Tria COVID-19 Datathlon

Seyyid Emre Sofuoglu

04/19/2020

COVID-19 Analysis and Simulations

Background and Motivation

For the last 4 months the pandemic has been roaming the earth, putting us to our places (literally). The whole world is affected by it and by far, we have no solutions but running and hiding. Still, a disaster is exactly defined like this and this is an experience to learn from.

So far, there have been many analyses on the spread and biology on COVID-19 by people who are the best of their fields. By no means, I aim to produce something more valuable than those studies but I aim to explore my own country, Turkiye, for which there are not many studies and not many useful data around. I will utilize the existing data for other countries in the world and several models, e.g. SIR, for this.

I will use publicly available COVID-19 Dataset provided kindly by European Centre for Disease Prevention and Control (https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide). In this dataset, we have daily cases and number of deaths. I will try to make a model that will be applicable on different countries with varying circumstances.

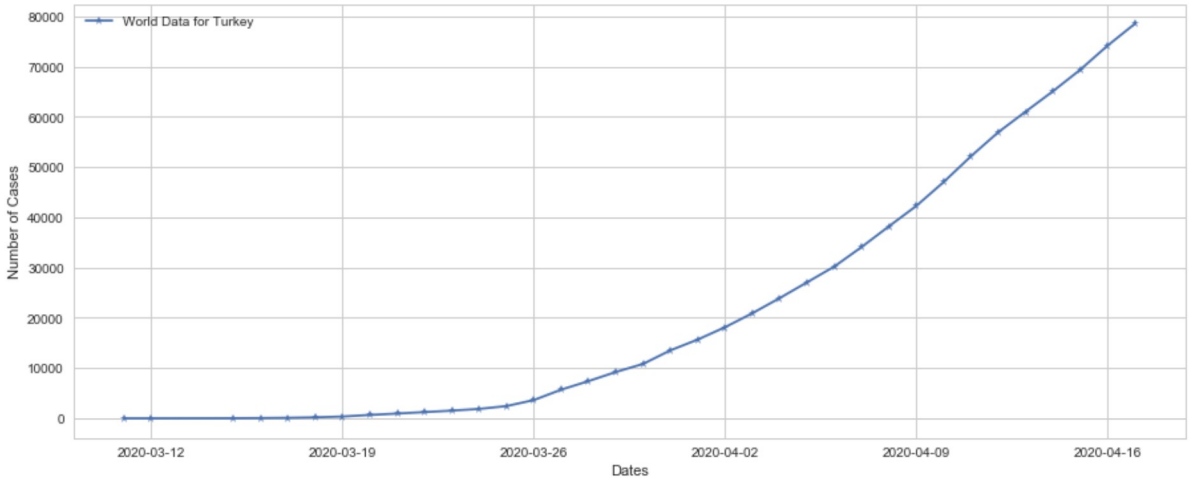
Data

The first data we have is the worldwide data for many countries. Let's have a look.

Out[2]:

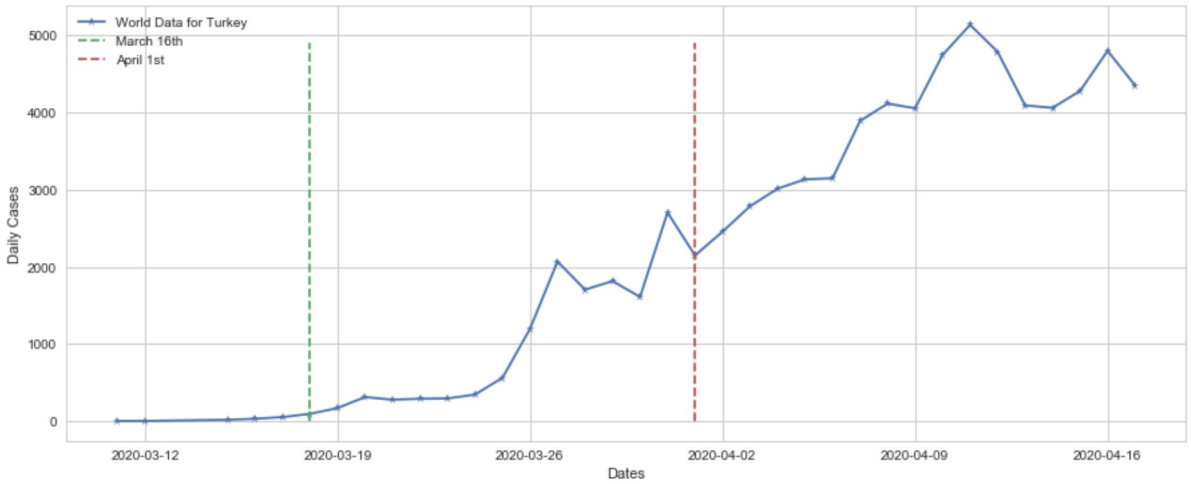
	dateRep	day	month	year	cases	deaths	countriesAndTerritories	geoId	countryterritoryCode	popData2(
0	4/18/2020	18	4	2020	51	1	Afghanistan	AF	AFG	3717238
1	4/17/2020	17	4	2020	10	4	Afghanistan	AF	AFG	3717238
2	4/16/2020	16	4	2020	70	2	Afghanistan	AF	AFG	3717238
3	4/15/2020	15	4	2020	49	2	Afghanistan	AF	AFG	3717238
4	4/14/2020	14	4	2020	58	3	Afghanistan	AF	AFG	3717238

