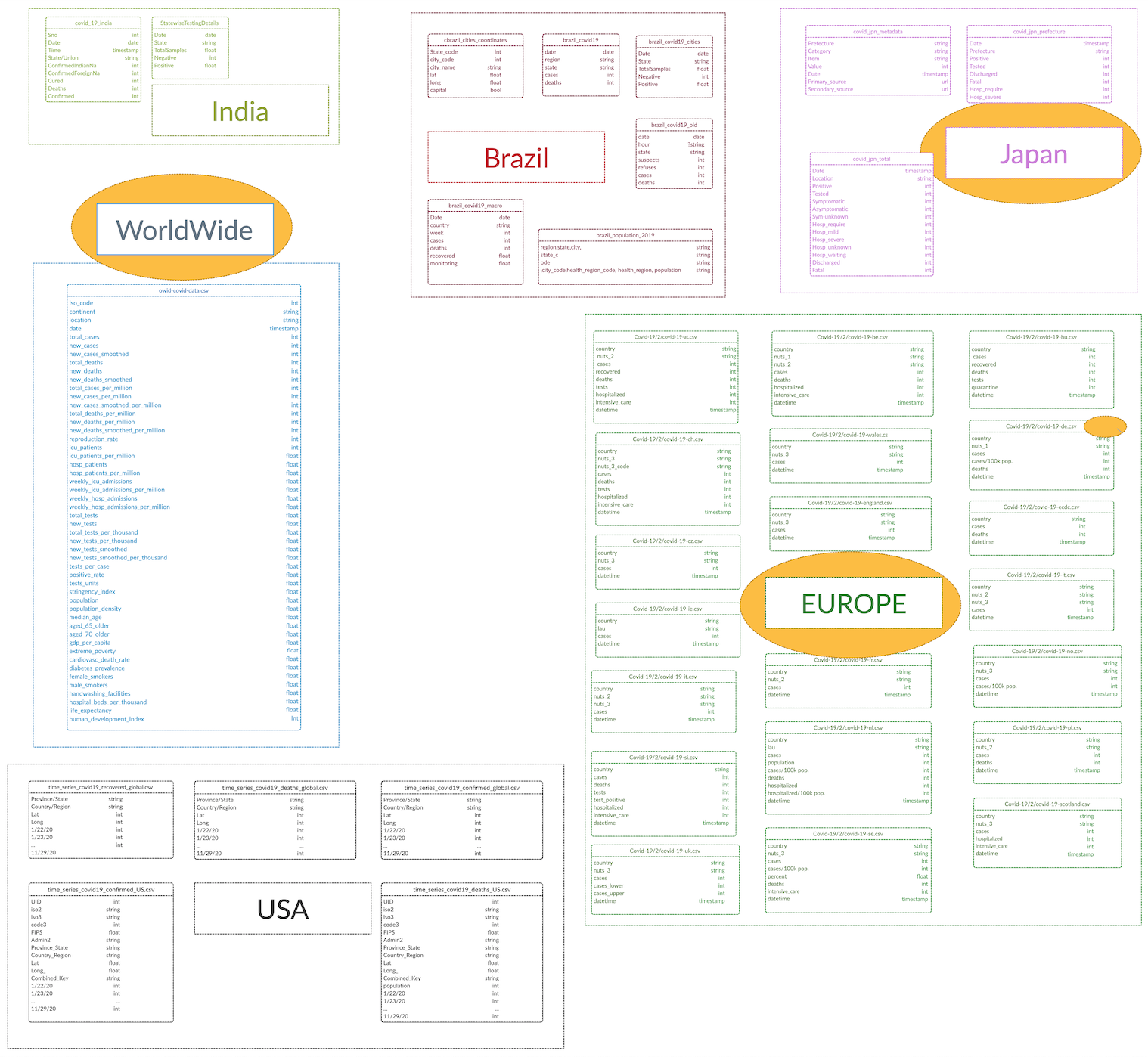


Figure 1: Complete ER Model

Within this report we focussed on World, Europe, and Japan datasets. The chart 2 shows as extract of the whole ER model Germany with all its columns:

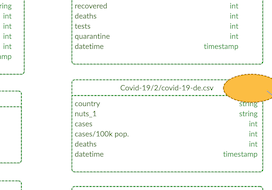


Figure 2: Extract of ER Model

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 countries sets into one file to enable direct comparisons via data analytic techniques.

