Tria COVID-19 Datathlon

Seyyid Emre Sofuoglu

04/25/2020

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.

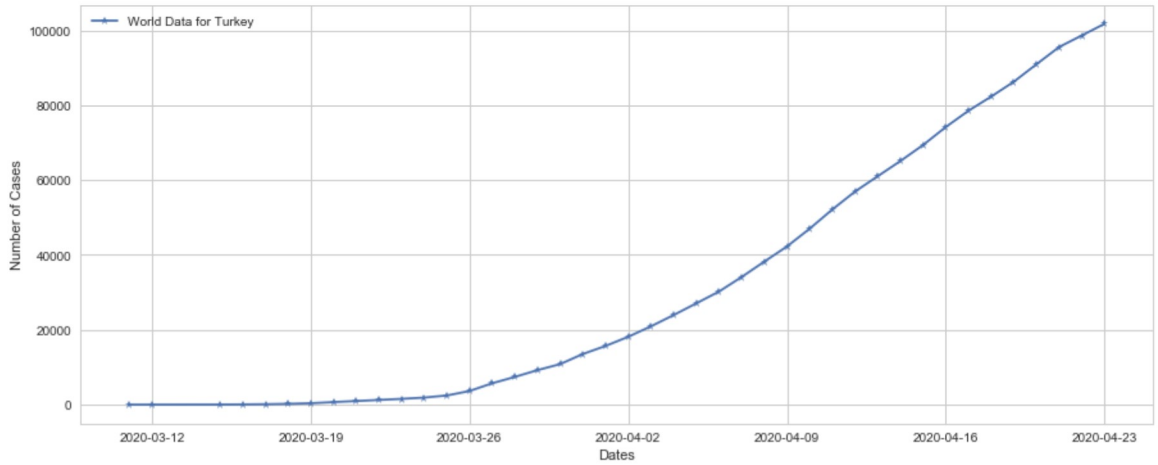
Data

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	countryterritory
0	4/24/2020	24	4	2020	105	2	Afghanistan	AF	
1	4/23/2020	23	4	2020	84	4	Afghanistan	AF	
2	4/22/2020	22	4	2020	61	1	Afghanistan	AF	
3	4/21/2020	21	4	2020	35	2	Afghanistan	AF	
4	4/20/2020	20	4	2020	88	3	Afghanistan	AF	

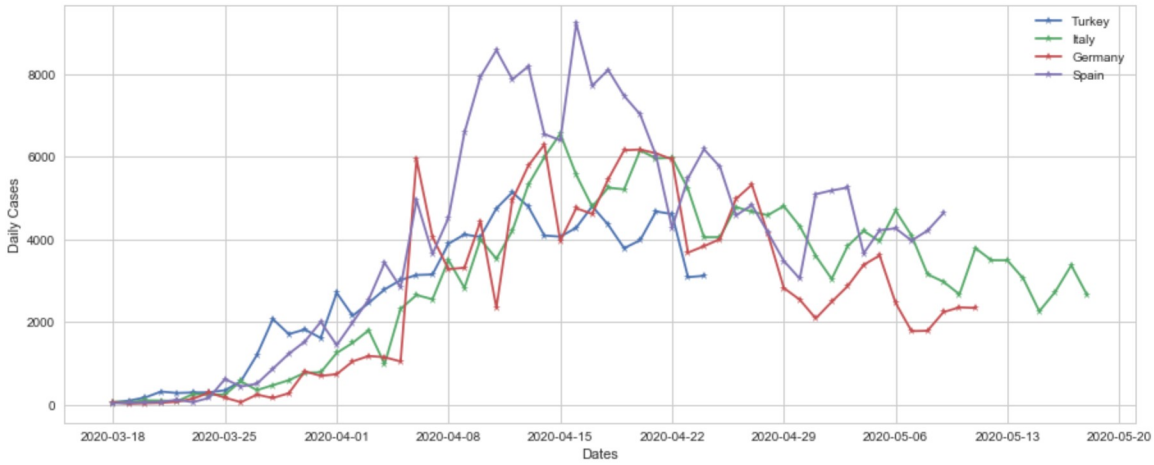
Now, let's have a look at the data on Turkiye.



