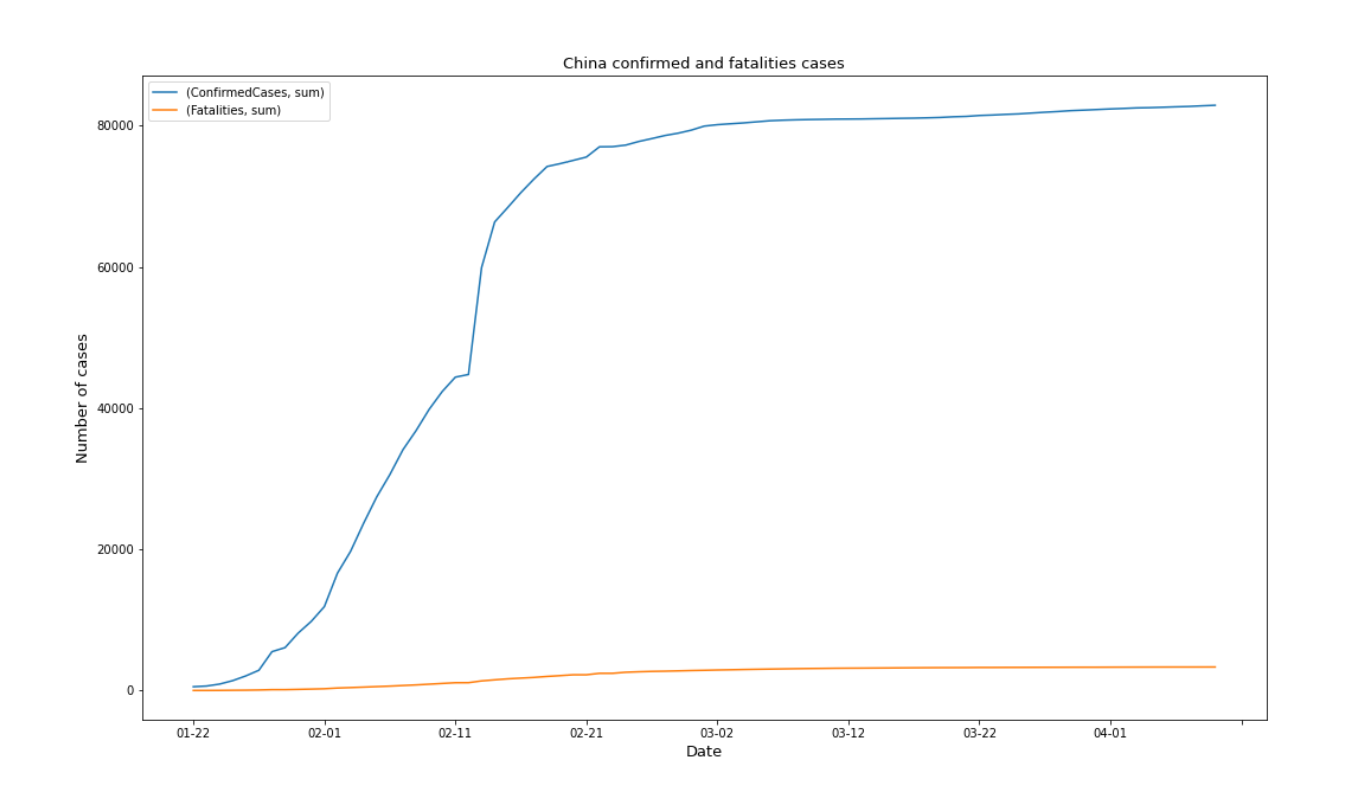