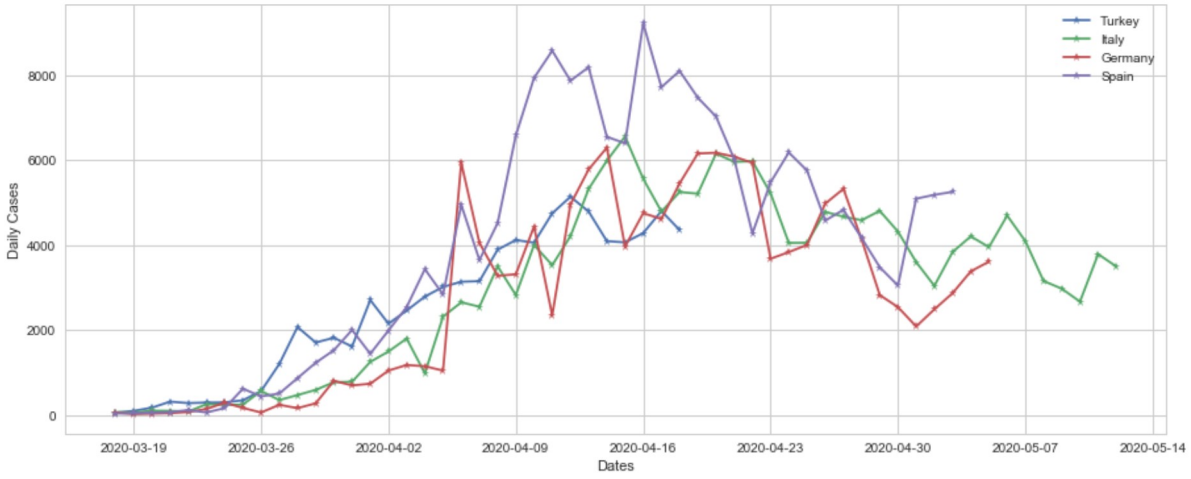
Now, let's have a look at the data on Turkiye.



It seems that a set of measures by government (Cancelling flights from Europe, Iran, South Kore etc., contact tracing, lots of tests, cancelling schools and universities at March 16th) had a big effect on slowing the pandemic at April 1st. Adding the influences of Italy, Iran and South Korea, these measures probably also helped more people decide on social isolation.

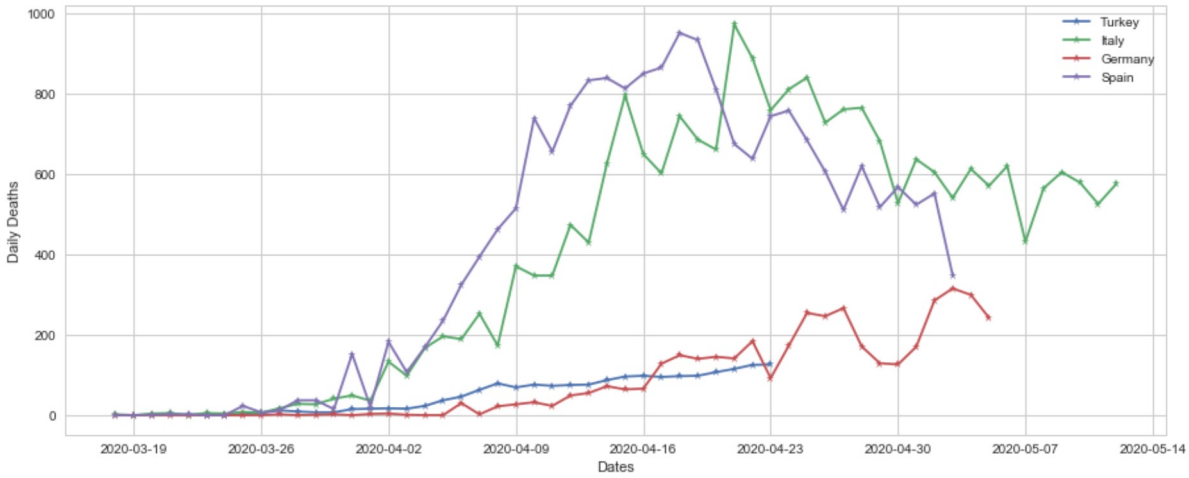
Now, let's see some comparisons between countries taking 100 total case counts as basis, or first day of the spread.