For comparisons using 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.

# 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 of limited new infected countries in the last week of February more than 30 additional countries started to deal with the virus. Five weeks later almost all countries had their first infection cases. Interestingly, in October some small Oceanian islands registered their first virus cases.

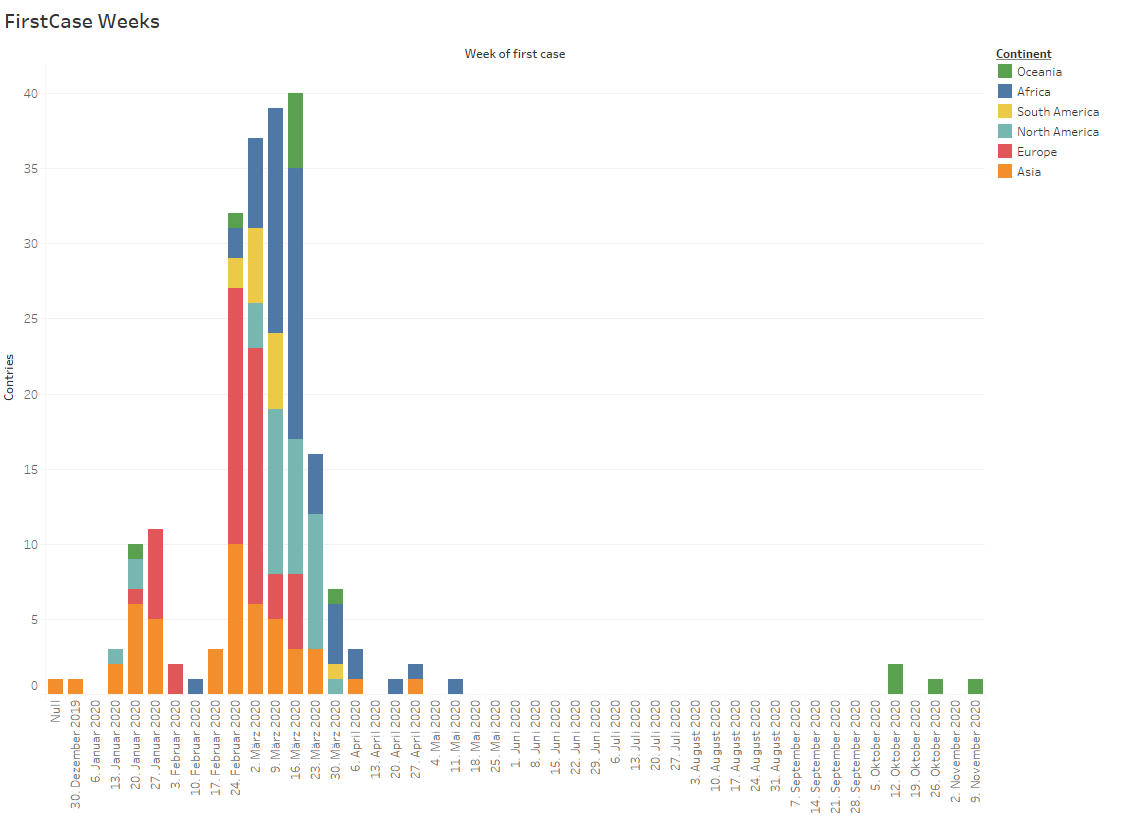


Figure 3: First Cases per Country

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.