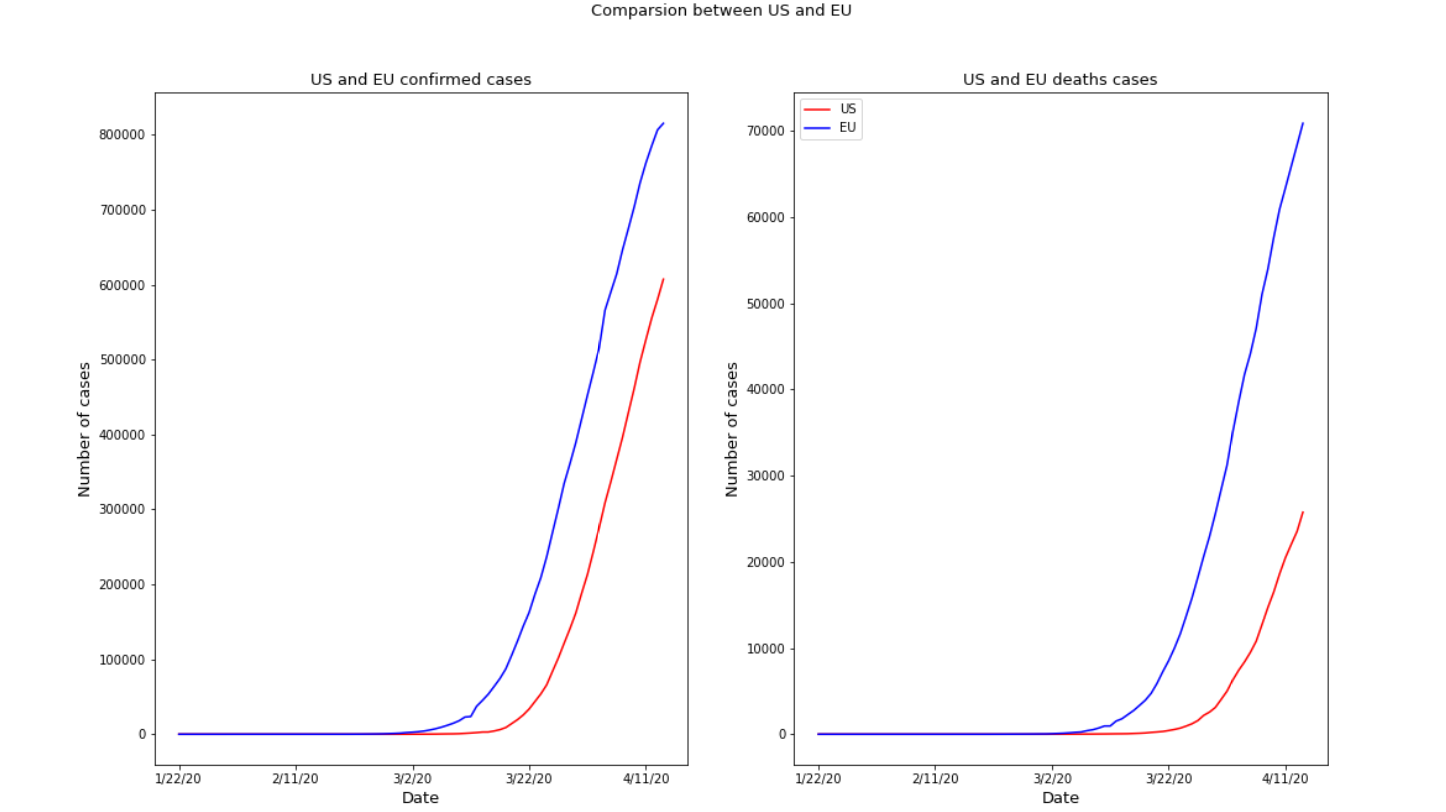
# 1.Exploratory data analysis (EDA)

