Data Science Capstone -

Monitoring COVID-19 Infections through Wastewater Surveillance

Kevin Villanueva | Helen Huang | Achyuth Varma

Fall 2020

Introduction and Motivation

Problem

How do different parameters such as how easily the disease can be transmitted over time (beta), how easily people can recover (gamma), and how people's reactions/precautions towards the disease (alpha) effect the SIR model.

Background Info

We use the SIR model in order to understand the effects of certain parameters.

We can use the viral load of the waste water to see how many people are infected at a given point in time.

We can use a tree to represent the city's sewer system where the leaf node represent a certain community.

For each leaf node we divide segments of the population through social vulnerability and comorbidity.

Usefulness/Challenges

Usefulness

As we mess with parameters, such as how easily the disease can change over time, we see that we are able to make predictions on how people easily people become infected which, in a real world situation, can be very useful.

Challenges

We have a limited amount of data that we believe is correlated to understanding how many people are infected.

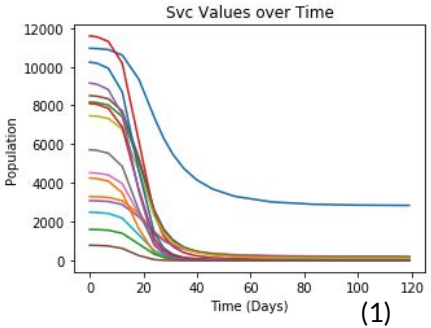
In an applied scenario we do may have a harder time dealing with the noise in the data.

Part 1

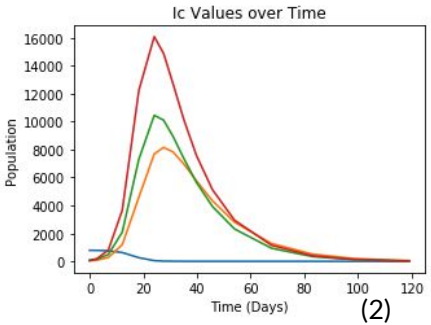
(a) Simulation of the behavior of the disease over 120 days; arbitrary values for the typical time between contacts.

When we *flatten the curve* we are referring to the **I value over time** because we want to limit the maximum amount of people being infected at the same time.

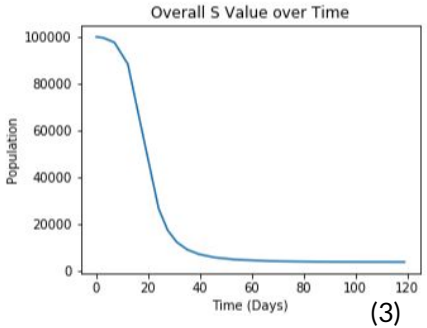
One of the Svc compartments does not converge to 0 because some people were never infected.



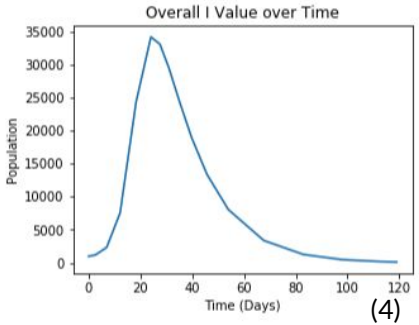
(1)



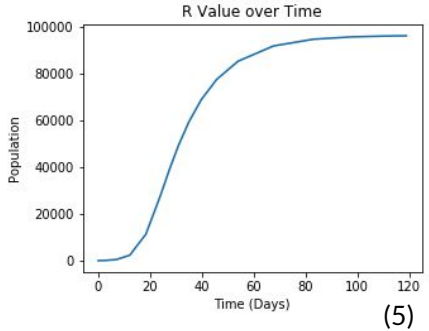
(2)



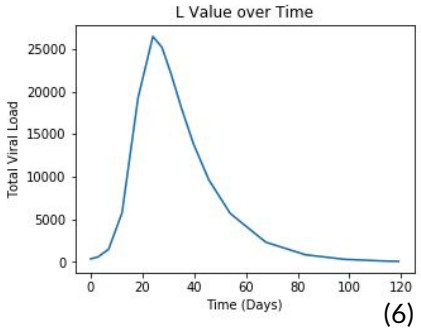
(3)