We can see that there are similar trends of them all.Spain has a higher peak and Germany has some irregularities. But this is weird, we know that Germany has much less number of deaths than Spain or Italy, so what gives?

Let's see the daily deaths comparisons.



We can see that Germany and Turkey now has much better picture compared to Italy and Spain. What might be the cause? What helped these countries achieve such low figures?

Now let's analyze the data using models like SIR.

Methodology

SIR (https://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology#The_SIR_model) is a compartmental model well known in epidemiology. Using this model, we will try to analyze various countries and try to see parameter variations among them.

In SIR models:

- **S** is very often used for "susceptible" to refer to people who can get a disease (some people may have some immunity).
- **I** refers to "infected", which means they have the disease; in many cases this is called **L** so that **I** can be used to refer to those that are infectious (i.e. can spread the disease).
- **R** refers to "recovered" or "removed".
- Another compartment we can use for our purposes is "deceased" **D**. The pandemic has a high death toll, so this is reasonable.

The model is governed by a set of differential equations which explain the transition between the compartments:

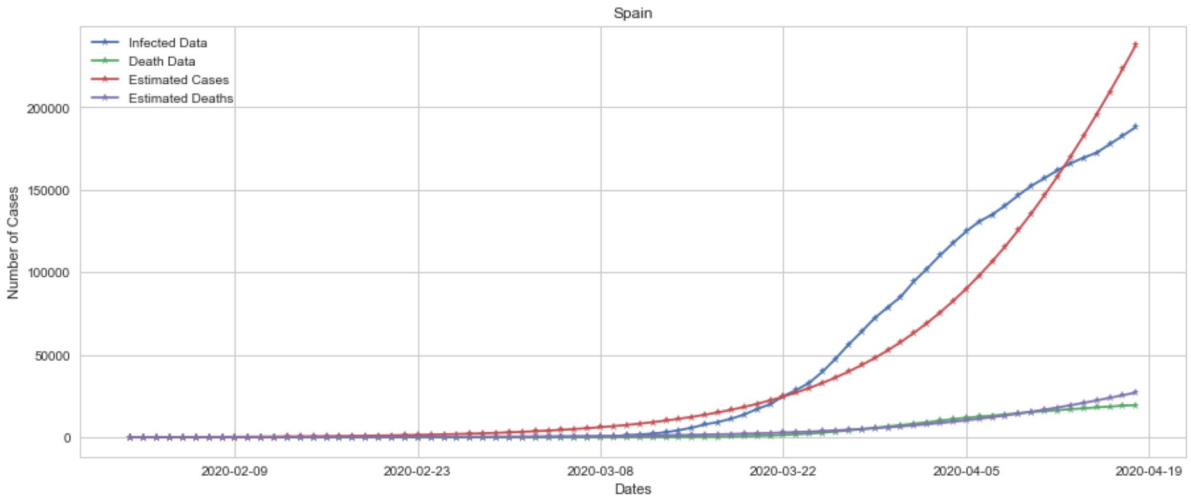
$$\frac{dS}{dt} = -\frac{\beta I S}{N},$$
$$\frac{dI}{dt} = \frac{\beta I S}{N} - \gamma I - \mu I,$$
$$\frac{dR}{dt} = \gamma I,$$
$$\frac{dD}{dt} = \mu I,$$

where N corresponds to the total population. We know that with increaseing number of cases, hospitals run out of ICUs and death rate μ increases. But let's assume that it doesn't and continue with our analysis.

The model is simplistic. It does not consider changing public dynamics and government measures. It also assumes we have knowledge of all infected. But we will start by this to get some ideas, then improve. Let's define an update function.