It is reasonable to first analyze and visualize China, the first country to be infected by the COVID-19 on a large scale. We should note that when the virus grows and spread in China, there is not enough information, data, effective methods of prevention, and attention. Besides, China took forced isolation to the suspected cases and confirmed cases which are not allowed in democratic states. Moreover, the WTO updated the definition of diagnosis. That’s the reason why the curve is not very smooth. So, it is also reasonable to split it outside the worldwide group.

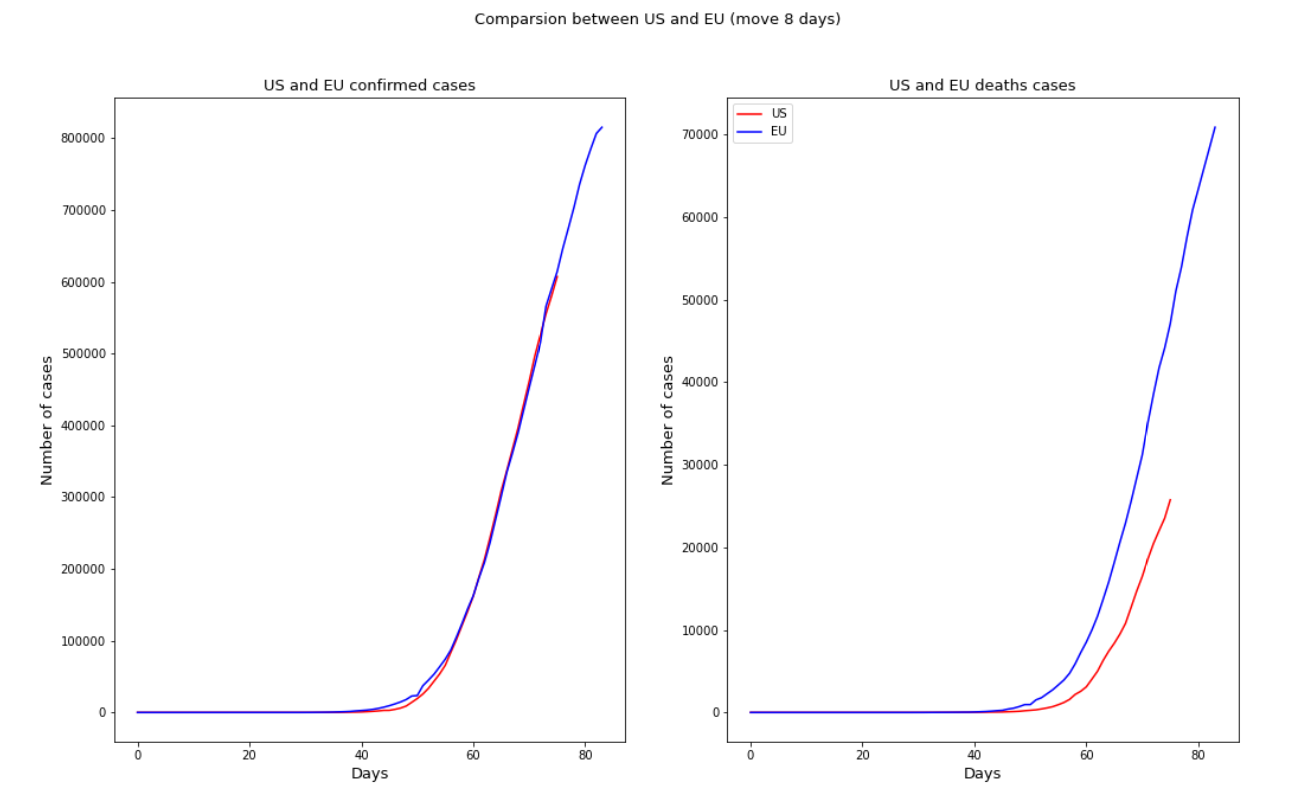


As we see, both the number of confirmed and fatalities cases have been growing very slowly since March, indicating that the epidemic is coming to the end. The graph also indicates that most of other counties in the world are placed in the range of Feb to the Mar.