(4)



(5)

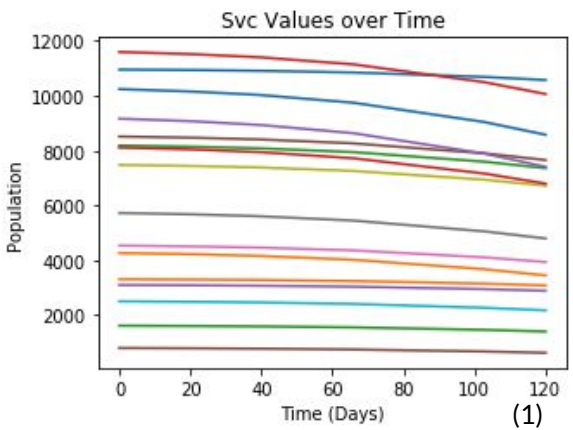


(6)

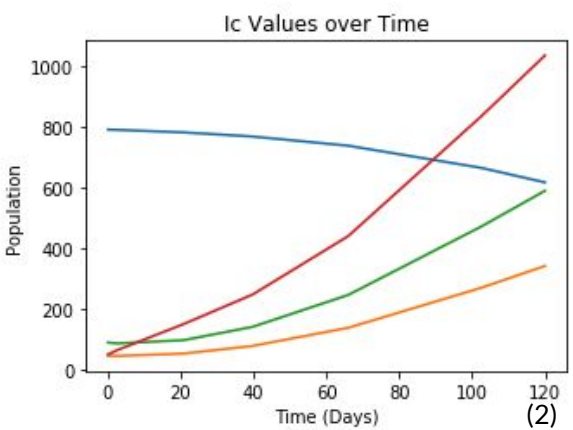
Figure 1. Simulation of the behavior of the disease over 120 days
(1. Svc values of every single compartment over time, 2. Ic values of every single compartment over time, 3. Overall S value over time, 4. Overall I value over time, 5. R value over time, 6. L value over time)

(a)

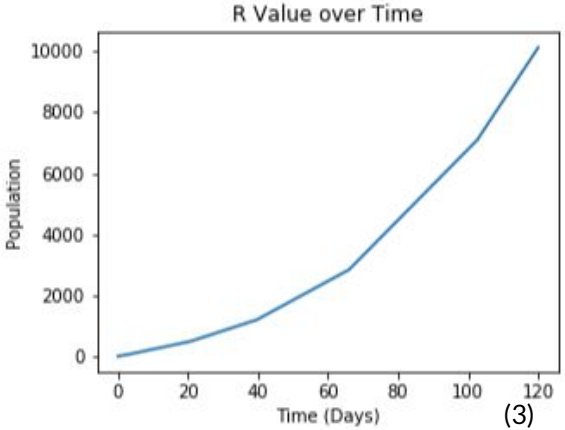
Beta values * ¼: Compared to the previous case, the curves flattened a lot, indicating that the percentage of population that never got infected increases a lot, which is 87.36% in this case. This can be explained by smaller beta values, which means lower time between contacts and easier transmission.



(1)



(2)



(3)

Figure 2. Simulation of the behavior of the disease over 120 days given the new beta values
(1. Svc values of every single compartment over time, 2. Ic values of every single compartment over time, 3. R value over time)

(b)

Given 5 leaf nodes for the vulnerability and the PMF for comorbidity, we tested different values of beta (from 0 to 1) that produced the closest values (minimal mean squared error) to the viral load density in comparison to the validation data that we were given using GridSearch method in Python.