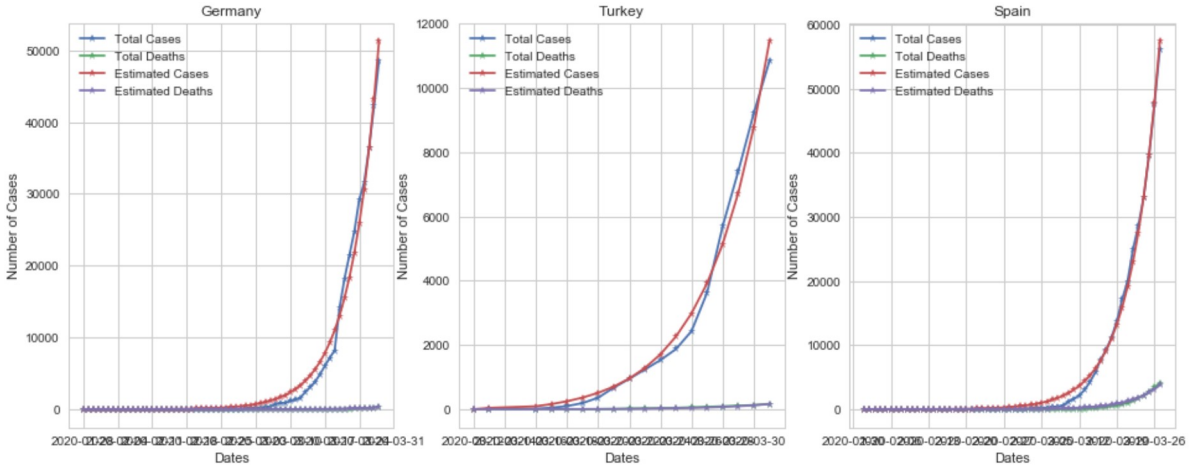
Now, the population of Turkey is about 85 millions, Germany is about 84 millions, Italy is about 60 million and Spain, 50. Hubei, in which a big majority of China's cases happened also is about 59 million.

Let's fit the parameters on a country, e.g. Spain, and see how it looks.



The fitted curve, while looking similar, is not very accurate for Spain. It is unable to model the curvature change. Cases change the curvature at some point, indicating the reaction of the people and government, so the model is undercomplex. Also, we know that the infected might not get symptoms for a while and do not get cured or die in a day, so there should be some delays in changes.

Let's see how good is the model until the effects of the measures are visible, a.k.a. 14 days after measures took effect. Let's also check new countries such as Turkey and Germany as their populations have similar size and their situation also look similar right now.



It seems the models fit the early stages very well, this brings the question:

```
How well will the model work if I assume there are two stages and fit my model accord
ingly?
```

Also, until now we only used visual interpretations but from now on, we will also see an estimation error e defined by:

$$e = \frac{\|y-\hat{y}\|_2}{\|y\|_2},$$

where y is the real number of cases as a time series vector, $\hat{y}$ is the vector, or array estimated by the model and $\|.\|_2$ is the l_2 norm, defined by:

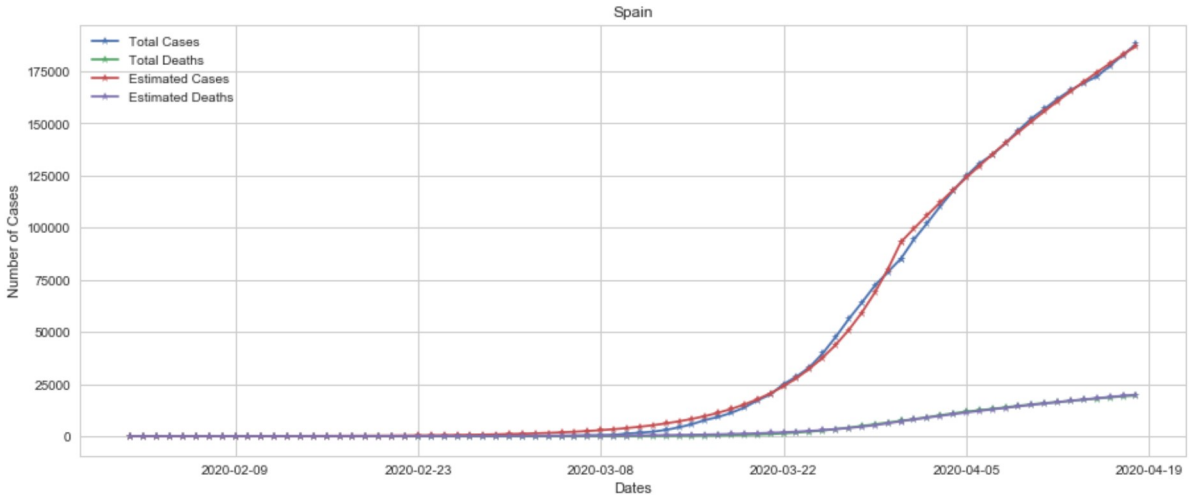
$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Let's check the error values of the above.

```
Error for Turkey is: 0.0797
Error for Germany is: 0.0967
Error for Spain is: 0.2066
Error for Spain when only cases until measures took effect is: 0.0502
```

As expected, the model parameters get effected by the measures.

Now, let's turn back to the model with two stages. We will need prediction later on, so let's add the capability.