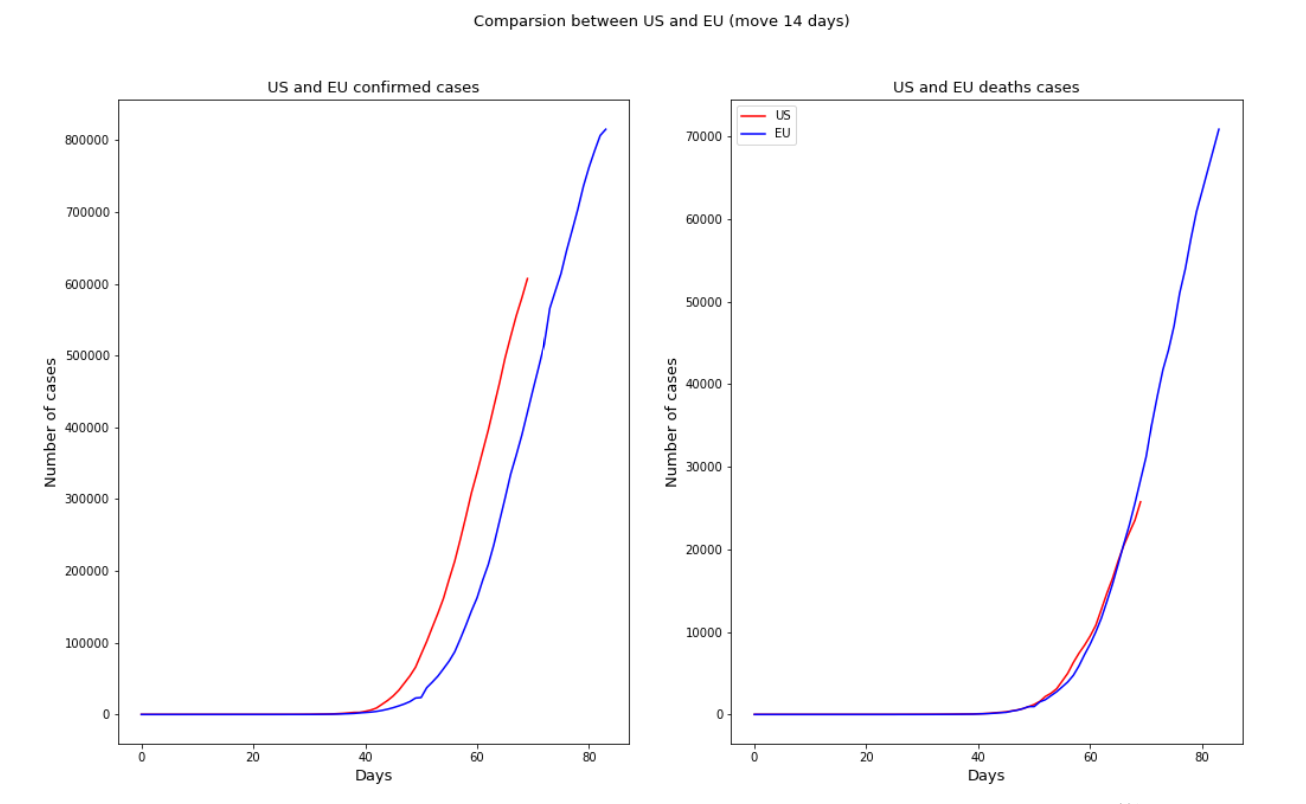


I want to group Europe as one entity. Europe's individual countries not only are very similar to the state governments of the United States, but also belong to the European Union as a whole (except the United Kingdom). Its overall population and land area can be compared with that of China and the United States in some extent.

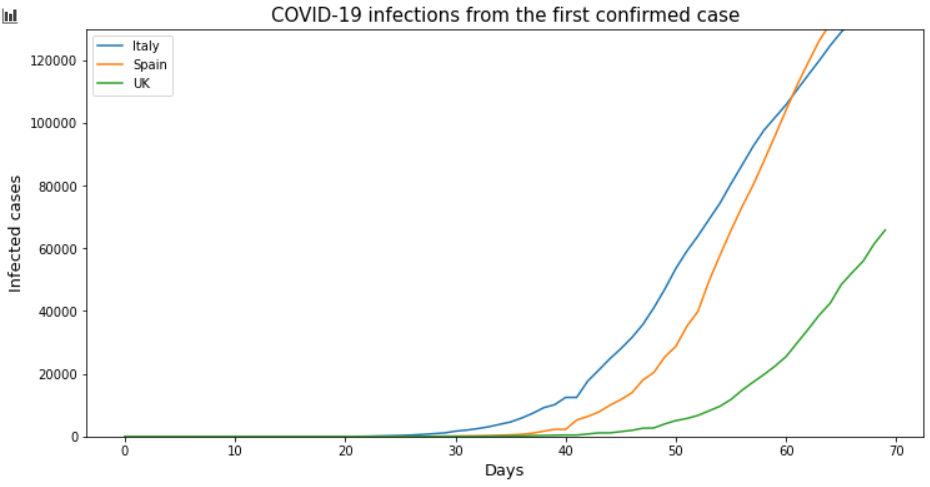
As the graph shown above, the Europe’ and US’ curve is very similar to each other’s. Intuitively, let me try to translate the curve of the United States to the left for a few days and see what will happens. The graph will be shown below.