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.

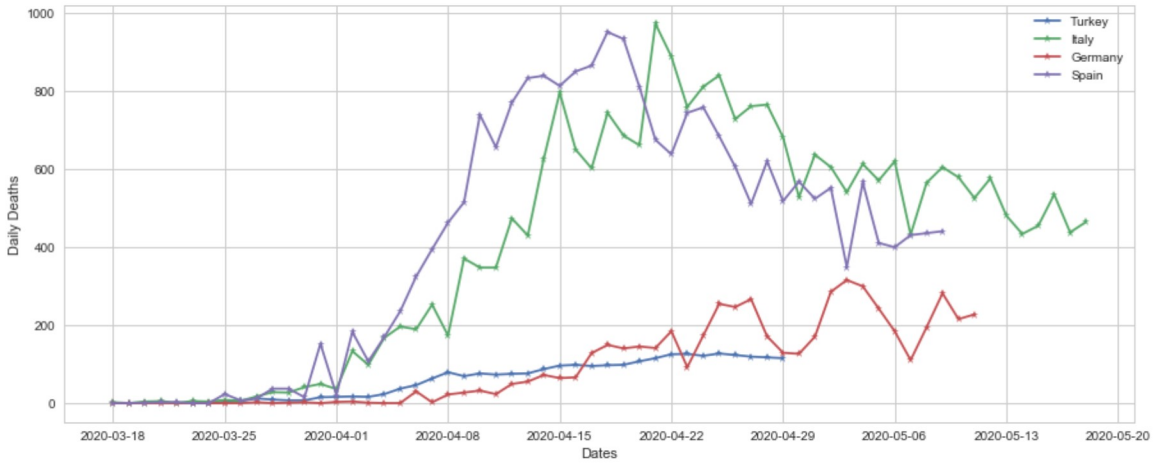
Now, let's see some comparisons between countries taking 100 total case counts as basis, or first day of the spread.

Processing math: 100%



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:

$$\frac{dS}{dt} = -\frac{\beta IS}{N},$$

$$\frac{dI}{dt} = \frac{\beta IS}{N} - \gamma I - \mu I,$$

$$\frac{dR}{dt} = \gamma I,$$

$$\frac{dD}{dt} = \mu I,$$

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.

So, assuming we found the parameters for a country, how do we quantify the results? We can use an estimation error e defined by:

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

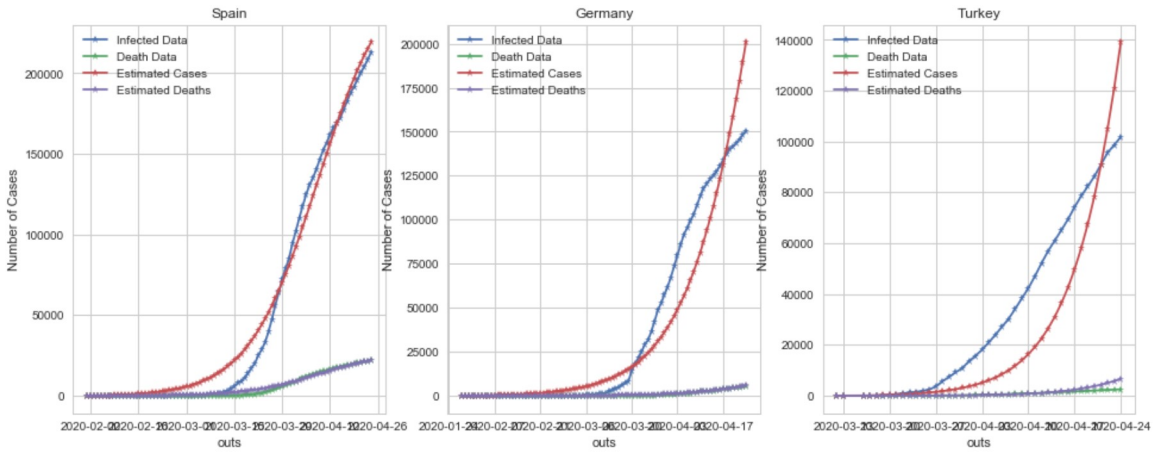
where y is the real number of cases as a time series vector, $\hat{y}$ is the vector, or array estimated by the model and $\|.\|_2$ is the l_2 norm, defined by:

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

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. These countries all have high test counts, so their data are reliable. They're our candidates for testing our models.

Let's fit the parameters on a couple of countries, e.g. Germany, Spain and Turkey, and see how it looks.

```
The parameters of Spain are: beta(S->I) = 34.59 , gamma(I->R) = 31.05
, mu(I->D) = 3.4468
The parameters of Germany are: beta(S->I) = 20.64 , gamma(I->R) = 19.
96 , mu(I->D) = 0.5978
The parameters of Turkey are: beta(S->I) = 33.02 , gamma(I->R) = 31.2
8 , mu(I->D) = 1.5765
```

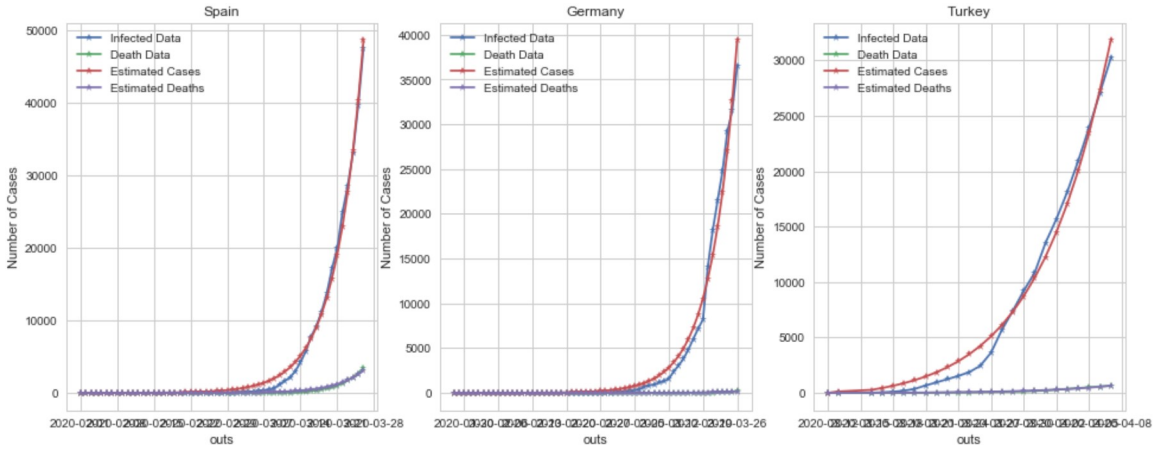



```
Error for Spain is: 0.0972
Error for Germany is: 0.2469
Error for Turkey is: 0.3585
```

The fitted curves, while looking similar, are not very accurate. They are 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, i.e. 14 days after measures took effect. For Spain we believe it's around 27th of March as this was 14 days after the government issued a state of alarm (https://en.wikipedia.org/wiki/2020_coronavirus_pandemic_in_Spain#State_of_alarm_(13%E2%80%9327_March)). In Germany, at 13th of March, the government transitioned to *Protection* (https://en.wikipedia.org/wiki/2020_coronavirus_pandemic_in_Germany) stage and wide closures took effect. In Turkey, widespread measures started at March 16th, when the first 100 case thereshold was passed. So let's use March 27th for all these countries as the change point.

```
The parameters of Spain are: beta(S->I) = 0.37 , gamma(I->R) = 0.16 ,
mu(I->D) = 0.0233
The parameters of Germany are: beta(S->I) = 0.13 , gamma(I->R) = -0.0
5 , mu(I->D) = 0.0006
The parameters of Turkey are: beta(S->I) = 111.42 , gamma(I->R) = 10
9.02 , mu(I->D) = 2.2317
```



The parameters for Spain and Germany look similar, but we can see that they are very different in Turkey.