Out[18]: 0.028858783532019202

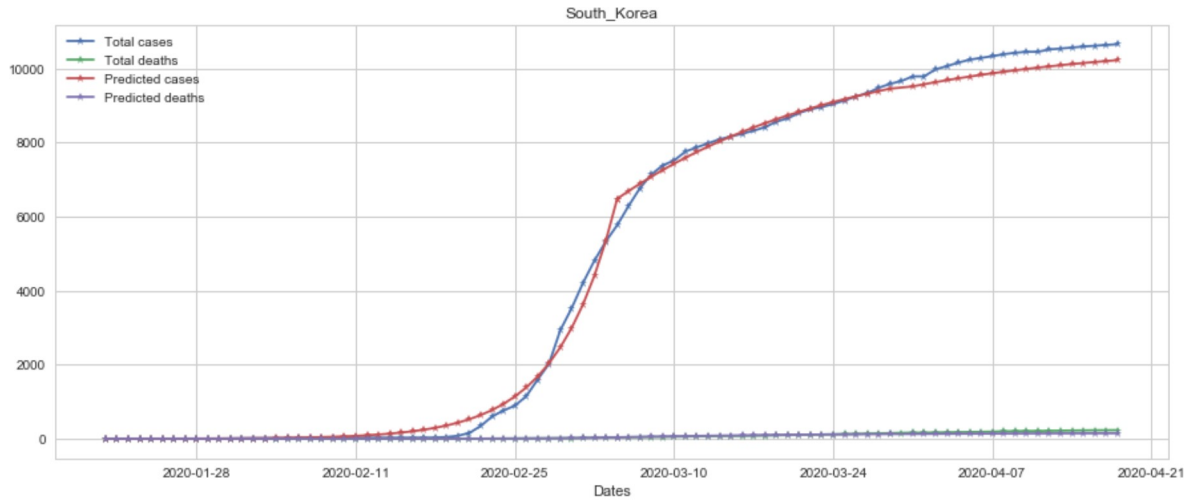
Now, this looks much better. Let us see the changes to the parameters.

```
[1.85322978 1.55940173 0.14219649 0.87074295 0.77823999 0.11806072]
```

We can see that spread rate dropped a lot. Weirdly enough, recovery rate fell down. This might be due to the lag in the deaths being much longer (about 5-10 days) than what the model assumes (1 day). Another thing is that we unfortunately do not have that much of a speedy recovery, these numbers are inflated to fit the curve as we didn't request the function to also fit the recovery data. So the R part of SIRD is not fitted to anything.

Let's leave these model issues to later and see how these parameters do in prediction.

In South Korea, there was a huge controversy regarding Shincheonji Church (https://en.wikipedia.org/wiki/Shincheonji_Church_of_Jesus) at around 20th February. Afterwards, although there were no lockdowns, people self isolated, as this was also strongly recommended by the government. So we can say that the spread rate might have changed at around 6th March.