After I shift the blue curve (EU) to the left for 8 days, it fits pretty well with the red line in confirmed cases. But the deaths cases curve is not as well as the confirmed cases curve. The reason maybe related to the diffident government’s policies and races. Again, I try to shift 14days, then I get a better result on the left part.



But this time the right part becomes worse obviously. So, maybe I need to 2 different models for each confirmed case and death case.



I also plot and compare some countries in the EU, the curves show quite different slope. So, the reason why the curves of large area are similar to each other is because the large area contains rich enough different features that related to COVID-19 (bad and good). So, their curves are more universal.

# 2.Model selection and visualization:

The main models of infectious diseases are SI, SIS, SIR, and SEIR. Each of them fits different infectious diseases. The capital letters inside these words stand for:

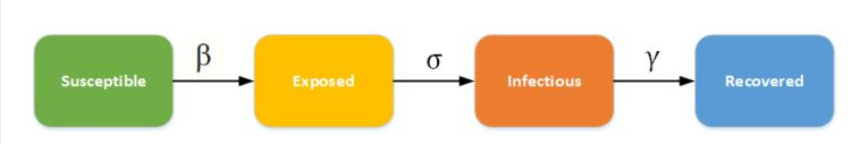
**·S (Susceptible):** the number of individuals that can get infected.

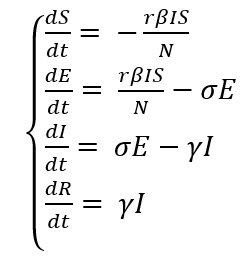
**·I (Infectious):** the number of individuals who are infected and can transmit the disease

**·R (Removed or Recovered):** the number of individuals who are immune.

**·E (Exposed):** the number of individuals who have touched the infected but are temporarily unable to infect others.

From the previous medical report, we can know that the COVID-19 most likely fits the SEIR model (the most complex one). SEIR model is a general model. If the infectious diseases studied have a certain incubation period, the healthy people who have contact with the patients do not get sick immediately but become the carriers of pathogens, which are classified as class E.





S: susceptible

E: exposed

I: infectious

R: recovered

t: time

β: Infection rate

N: total number

: recover rate

rate of the E change to the Infectious.

Besides, researchers not only found that the exposed people can also be infected susceptible and be recovered but also the recovered can become susceptible again. These features of COVID-19 make the model more complex than normal SEIR. Then, we have to use the equation below to fit our model.

Let me quickly introduce the equations set. The Susceptible can change to the Exposed and come from the Recovered. The Exposed comes from the Susceptible and can change to the Recovered or the Infectious. The infectious come from the Exposed and can change to Recovered. The recovered can change to the Susceptible and come from the Exposed or the Infectious.

S: Susceptible

E: Exposed

I: Infectious

R: Recovered

t: time

β­: infection rate for Susceptible

N: total number

: rate of the Exposed change to the Infectious.