In Germany and Spain, the number of cases were staying still for a long time. This was probably because at early stages of the pandemic, testing was not widespread and goverments were not alert. Though Turkey, as the pandemic started at around March 10th, started contact tracing and widespread testing immediately. In fact at March 3rd it had 900 tests with no cases, while US had 400 tests and a death. This clearly illustrates the rapid reaction, which can also be seen in the figure. Thus, the spread rate looks faster for Turkey as it has reliable data, whereas for Germany and Spain, earlier days do not really look realistic.

Let's check the error values of the above.

```
Error for Spain is: 0.0565
Error for Germany is: 0.1077
Error for Turkey is: 0.0775
```

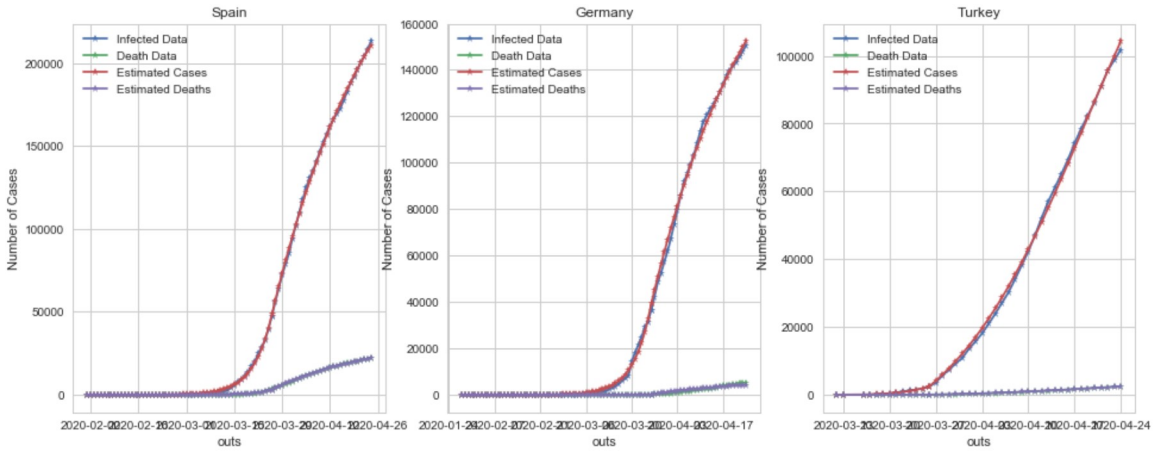
It seems the models fit the early stages very well, this brings the question:

```
 How well will the model work if I assume there are two stages and fit two
completely different models for these two stages?
```

Results

So let's apply a two stage model and check the errors. We can also look at the prediction errors where we estimate values for days we did not see the relevant data to help us understand the effectiveness of the model in terms of prediction.

```
The parameters of Spain at initial stage are: beta(S->I) = 0.37 , gam
ma(I->R) = 0.16 , mu(I->D) = 0.0233
The parameters of Spain at latter stage are: beta(S->I) = 0.35 , gamm
a(I->R) = 0.34 , mu(I->D) = 0.0408
The parameters of Germany at initial stage are: beta(S->I) = 0.13 , g
amma(I->R) = -0.05 , mu(I->D) = 0.0006
The parameters of Germany at latter stage are: beta(S->I) = 0.11 , ga
mma(I->R) = 0.13 , mu(I->D) = 0.0039
The parameters of Turkey at initial stage are: beta(S->I) = 32.79 , g
amma(I->R) = 32.31 , mu(I->D) = 0.2345
The parameters of Turkey at latter stage are: beta(S->I) = 80.42 , ga
mma(I->R) = 78.47 , mu(I->D) = 1.8784
```