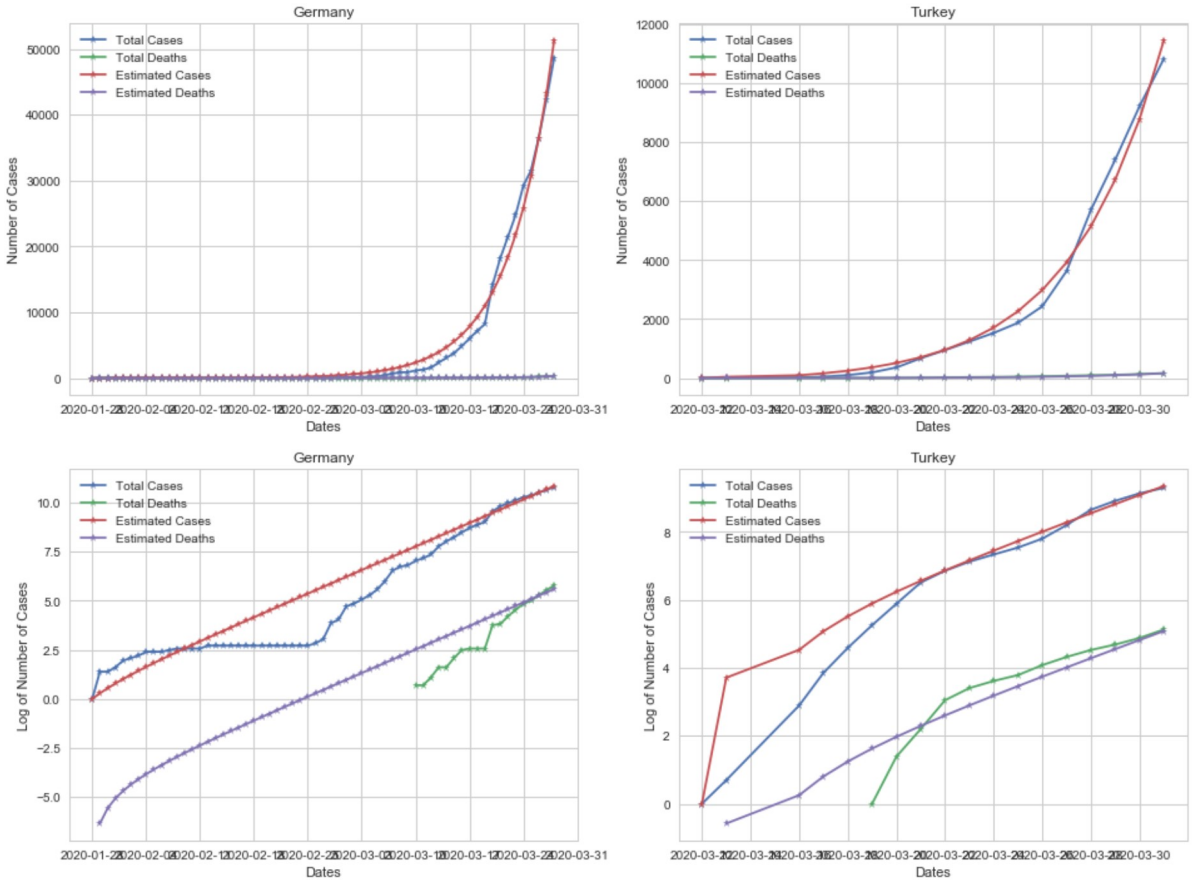


Out[21]: 0.040487744994050216

This is a pretty good estimate for the last 20 days. Now let's return to our objective, what will happen in Germany, Italy, Spain and, finally, Turkey.

Results

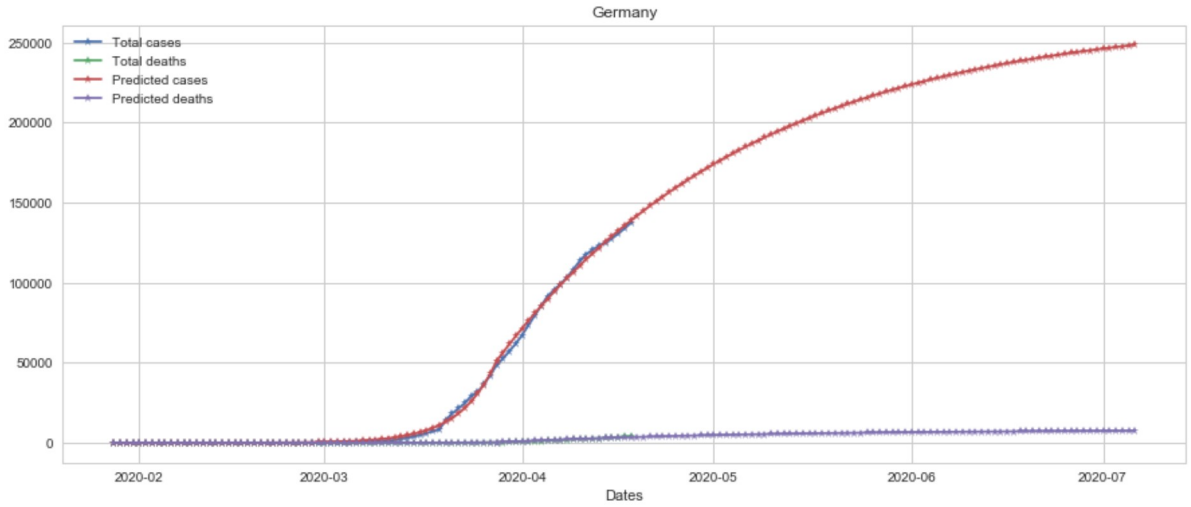
Let's see how SIRD model fits the data for Turkey and Germany.



The errors are not similar even though the graphs for total cases are similar for these countries. One possible explanation is that the number of cases increase very quickly after the first case in Turkey. This is due to excessive contact tracing. A clearer contrast can be seen in the figures with logarithmic scale. The number of cases stay frozen in Germany for a while which is unlikely considering how fast the pandemic spreads. Turkey, with the advantage of a late start, was more prepared and people were more alert.

Now, let us see if we the model is useful when there are not many measures, i.e. early stages. In Germany, the social distancing measures took effect starting March 13 with closure of schools, etc. One day later, borders were closed and a week later curfew was imposed. There is no curfew for Turkey but the social distancing measures took effect aat March 16th.

Now, let us try to predict the future values for several countries.