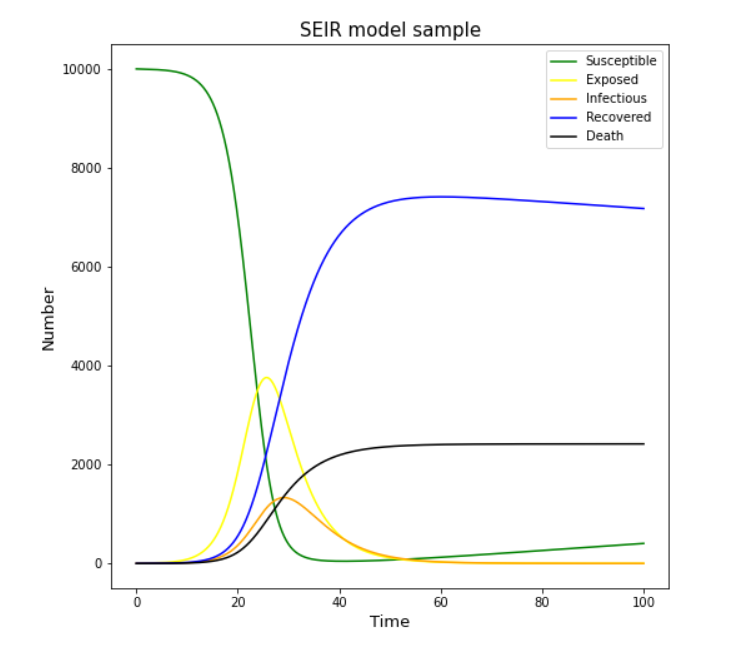
: recovery rate for the Exposed

: recovery rate for the Infectious

: death rate for the Exposed

: death rate for the Infectious

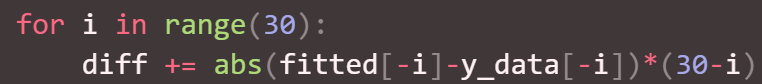
The sample graph is shown below. The paraments are N = 10000, r = 100, S0 = N-I0, E0 = 0, I0 = 1, R0 = 0, β = .03,  = .05,  = .2, 1 = .1, 2 = .001, = .03 .03.

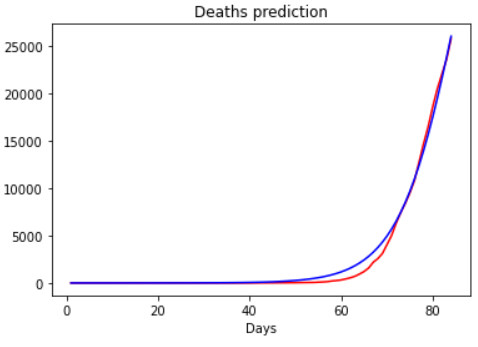
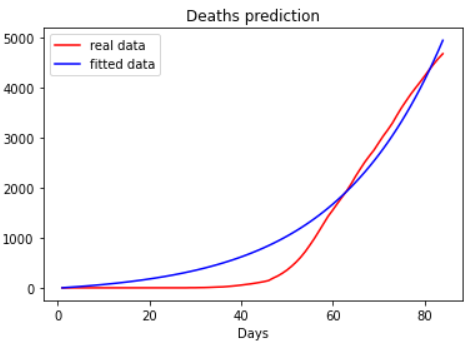


# 3. Experiment and prediction

First part I want to focus on the death cases.

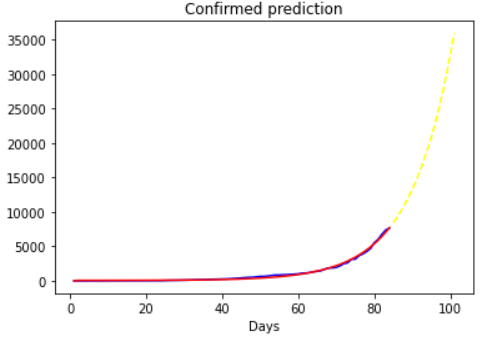
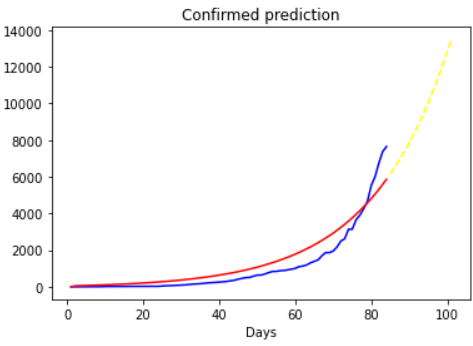
The way I adjust the parameters is brute force searching. Then compare the fitted data with the real data and choose the one that has the smallest differential. Also note that I set a wright to the day close the final day, that I can get a better trend in the future.