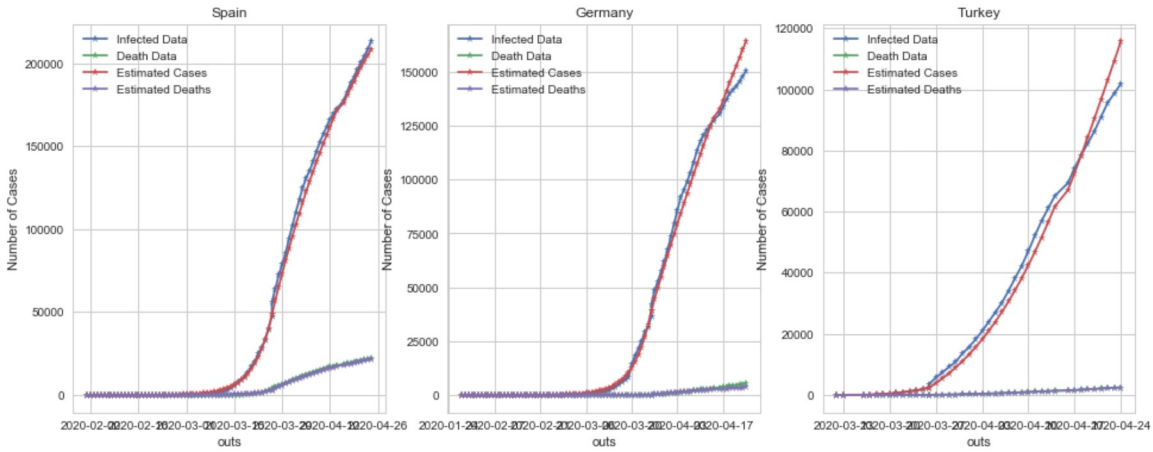

```
Error for Spain is: 0.0116
Error for Germany is: 0.025
Error for Turkey is: 0.0218
```

Now, this looks much better than one stage SIRD. Also, we can see that parameter β dropped to less than half its value at the latter stage. These parameters can be thought of the quantified results of the measures.

We can see that spread rate dropped a lot for Germany and Spain. Weirdly enough, recovery rate fell down for Spain. In the case of Turkey, on the other hand the spread parameter increased even further and recovery rate also jumped. These parameters might be overfitted to 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. We clearly need data for recovered for these parameters to make sense.

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

```
The parameters of Spain at initial stage are: beta(S->I) = 0.37 , gam
ma(I->R) = 0.16 , mu(I->D) = 0.0233
The parameters of Spain at latter stage are: beta(S->I) = 0.35 , gamm
a(I->R) = 0.34 , mu(I->D) = 0.0406
The parameters of Germany at initial stage are: beta(S->I) = 0.13 , g
amma(I->R) = -0.05 , mu(I->D) = 0.0006
The parameters of Germany at latter stage are: beta(S->I) = 0.1 , gam
ma(I->R) = 0.11 , mu(I->D) = 0.0027
The parameters of Turkey at initial stage are: beta(S->I) = 32.79 , g
amma(I->R) = 32.31 , mu(I->D) = 0.2345
The parameters of Turkey at latter stage are: beta(S->I) = 68.33 , ga
mma(I->R) = 66.74 , mu(I->D) = 1.5001
```

