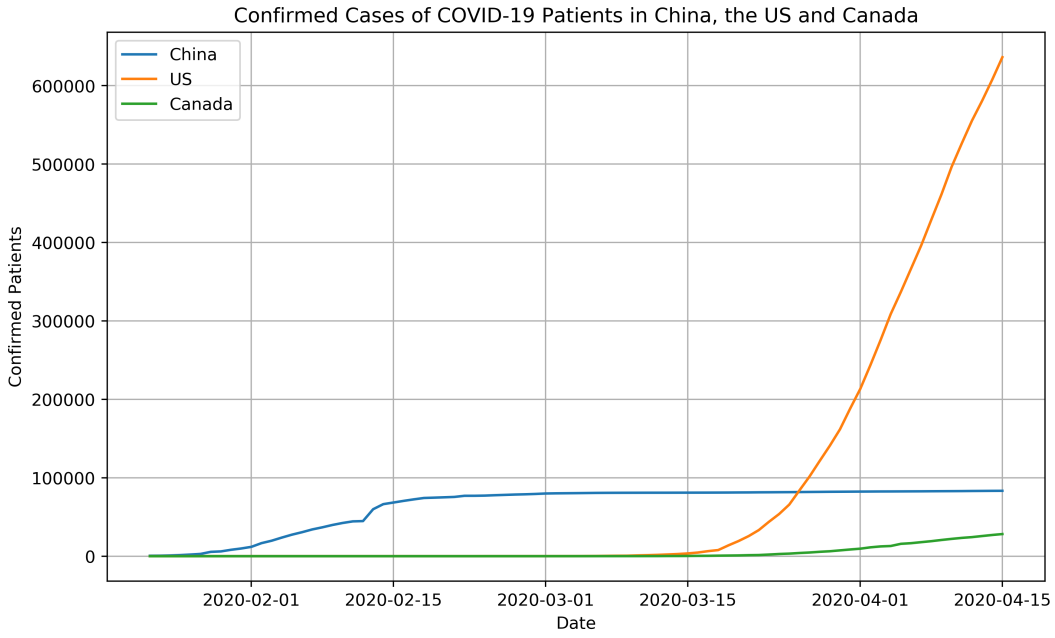
Forecasting the COVID-19 Time Series

ZhenDi Pan 1003241823 04/17/2020

Introduction

The Coronavirus, also known as the COVID-19 is a newly erupted disease caused by SARS-CoV-2 [1]. The first case was identified in December 2019 in Wuhan, China. Due to the virus being airborne and highly transmissible among humans, controlling the disease is quite difficult. Since the outbreak in January, the virus has spread rapidly to almost every country, thus causing a global pandemic. For a brief visualization of the confirmed patients, the data collected from January 22^{nd} 2020 to April 15^{th} 2020 for three countries is shown below:



Although I'm sure that mankind will beat this invisible enemy eventually, it is important and useful for people to have a general idea of when approximately will the disease be contained. Therefore, for an attempt to forecast the spread of this disease, I propose to develop a predictive model specifically designed to capture the properties of this virus. To do this, I will modify the existing SIR model [2], which is one of the simplest epidemiological model that computes the theoretical number of people infected with a contagious illness in a closed population over time. The model has three variables:

$$S = S(t)$$
$$I = I(t)$$
$$R = R(t)$$

where S is the number of individuals susceptible to the disease, I is the number of infected individuals and R is the number of recovered individuals. Then, we define the infected individuals to be the source of the spread, which has a certain probability of transmitting the disease to the susceptible individuals and also the infected individuals have chances of recovery or death, but in either scenario, the spreading stops. Next, we will modify the model accordingly to forecast the time series and perform an analysis on our results.

Analysis

Before employing our model for predictions, we have to set some assumptions for forecasting the time series:

- The population remains constant, we do not take into the account of birth or death by means other than the disease.
- As long as contact is made, there is the probability of infection
- The number of individuals removed from the infected individuals is proportionate to the total number of infected individuals.

