Estimating the True Number of COVID-19 Cases in the US

Ethan Greenberg, Daphne Poon, Sabrina Shen, Will Warfield

May 15 2020

1 Introduction

A large hurdle in the fight against COVID-19 is the simple lack of available data on the number of individuals that have been exposed to the virus. It is important to know how many people have been sickened or exposed to COVID-19 to better understand how lethal the virus is, who is most susceptible to it and to make decisions about when it is safe to return to normal life. In recent weeks, literature suggests that the virus may have arrived in the US long before was initially thought. The team desired to apply some of the data analytics tools they learned about in the course to existing COVID-19 datasets to gain a better understanding of how many people have been exposed to the virus in the US before social distancing measures were put in place. Specifically, the team used a coupled DEs model and an error function to optimize model parameters and fit the existing datasets. The team fit the model to data on test results for different countries up to the time that each country instituted social distancing. Finally, the model was used to predict the number of cases in the US before social distancing measures were introduced, assuming different initial viral arrival dates.

2 Methods

The DEs model the team used is the susceptible, exposed, infected, recovered – or SEIR – model. This model is commonly used to understand and model the spread of diseases and there is a lot written about it. The model is defined by 3 parameters that the team needed to determine. To begin, the team carefully considered what data to use to determine parameters for the model. Ever conscious of the principle "garbage in, garbage out," the team used data for countries with the highest testing rates to determine model parameters. The team selected Germany for fitting the model.

To optimize the parameters for the model, the team used a built in package with the Adam optimization algorithm to perform gradient descent. The team used log squared error for the error function. To calculate the derivatives, the team used local adjoint sensitivity analysis with numerical differentiation enabled. With the parameters selected, the team had a model

to use for making predictions about US case load.

To test the performance of the model, the team examined the performance of the model against Israel's data, as Israel also tested well.

The team determined a handful of potential US virus arrival dates. Then the team used the model and these various start dates to estimate the number of cases active in the US on March 22 – a date selected for the start of widespread social distancing in the US. The team initialized the models on these starting dates with 1 infected and 1 exposed persons.

3 Results

We were able to fit the data to the German growth curve quite well. Due to concerns about fitting our data to small numbers in the initial stages of the virus, we trained from the period in which the infected population reached 500 up to when widespread social distancing was implemented. **Figure 1** shows the result of this training.

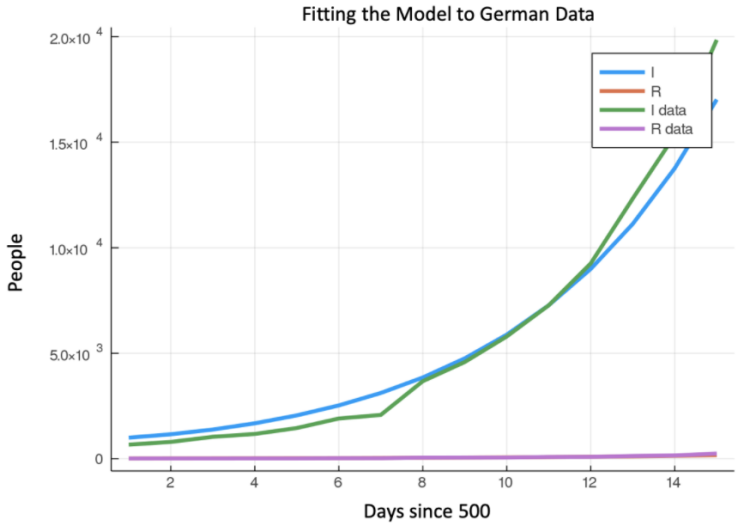

Figure 1: Fitting the SEIR Model to German data. I=infected, R=recovered

As discussed above, we tested the performance of our model on Israel. While the infected Israeli population remains below 500, the early data still provides a viable grounds for testing our model, even if it might not be as suitable for fitting our model parameters. **Figure 2** shows the resulting comparison. The cost function computed against the Israel data was .019. However, one must keep in mind this is partly smaller than the cost function on German data, because the model and data take on smaller values, generating smaller differences.