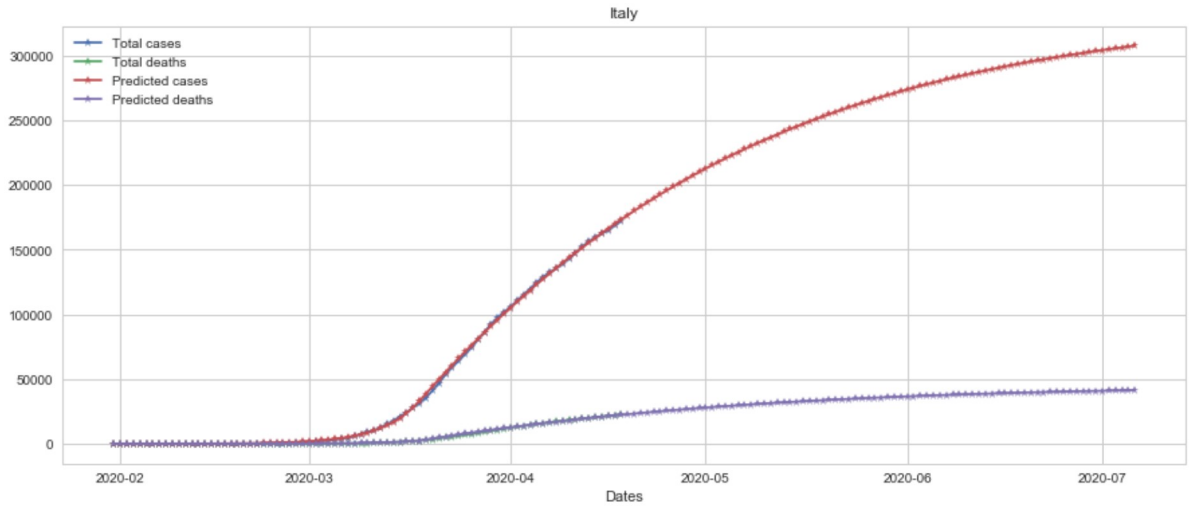



```
[0.30334233 0.13051881 0.0015962  0.18472814 0.20245199 0.00675232]
```

Out[24]: 0.03047678767098857

We can state that although case number is projected to increase, deaths do not increase that much. Also it looks like we are probably still at an early phase to model the second part. Still, what we see after the measures is not similar to China, so there might be a need for more precautions. The model looks matching to the current cases but a chnage in societal dynamics, as shown in earlier examples is highly affecting the parameters.

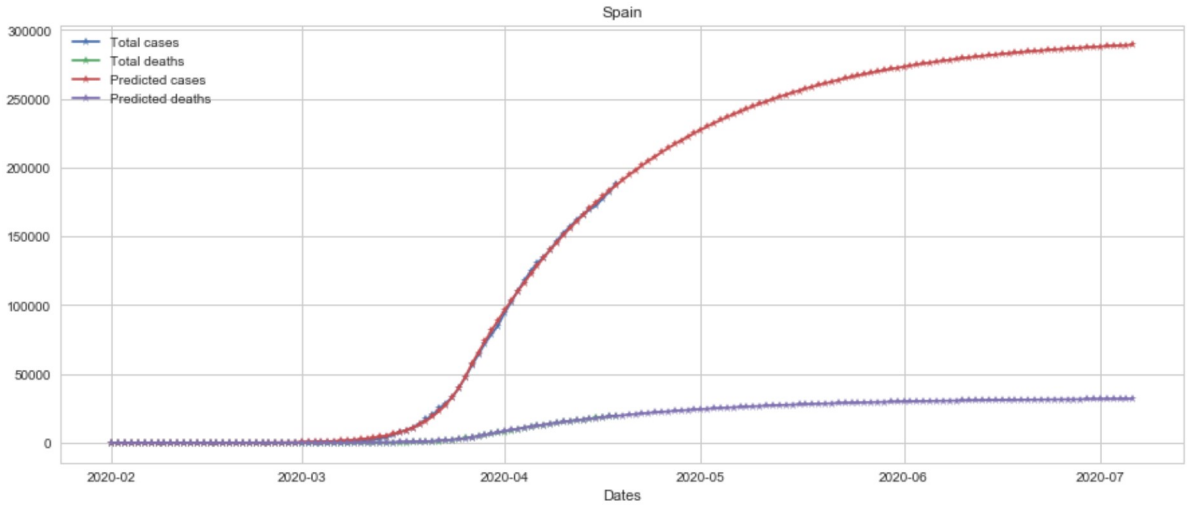
Now, let's see Italy.




```
[1.96949976 1.65711803 0.143077   2.01280662 1.74064573 0.28486868]
```