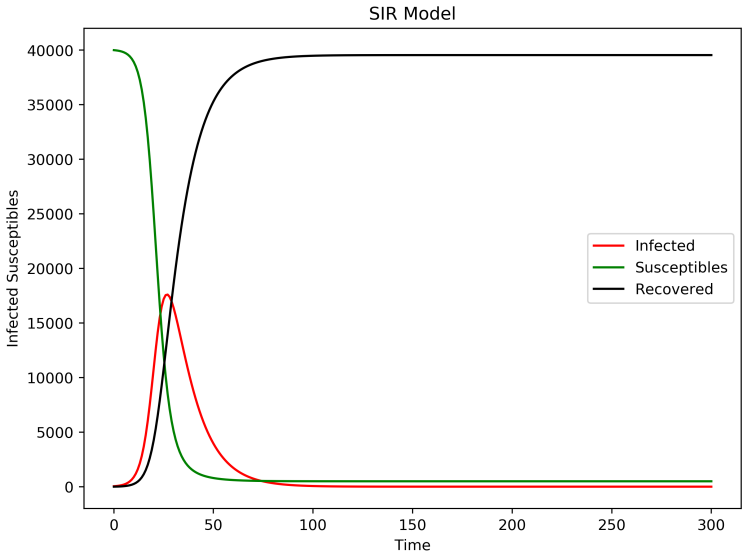
From these assumptions, the model gives the following differential equations:

$$\frac{dS}{dt} = -\beta S(t)I(t)$$
$$\frac{dI}{dt} = \beta S(t)I(t) - \gamma I(t)$$
$$\frac{dR}{dt} = \gamma I(t)$$

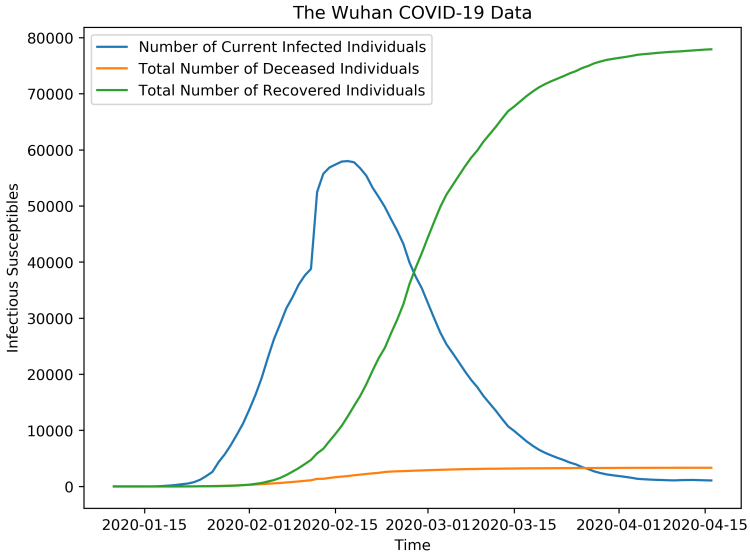
The first equation is the susceptible equation, where β is the infection coefficient that represents the probability of infection upon contact. The second equation is the infected equation, where γ is the rate of recovery. The third equation is the recovered equation, which implies we are interested in $\frac{1}{\gamma}$, the average infection period [3]. Given these three time series, we can simply compute the differential equations using Python:

```
1     def equations(input, t):
2         equation1 = -beta * input[0] * input[1]
3         equation2 = beta * V[0] * V[1] - gamma * V[1]
4         equation3 = gamma * V[1]
5         return equation1, equation2, equation3
```

We can then compute the result by using scipy.integrate. A simple simulation of the model can be plotted to visualize the characteristics:

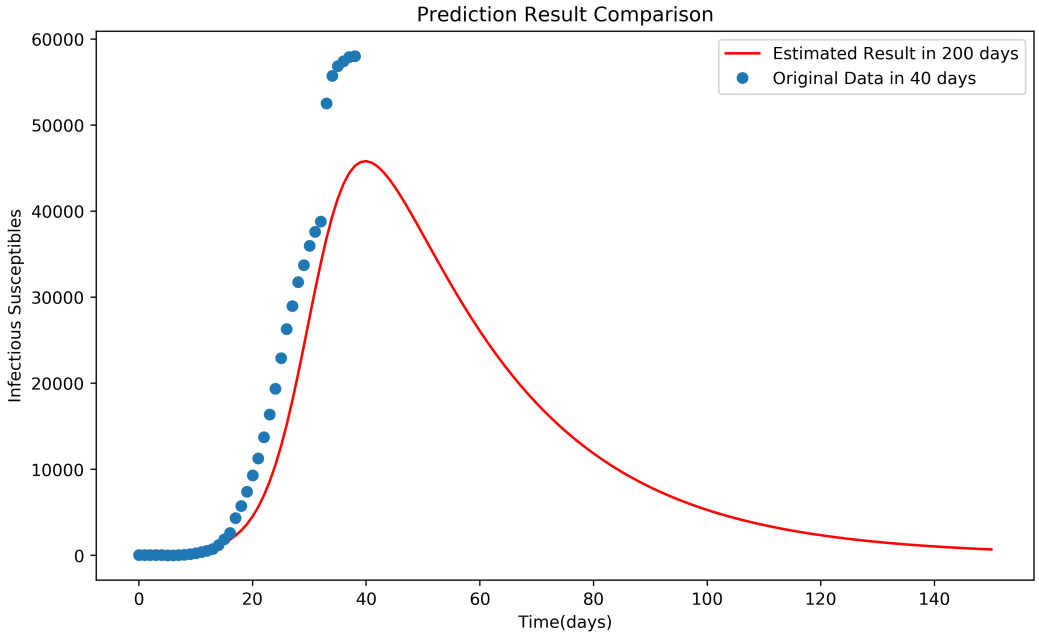


Furthermore, there is the basic reproduction number $R_0 = \frac{\beta}{\gamma}$. This number represents the expected number of susceptible individuals one infected individual will infect before death or recovery. The disease will only spread when this number is greater than 1. Theoretically, if $R_0 > 1$ then the virus will infect everyone on earth if no action is taken to prevent the spreading. The general idea is that the number of infected will first increase until it reaches a maximum where all the susceptible individuals are infected and then people start recovering. For the purpose of this report, I used the data from Wuhan, China to fit our model. Now, we only need to set our initial number of infected people I_0 and susceptible S_0. The first day of the collected data is January 10^{th} 2020 with 41 initial infected individuals. Let us first visualize the data by plotting it:



Note that the number of current infected individuals is the total number of infected minus the total number of death and recovered individuals. From this plot, we can see that mortality rate is quite slow, which is consistent with the new reports. The deceased patients are usually the

elderly people or with severe health issues before hand. In addition, there is also significant drop since February 15^{th}, which means the disease is somewhat controlled. There are many ways to estimate or investigate the coefficients for β and γ, and also the initial number of susceptible individuals. The simplest method is by using least squares. We define an objective loss function as $L = \sum_{i=0}^{N} y_{prediction} - y_{actual}$, where N is the total number sampled time series. For the purpose of testing the SIR model, let us use the data from the first 40 days and pretend that we do not know happens after February 20^{th}, I estimated $\beta = 3.5e-6$ and $S_0 = 85000$. Thus, we can finally plot our predictions:



We can see that the estimated time series captures the original time series quite well, they show almost identical trend especially at the start. The only imperfection is at the maximum, there seems to be a spike in the original data and the estimated result is visibly lower than the original data.

Discussion

From our results, we can see that the number of infected patients decreases significantly after 40 days of the initial outbreak, and after 120 days, the disease can be considered controlled. Given that our data starts from January 10^{th} 2020, 120 days approximately take us to the present moment. On April 14^{th}, the Wuhan city has finally ended its 76 days lock down and people are allowed to go outside of their neighborhood. The people still remain vigilant but the situation is finally under control. This is consistent to our model predictions and thus we can conclude that our results are quite reasonable and also the model performance is actually beyond my expectation. In reality, there are many more factors to consider than just the coefficients we estimated in our model. Therefore it is reasonable to have deviations in the results. To forecast the spread of a pandemic disease is quite difficult since there are just too many unpredictable

factors. For example, how quickly does the government react and how deadly is the disease are all contributing factors to this problem. I would imagine that just a month ago, none of us figured that the outbreak in US would be this terrible. Employing the SIR model on the Wuhan data is shown to be a success, which means we can use the similar method on other cities or countries that are currently being plagued by this virus. However, I doubt that we will get such promising results if we use the model on other cities, perhaps for some certain cities it might even be a completely off the charts failure. Nevertheless, time series analysis is always useful and many times even necessary to our society, with this tool the governments and the related institutions can make quick and effective actions.

Conclusion

In conclusion, we employed a SIR model to forecast the future number of confirmed cases of COVID-19 patients in the city of Wuhan. The results were consistent with the actual trend of the disease. We can also employ the same technique on many other cities, which can provide insights to the governments and the people on how to deal with the situation. Predictive modeling is always useful to our society, because making accurate predictions provides us the initiative to take actions in foresight.

References

[1] Wikipdepia: https://en.wikipedia.org/wiki/Coronavirus_disease_2019.

[2] Callahan, J. "The Spread of a Contagious Illness." https://maven.smith.edu/ callahan/ill/pde.html.

[3] David Smith and Lang Moore. "The SIR Model for Spread of Disease - The Differential Equation Model".

[4] "Novel Coronavirus 2019." https://datahub.io/core/covid-19resource-us_confirmed.

[5] China CDC (CCDC): http://weekly.chinacdc.cn/news/TrackingtheEpidemic.html.