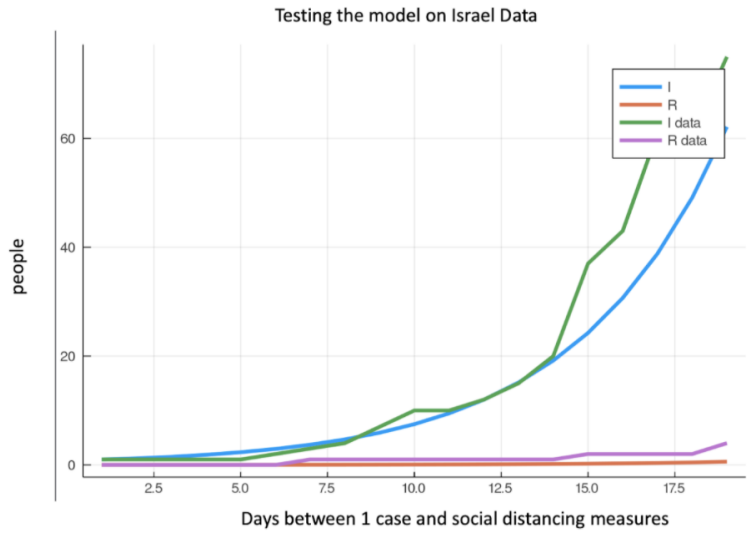

Figure 2: Testing the model against Israeli data. I=infected, R=recovered

Next the team proceeded onto the predictive part of our modeling. We used the model to estimate how many cases the United States would have when we implemented widespread social distancing, assuming that the virus spread unchecked since Jan 22 (**Figure 4**) and Jan 1 (**Figure 3**).

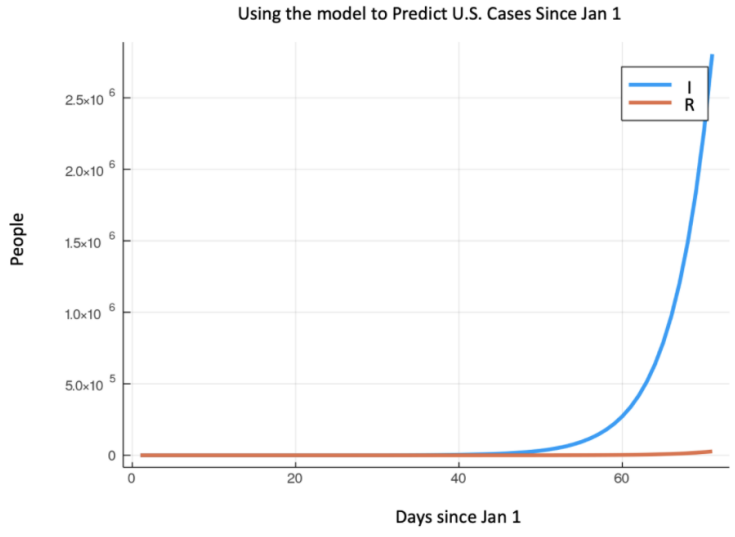

Figure 3: Predicted Cases with COVID-19 spreading unchecked from Jan 1 to Mar 22. I=infected, R=recovered

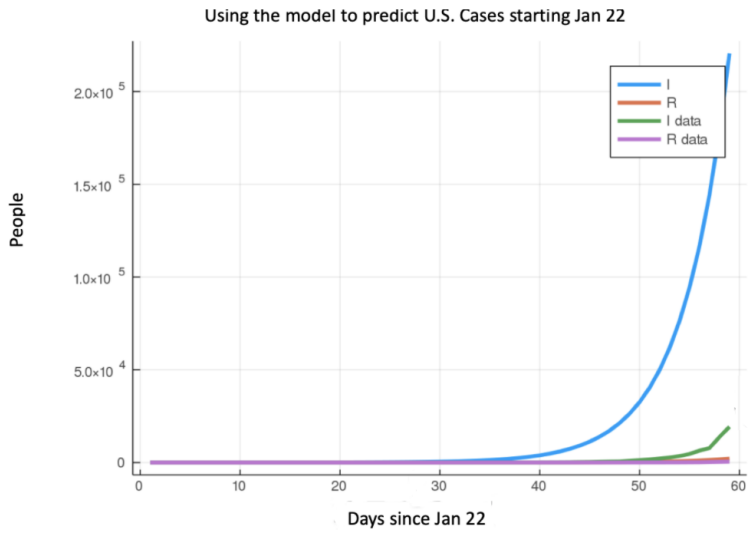

Figure 4: The data together with our predictions, assuming Covid began spreading in the U.S. on Jan 22. I=infected, R=recovered

The **Figure 4** graph also shows the actual data for comparison. If COVID-19 had spread unchecked in the United States since Jan 22, according to our model, 2.5×10^5 people would've been infected when we began social distancing (10 times more than the confirmed number of cases). If COVID-19 began spreading on Jan 1, according to our model 3.1×10^6 people were infected when we began social distancing (100 times more than the confirmed number of cases).