Out[26]: 0.013417838126834182

That's a bleak picture, which is hopefully wrong. We know that number of new cases is dropping but if the model is correct, we will see 250 thousands. We also need to keep in mind that the model thinks the whole population of Italy is connected, although we know right now that the regions are separated with travel restrictions. This might have an impact on the model as the population is a big factor.




```
[0.43270484 0.22046514 0.02855114 0.34805433 0.34014408 0.04205364]
```

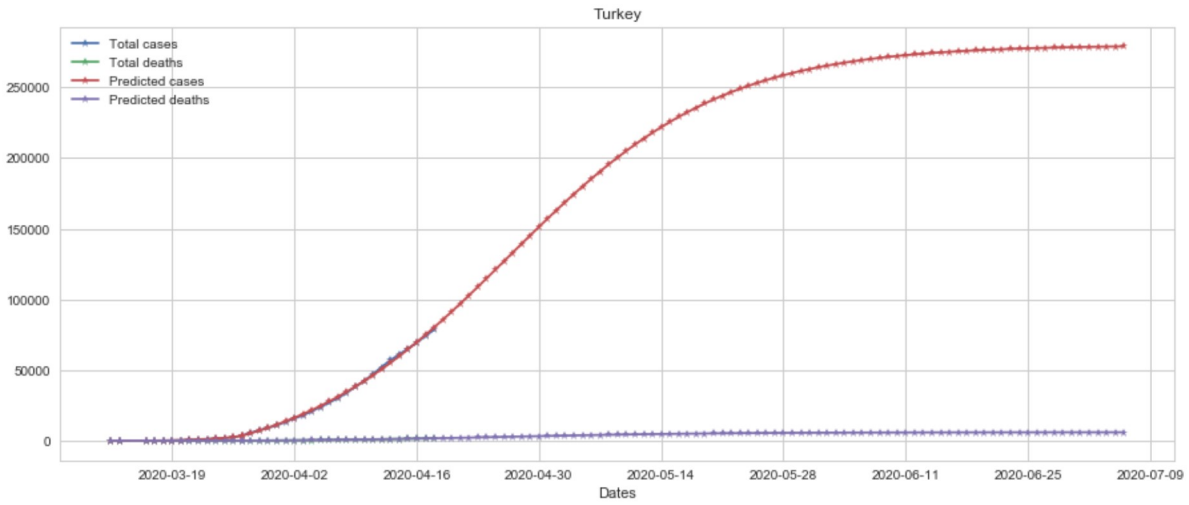
Out[28]: 0.013733665097188597

Another scary picture. Although the number of deaths do not increase, the model shows that the cases will keep increasing.

Now let's look at Turkey. But before that, let's choose the day with some quantitative selection. We will check 5 different days of inflection and take the best fit for current data.

Out[29]:
```
[0.041304332087760916,
 0.02267094444441011,
 0.01956404178706028,
 0.024059701649302222,
 0.022048107082470073]
```

Looks like March 28th gives the best fit. Let's see the predictions now.




```
[26.70982795 25.83049435  0.59991655 45.77941644 44.72374519  0.98463048]
```

Out[31]: 0.01956404178706028

It looks like it's not going to end pretty soon. Hopefully the model is overestimating, just like the case with China. Also, the government seems to be imposing more restrictions now, so let's hope some parameters change and it gets better.

I am not going to 80 days. With this many cases it is not very reasonable. Let's hope that the model is overestimating.

Discussion

- In this report we analyzed the COVID-19 Data for case and death count prediction using SIRD model.
- As the model does not take the parameter changes into account, for an ongoing pandemic it is hard to estimate the best values.
- Yet, making an assumption in regards to the stages of the pandemic, we were able to create a model that can fit and predict better than the original model.
- Using the model parameters some predictions were made.
- **The predictions have no implications as the author is not an expert in the subject.**
- It can be seen that the model tends to overestimate and is not taking many underlying information into account, still some predictions can easily made regarding the implications of an early lockdown or public awareness.

Conclusions

1. Earlier lockdowns, or other measures effect the number of infected greatly, reducing also the death toll.
2. The two stage model approach can be extended to three or higher stages as sometimes countries take extra measures after a new discovery.
3. The model might be improved using number of tests, recoveries and other data into account. This might provide better and more trustworthy predictions.
4. The model also can be improved to focus on states, cities and counties and their relationship to better explain the underlying dynamics such as connectivity and population density.

References

[1] “Download Today's Data on the Geographic Distribution of COVID-19 Cases Worldwide.” European Centre for Disease Prevention and Control, 18 Apr. 2020, www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide.

[2] “Compartmental Models in Epidemiology.” Wikipedia, Wikimedia Foundation, 15 Apr. 2020, en.wikipedia.org/wiki/Compartmental_models_in_epidemiology#The_SIR_model.