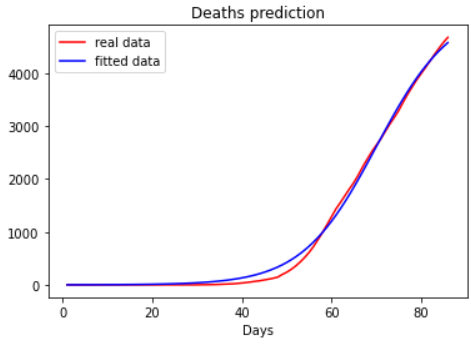
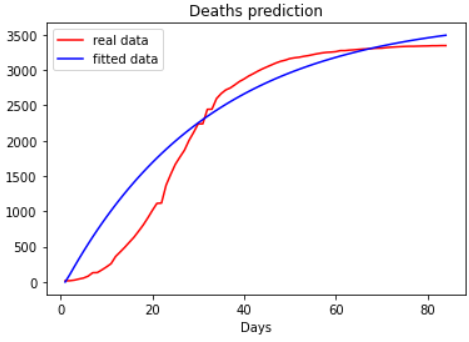


The above two graphs are two samples, both they are highly fitted in the days 60 to 80 but the second one did better at days 0 to 60 and also has better performance at latter prediction (there is future data available every new day). So, the range of differential is more reasonable focus on a larger range of data rather than the last several days.

I can also judge the performance by the last several days’ trend. 



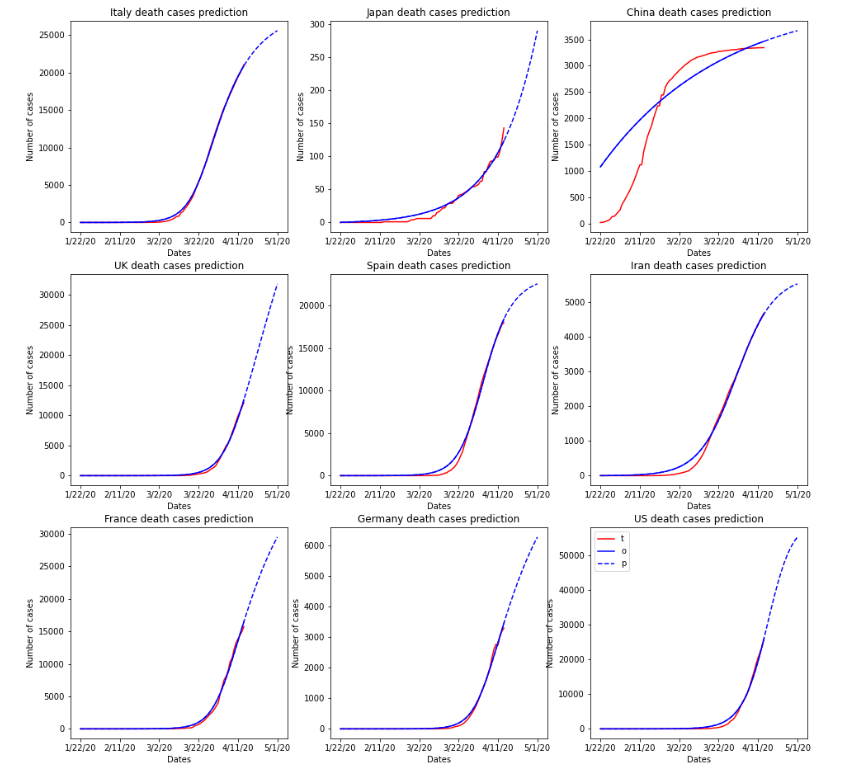
By this way, I will not select the one fitted better (right), but the one (left graph) perform better in the future. We can find that the left one is more reasonable though it doesn’t fit the blue curve as well as the right one.



Another point I found is some Asia countries, like Iran and China, are earlier hold Exposed and Infectious that European countries. So, they did not fit well because the data provided is not accurate (the left one is China). I need to search and increase more data than provided. As the right picture (Iran), it fits better after I add only one more day.

# 4. Result and conclusion

I used date interval form 1/22/20 to 4/14/20 as actual data. The prediction interval starts form 4/15/20 to 5/1/20.



An interesting point is that my prediction curve of China fits not very well, but precisely predicts a precipice increase sed on the change of Chinese government statistical model. So, the curve more fits to currency number.

# 5. Reference and datasets