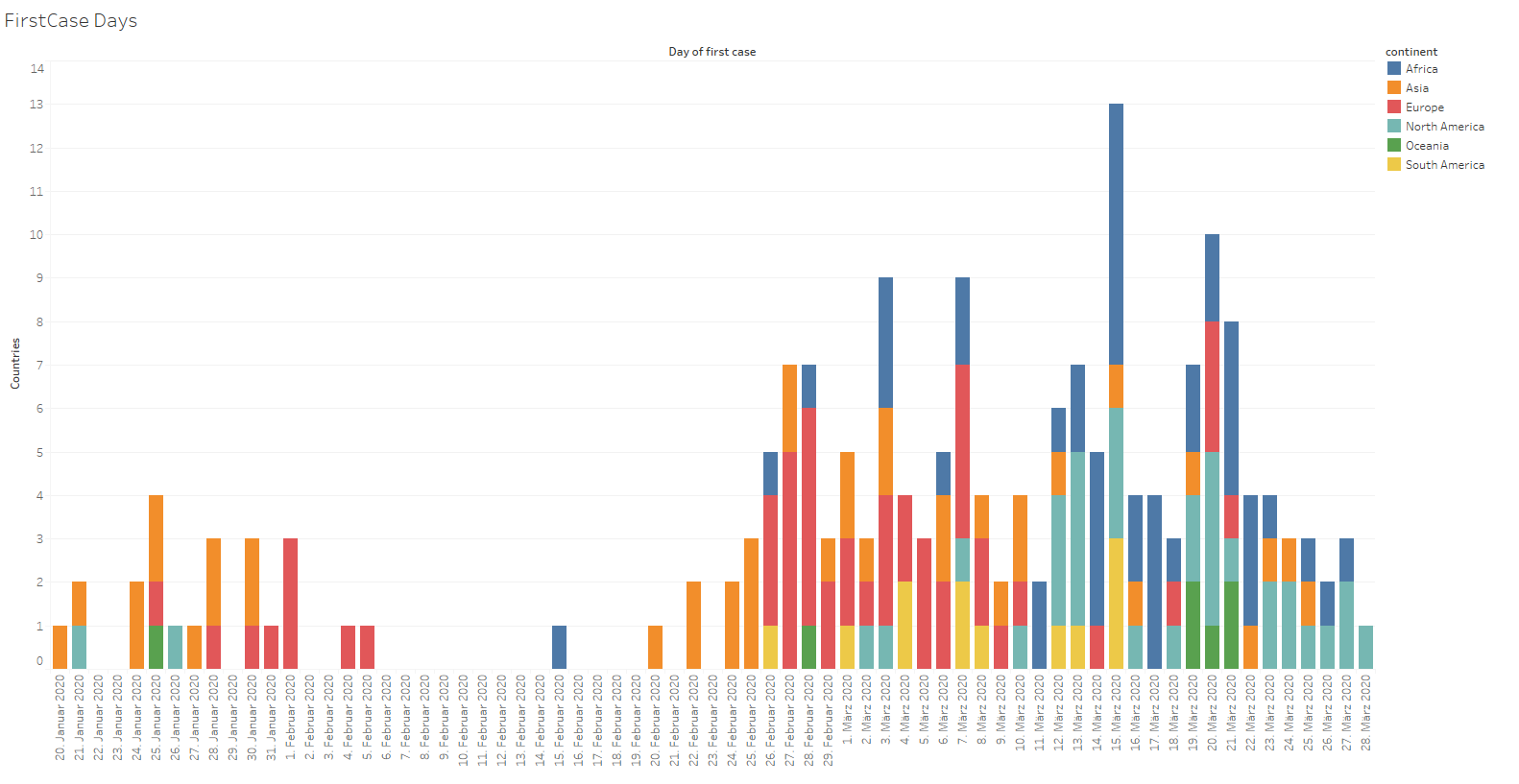


Figure 4: Number of Countries with first registered case of infection (Q1 2020)