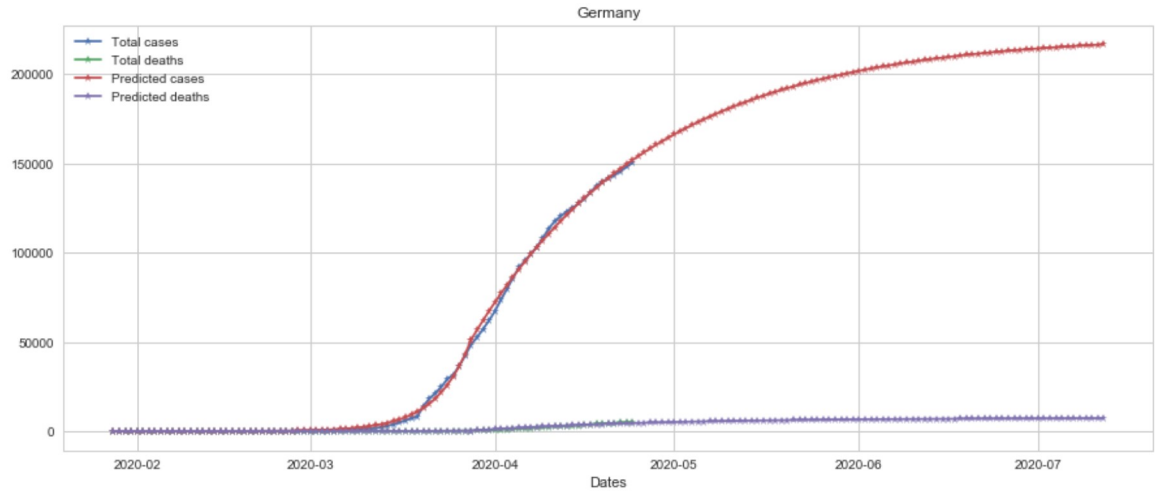



```
Error for Spain is: 0.0362
Error for Germany is: 0.0558
Error for Turkey is: 0.0835
```

We can see that for Turkey and Germany, the model tends to overestimate the last 10 days. But it is not too far off and we have seen that with more data the model fits really well.

Prediction

Now, using our model, let us try to predict the future values for several countries. We are going to predict 80 days which even [experts (https://www.nytimes.com/interactive/2020/04/22/upshot/coronavirus-models.html?algo=top_conversion&fellback=false&imp_id=802743813&imp_id=788768220)](https://www.nytimes.com/interactive/2020/04/22/upshot/coronavirus-models.html?algo=top_conversion&fellback=false&imp_id=802743813&imp_id=788768220) refrain from so I would suggest you do not take this seriously.

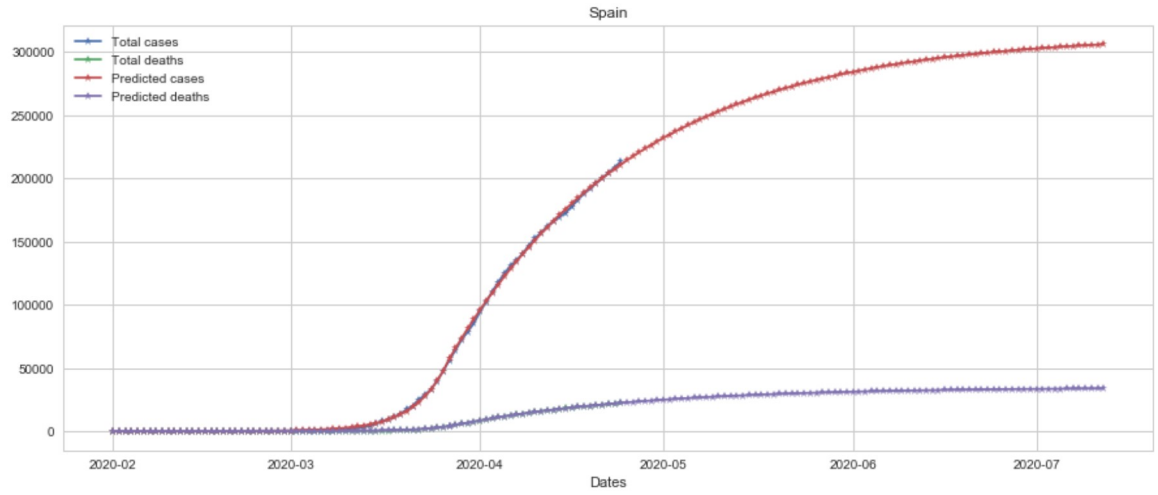



```
[0.30334233 0.13051881 0.0015962  0.19429423 0.21875315 0.00811525]
```