After making sure our code and logic work fine for estimating the beta values, we estimate the beta values for the "test" data with 20 L values for 20 days, and we find the following estimation (without ground truth):

With ground truth

```
Ground truth betas:
 [[0.1  0.15 0.2  0.25]
 [0.2  0.25 0.3  0.4 ]
 [0.35 0.45 0.5  0.6 ]
 [0.4  0.5  0.6  0.8 ]]

Estimated betas:
[0.1  0.15 0.2  0.25]
[0.2  0.25 0.3  0.4 ]
[0.35 0.45 0.5  0.6 ]
[0.4 0.5 0.6 0.8]
```

Without ground truth

```
Estimated betas:

[0.05 0.1  0.15 0.25]
[0.2  0.25 0.3  0.4 ]
[0.4  0.45 0.5  0.6 ]
[0.45 0.65 0.75 0.85]
```

(b)

Using the estimated beta values in the last slide, we predicted disease behavior for the next 100 days. The four sets of graphs below correspond to each level of vulnerability, with the orange line on the viral load graph being the given observed values.

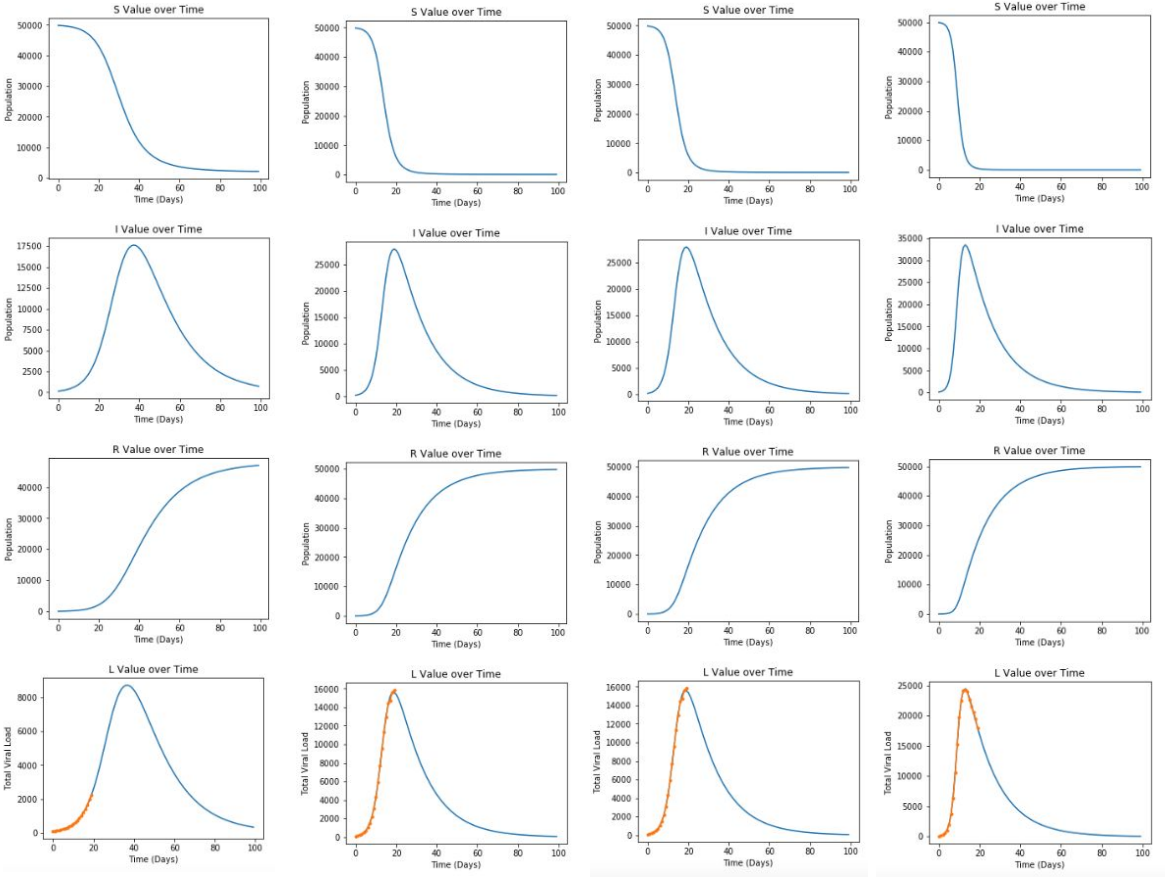


Figure 3. Simulation of the behavior of the disease over 100 days for the first node with the estimated beta values
Column 1: group with vulnerability = 0.2, Column 2: group with vulnerability = 0.4, Column 3: group with vulnerability = 0.6, Column 4: group with vulnerability = 0.8
Row 1: S values, Row 2: I values, Row 3: S values, Row 4: L values with the corresponding observed values for the first 20 days