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

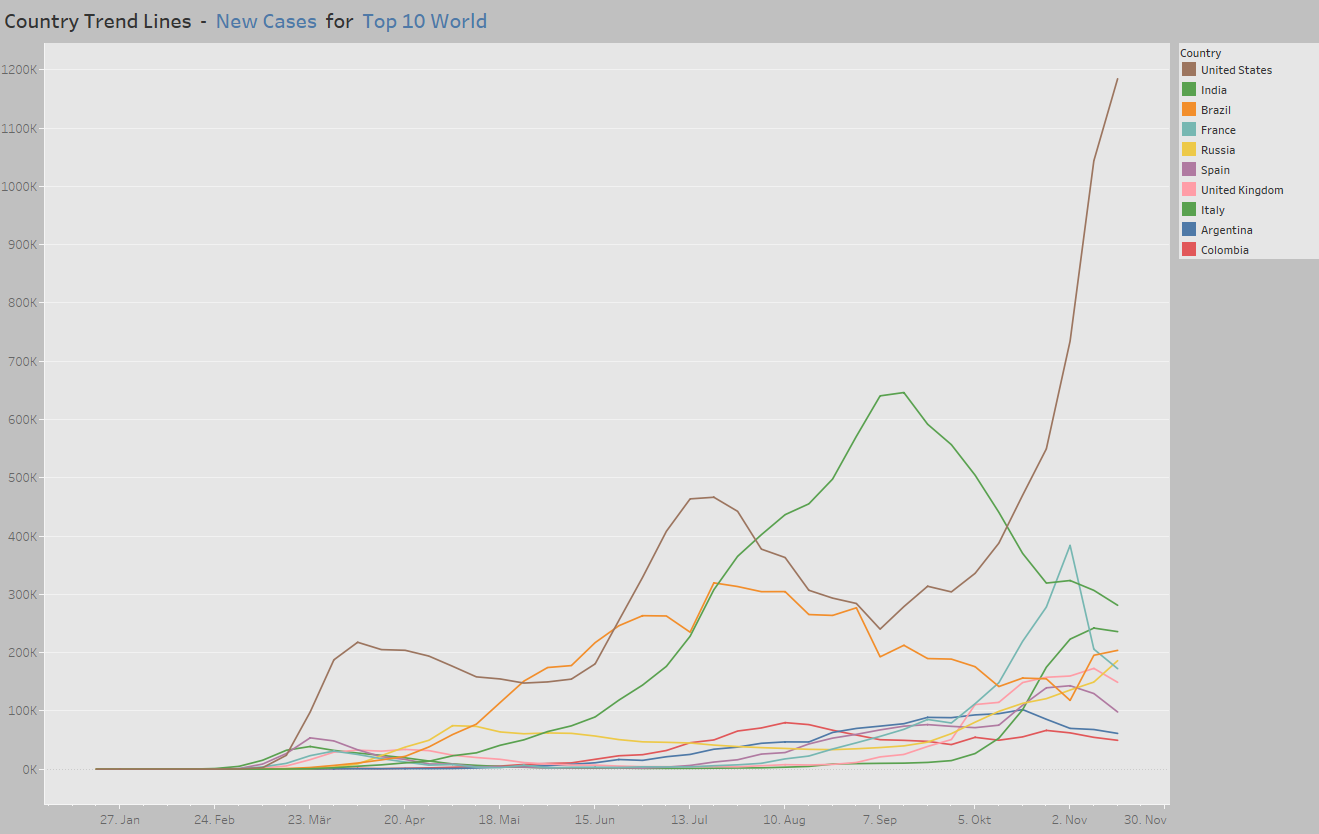


Figure 5: Development new cases per country

# 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
   1. in total 7 parameters to be chosen
   2. choose error function to evaluate the model accuracy
   3. per grid search calculate error for different combinations of parameters
   4. select parameter set which gives the lowest error
2. Fit Model
   1. run the model with the chosen parameters with the training data, the actual new cases
   2. returns a set of figures which can be compared with the actual data
3. Predict future development
   1. run the model with the number of time steps desired.
   2. 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 graph 4 shows the development of new cases within Germany:

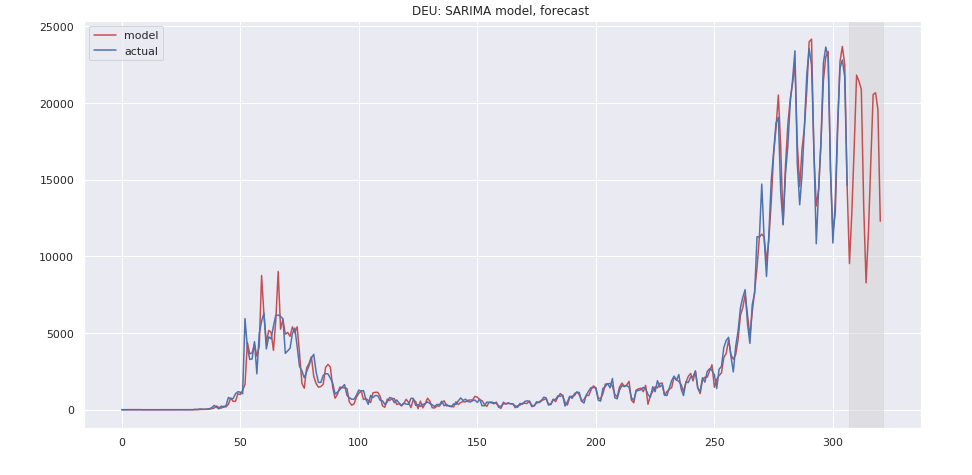


Figure 6: Development and 14 days prediction for Germany

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 7, 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.