4 Discussion and Conclusion

The greatest obstacle in developing a more satisfactory model is the lack of reliable data that can be used as input to the system. As the epidemic continues to progress, researchers increasingly need more accurate data in order to make accurate predictions and assessments, however there seems to be a lot of conflicting information among published data. For example, there have been studies that indicate that the percentage of asymptomatic individuals within the infected group could be as low as 4% [1] while other sources suggest that up to 80% of infections are mild or asymptomatic [2]. Although this is just one out of many factors in the spread of COVID-19, the drastic difference in numbers are representative of the overall level of uncertainty in published statistics.

In order to try and find countries with the most reliable information to train the model, the number of COVID-19 tests done per capita was used as a proxy to accuracy. As a result, 25 countries were found to have a higher testing rate than America: Australia, Austria, Bahrain, Belgium, Canada, Republic, Czech, Denmark, Estonia, Germany, Iceland, Ireland, Israel, Italy, Latvia, Lithuania, Luxembourg, New Zealand, Norway, Portugal, Qatar, Slovenia, Spain, Switzerland. Since the goal of the SEIR model is to predict the spread of COVID-19 in America before social distancing was implemented, only data between the first

detected case of COVID-19 in the country to the first mandate of social distancing can be accepted as inputs to the model.

Perhaps unsurprisingly, almost all of the countries that have a higher rate of virus testing developed their first case of COVID-19 relatively recently and were much quicker to implement social distancing measures. Therefore, data from many of these countries was too short in duration to adequately train the SEIR model. Additionally, data before a country reached an infected population of 500 was also eliminated to avoid initial testing errors, further limiting the available pool of data. After attempting to train the model on shorter datasets, the results showed that a model run on Germany's dataset exclusively performed comparatively well. In order to evaluate the performance, the model was used to predict the infection and recover rate in Israel, which is also listed as one of the countries with higher testing rates, by using the date of the first confirmed case in Israel as the basis of the simulation.

When implementing this model to estimate the number of infected people in America, the date of the first confirmed case [3] used was January 22. Based on this information, the model showed that the predicted number of Coronavirus Cases in the United States was nearly 10 times the reported cases at the time. As mentioned in other papers that have attempted to model the COVID-19 spread [4], an extremely accurate model much will be difficult to create until more data is available or, in the worst case, after the pandemic is over and a retrospective count is published. While taking into account the margin of error of the simulated values, the model still suggests that America had severely underestimated the number of COVID-19 cases.

References

[1] Zhou X et al., Follow-up of asymptomatic patients with SARS-CoV-2 infection, Clinical Microbiology and Infection, https://doi.org/10.1016/j.cmi.2020.03.024.

[2] "Q&A: Influenza and COVID-19 - similarities and differences". *World Health Organization*, 17 March 2020, www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-similarities-and-differences-covid-19-and-influenza.

[3] Holshue, Michelle L. and DeBolt, Chas and Lindquist, Scott and Lofy, Kathy H. and Wiesman, John and Bruce, Hollianne and Spitters, Christopher and Ericson, Keith and Wilkerson, Sara and Tural, Ahmet and Diaz, George and Cohn, Amanda and Fox, LeAnne and Patel, Anita and Gerber, Susan I. and Kim, Lindsay and Tong, Suxiang and Lu, Xiaoyan and Lindstrom, Steve and Pallansch, Mark A. and Weldon, William C. and Biggs, Holly M. and Uyeki, Timothy M. and Pillai, Satish K., "First Case of 2019 Novel Coronavirus in the United States", New England Journal of Medicine, 382, 10, 2020, 929-936. https://doi.org/10.1056/NEJMoa2001191.

[4] Fairoza Amira Binti Hamzaha, Cher Han Laub, Hafeez Nazric, Dominic Vincent Ligotd, Guanhua Leee, Cheng Liang Tanf, Mohammad Khursani Bin Mohd Shaibg, Ummi Hasanah Binti Zaidonh, Adina Binti Abdullahi, Ming Hong Chungj, Chin Hwee Ongk, Pei Ying Chewl and Roland Emmanuel Salungam, "CoronaTracker: World-wide COVID-19 Outbreak Data Analysis and Prediction", CoronaTracker Community Research Group, 19 March 2020. www.who.int/bulletin/online_first/20−255695.pdf.