Part 2

(a)

Until now, we'd only assumed that people would continue their behaviors even after the onset of an epidemic. Now, we consider what happens when people shift their behavior once the epidemic starts. Thus, we introduced a variable (alpha) which delineates the effect of people being more careful.

We implemented the effect of this alpha value on day 10. Using the day 9 values as the initial values, we went on to simulate different possibilities for days 10-29. We guessed the alpha values and used the one that minimized the MSE between our predicted viral load density and the given viral load density for each node.

```
Alpha_Predicted Values from Lowest to Highest:
 [0.12352142 0.13822021 0.14249268 0.14443359 0.14606934 0.14807129
 0.15137939 0.15171814 0.1552803  0.16387329 0.16503143 0.16713867
 0.17045288 0.17644043 0.18051758 0.19622192 0.20018311 0.20029602
 0.20103149 0.20452881 0.20931396 0.20965576 0.21018066 0.22185059
 0.22246094 0.22650146 0.22669983 0.23173828 0.23728027 0.24645996
 0.24889221 0.25011292 0.25706787 0.26390381 0.26639404 0.26811523
 0.27178955 0.27543945 0.27561264 0.27568665 0.28189697 0.28358765
 0.28780823 0.2932373  0.29797363 0.3010498  0.30296936 0.31580811
 0.32045898 0.3225708  0.32358704 0.33059082 0.330896   0.33418884
 0.33525391 0.33642883 0.34067383 0.35007172 0.35155029 0.35754395
 0.3601532  0.36387939 0.37408447 0.3757782  0.38779907 0.39297028
 0.39484253 0.41170807 0.41690369 0.41756592 0.42473755 0.43126221
 0.43154907 0.44144897 0.44838867 0.45488892 0.46865234 0.49281006
 0.51626587 0.51790771 0.52313232 0.53582764 0.54241028 0.56386719
 0.56835938 0.56982422 0.59924927 0.61907349 0.66160202 0.66409302
 0.66747131 0.68422394 0.69208984 0.70143738 0.71300049 0.72009888
 0.73707123 0.89589844 0.90992126 0.98475342]
```

Figure 4. Predicted alpha values for the 100 nodes, ranked from lowest to highest

(b/c)

We learned that non-pharmaceutical interventions (NPI) can help slow the spread of the disease. To do this, the local administration allocate a budget of $1,000,000 for all 100 nodes. We worked through 4 policies (criteria) to distribute the budget, study the relationship between the distribution of budget and alpha values, and tested fairness of all four policies.

To visualize how the policies work, we plot the alpha values before and after NPI where marker sizes are determined by population sizes and marker colors are determined by average beta values for that node. To see how effective the NPI is, we calculate the difference in the number of recovered people in total after 200 days with and without NPI.

(b/c)