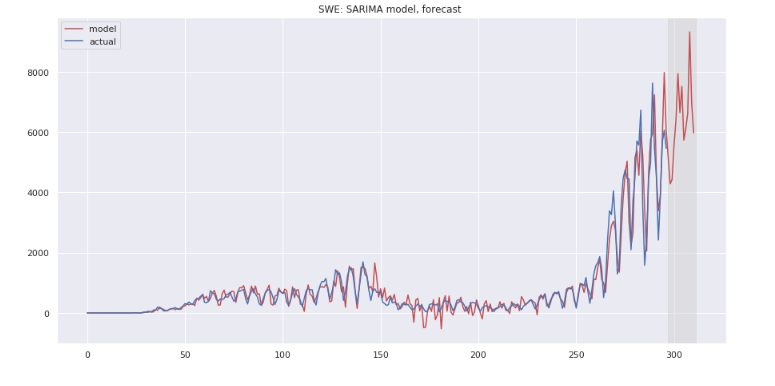


Figure 7: Development and 14 days prediction for Sweden

# 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 8 shows the generated dashboard showing development of infection across the world:

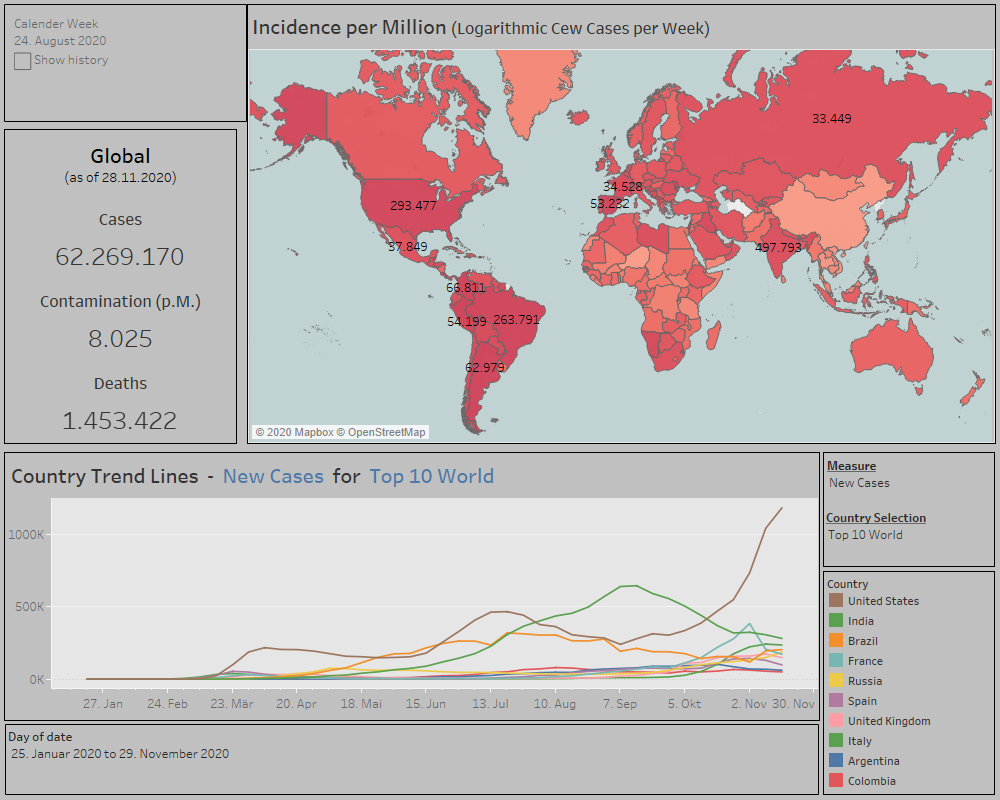


Figure 8: Dashboard Entire World

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.

# Conclusion

# 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