Out[27]: 0.02609954227937675

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.

Let's see Spain, a country that struggled and lost a lot.




```
[0.43270484 0.22046514 0.02855114 0.34077737 0.33081015 0.04090238]
```

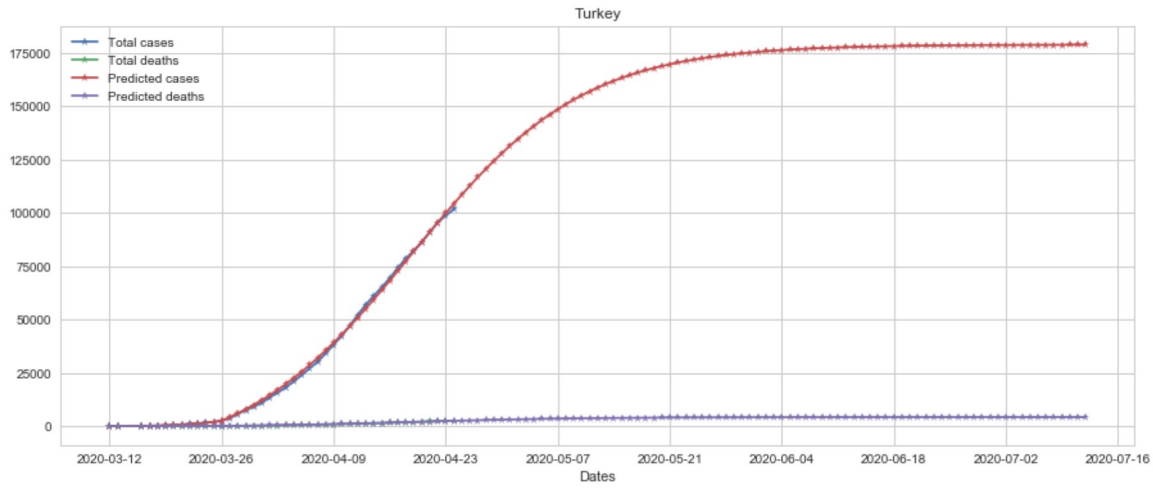
Out[31]: 0.011902452440405296

Although the number of deaths do not increase too much, the model shows that the cases will keep increasing. Looking at Spain and Italy, if we compare them with Germany and South Korea, an early lockdown, curfew or other strong measures affected the end results a lot. Thus, we repeat the experts, every day is critical.

Now let's look at Turkey, a late starter in terms of the first case. But before that, let's choose the day with some quantitative selection. We will check 9 different days of inflection (4 days to the past and future) around our guess (March 29th) and take the best fit for current data.

```
Out[32]: [0.03878347051763411,
          0.027407359322468725,
          0.02180130550391481,
          0.02674537409666896,
          0.033663119992450204,
          0.03317425564871676,
          0.1712116026189017,
          0.026950635125211047,
          0.02988516840198759]
```

Looks like March 27th gives the best fit. Let's see the predictions now.




```
[32.79015356 32.30625565  0.23446839 80.42289203 78.46670042  1.87835
521]
```

```
Out[34]: 0.02180130550391481
```

The parameters are very high which puts the estimates into question. Still, it looks like it's not going to end pretty soon. Also, the government seems to be imposing more restrictions now, so let's hope that the real parameters change and it gets better in a stage 3.

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 is observed 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.