Policy number 1: give each node (community) the same amount of money

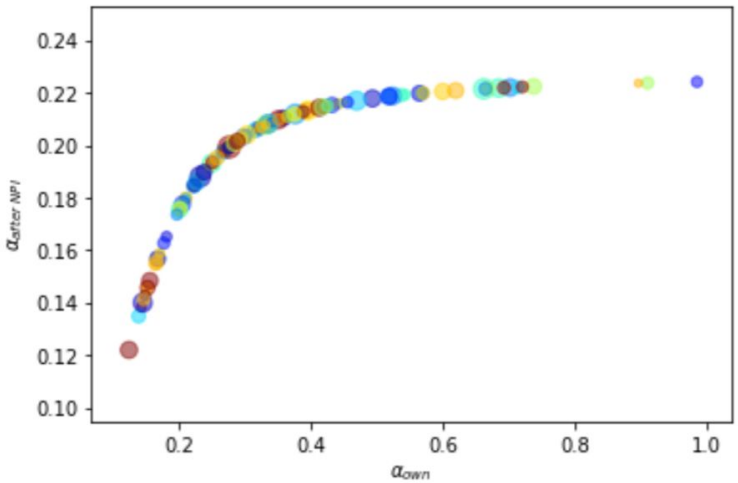


Figure 5. Alpha values vs. alpha values after NPI with policy 1

Analysis

Total Difference in Recovered: -329893

We give each node the same amount of money which makes it so that there is less bias involved.

However, one may also have to consider the fact that some of the nodes may need the money more than others, for example places that are already doing well may not need the money.

(b/c)

Policy number 2: spend the budget proportional to each node's population

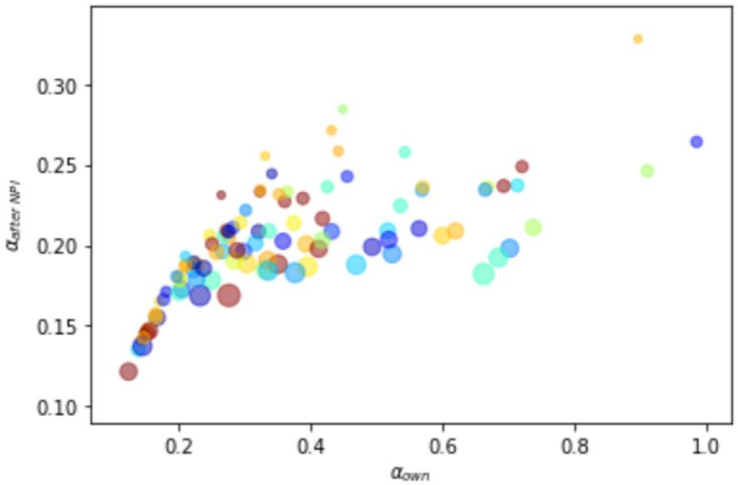


Figure 6. Alpha values vs. alpha values after NPI with policy 2

Analysis

Total Difference in Recovered: -346022

Spending according to the population of the node can be seen through the lense of fairness because it is a logical assumption to assume that in areas that are more densely populated more money may be needed in order to help the population at large.

On the other hand, it is possible to see it from the perspective that an area with a smaller population may need more resources than an area with a greater population.

(b/c)

Policy number 3: the change in alpha induced by the NPI is the same for all nodes

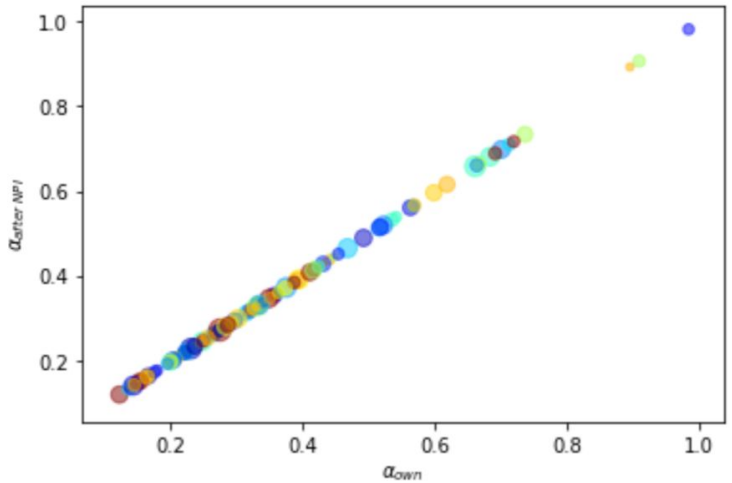


Figure 7. Alpha values vs. alpha values after NPI with policy 3

Analysis

Total Difference in Recovered: -7836

We are effectively offsetting the 'disadvantage' certain populations may have as a result of having a lower alpha value. In doing so, we level out the playing field.

(b/c)

Policy number 4: by equalizing the alpha values, our situation results in equity for everyone

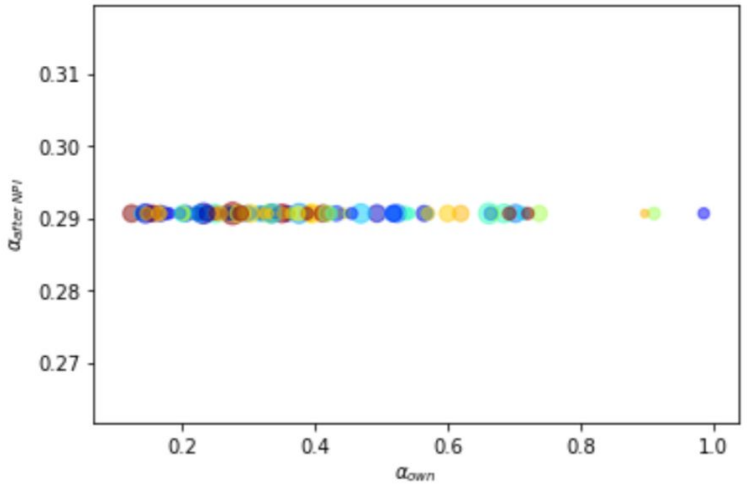


Figure 8. Alpha values vs. alpha values after NPI with policy 4

Analysis

Total Difference in Recovered: -113080

Regardless of N, v, and c values, we will have the same outcome for every node.

This can seem fair at first glance, but if you look closer you'll notice that there will undoubtedly be nodes that start off with a higher value. Thus, this is not an optimal way to disburse funding for the NPI, as many people will not be happy.

Future Work

Improvements

Although we have a pretty intricate model, there are many factors that we have not covered. For example, we would not know how effective our model would be in areas with varying levels of ICU capacity, or how different communities interact with each other. If we were given more factors to fine tune our solutions, we could probably create a very individualized resolution for specific cities.

However, our model would also not be terrible at predicting the spread of the disease over many different populations. It wouldn't be a definite answer for any one community, but as a general indicator it may provide some insight.

Summary

We achieved what we expected to achieve at the beginning of the quarter. From exploring the SIR model, to better understanding how to model the effects of human behavior during a pandemic, we were able to see the impact of something taking over the world right now.

As we all had a pretty solid understanding of coding, that part came easier to us. We also worked together really well so it made overcoming any walls we ran into a lot easier.

We struggled more with understanding the underlying concepts behind our code. It took us a little longer to understand some of the math that went along with modeling the SIR model, but over the course of a few days we were able to figure it out.

Coordination

As a team, we would hop on zoom calls and work through each part of the project together. Most of the time, we would end up spending multiple days across a week or so working on each part. Outside of our meetings, we'd research independently into the parts we got stuck on and go over them when we reconvened. Ultimately, our collective ability to do what was needed contributed the most to our success.

Future Work

If we were to apply our learnings and we were able to collect similar measures for how people are infected by covid, we might be able to apply the knowledge we have and help certain communities experiencing outbreaks.

One of our biggest challenge would be the actual retrieval of data and dealing with noisy data because we would need census data to understand factors such as social vulnerability and comorbidity, and we would also need specialized equipment to collect information on the viral load density.

References

[1] JA Patel, FBH Nielsen, AA Badiani, S Assi, VA Unadkat, B Patel, R Ravindrane, and H Wardle. Poverty, inequality and covid-19: the forgotten vulnerable. Public Health, 183:110, 2020.

[2] Emily E Wiemers, Scott Abrahams, Marwa AlFakhri, V Joseph Hotz, Robert F Schoeni, and Judith A Seltzer. Disparities in vulnerability to severe complications from covid-19 in the united states. Technical report, National Bureau of Economic Research, 2020.

Thank you for your time!