COVID-19 FORECASTING

ELIAS GHANTOUS , EFRAN HIMEL , DERRICK LIU , TIFFANY QUANG , AND AUBREY WIEDERIN

Abstract. The most common model that is used for modeling the progression of infectious diseases is the SIR model. This model tracks the number of susceptible, infected, and recovered individuals in a closed population. Because this SIR model is for a closed population, it does not take into account immigration and emigration as well as geographic components. To address the weaknesses of the basic SIR model, we decided to build our own SIR model, and we hypothesized that our model would provide more accurate forecasting of the morbidity and mortality due to COVID-19 than the basic SIR model would provide. In approximating the rate of a susceptible-infected contact resulting in a new infection and the rate that an infected recovers and moves into the resistant phase, the predicted values that the neural nets produced had very high residuals, so it was difficult to reintroduce the predictions back into our SIR model.

Key words. COVID-19, SIR model, Multiple Linear Regression, Neural Nets

The paper is organized as follows. Our methods are in section 2, our main results are in section 3, and the potential issues and summary of our project follow in section 4.

1. Introduction. The coronavirus pandemic has been the most central topic in the world over the past few months. It has impacted every country in the world, and the cases and deaths due to this virus continue to rise. The United States of America has become an epicenter of this pandemic, and as of April 10, COVID-19 has infected 460,287 people and has killed 16,535 deaths in America [2]. With such a devastating tragedy at the center of our lives right now, our group took a tremendous interest in devoting our project to building a model that would help forecast COVID-19 in the US.

Different countries have responded to the outbreak of COVID-19 in different ways. South Korea was one of the first countries that was able to effectively contain the spread of COVID-19 by installing public phone booths outside of hospitals. At these phone booths, people getting tested for the virus are on one side of the glass, and a

hospital worker is on the other side. The hospital worker can swab the patient quickly, disinfect the booth, and get a result within seven minutes. These phone booths allow almost ten times as many samples to be tested. In Senegal, as of March, hospital workers were using test kits to diagnose patients within 4 hours, which is faster than the kits in the U.S. were able to make diagnoses. In comparison to South Korea, countries like Italy and the U.S. responded too moderately and set social distancing measures once the virus had already spread too much [1]. Social distancing measures have been especially important to implement because COVID-19 is primarily spread directly through contact with infected people (in the form of respiratory droplets) or indirectly with surfaces that have been touched by infected people [3].

2. Methods.

2.1. Building the Database. Our initial step for our project involved gathering data regarding the coronavirus as well as various metrics from each state we believed would have some relationship with the incidence and deaths due to the virus. Fortunately, a reliable GitHub repository had already aggregated all the individual state COVID 19 reporting into well organized CSV files, and we incorporated the data in this repository for our project's purposes. After conducting some research, we also found valuable datasets on state population and density, transportation data, population based age categories, and number of physicians and hospitals per state from various reliable sources such as the U.S Census Bureau and the Kaiser Foundation. After our research, we created a database using MySQL that combined all our aggregated datasets into one master table, which we would use for our project. This table was regularly updated with new information, and we used GitHub to store and share the information with one another.

2.2. Multiple Linear Regression. We used multiple linear regression to try to find predictor variables in our data set that we could use to plug into our neural nets. For the regression, we checked the no multicollinearity assumption in order to remove predictor variables that were highly correlated with one another from our model. We also checked the linearity assumption by plotting the outcome variables (number of

COVID-19 deaths and number of COVID-19 cases) against each of the individual predictor variables, and we lastly checked the normality assumption by plotting the residuals for each predictor variable on qq-plots and looked for a linear trend.

2.3. Naïve β Estimation. In the basic, analytic SIR model, the total number of cases follows this differential equation:

$$\frac{dC}{dt} = \frac{\beta IS}{N} = \frac{\beta I(N-C)}{N}$$

where C is the cumulative number of cases, N is the total population, I is the infected population, and S is the susceptible population. Solving for β produces

$$\beta = \frac{1}{I(1-C/N)}\frac{dC}{dt}.$$

Thus, given the population and time-series data of the number of cases and currently infected, β can be estimated day-to-day.

For real data, there is the issue of under-reporting due to less-than-perfect coverage in testing. While this would be best modeled using arithmetic over negative hypergeometric distributions, we found the point estimates of just multiplying by the appropriate ratio to be sufficient.

2.4. Basic SIR Model. We wanted to find and further develop a basic SIR model to help us understand the dynamics of COVID-19. Due to the restraints of the model, we were only able to run simulations on closed populations based off a time-series database. With the help of our SIR model, we were able to run simulations to help us predict the number of susceptible, infected, and recovered cases in the United States for the next 150 days from the start date of January 22, 2020. We were not completely satisfied with the outputs of our basic SIR model, so we determined that it would be best to adjust the basic SIR model in order to omit vital dynamics such as birth and death.

2.5. Neural Nets. While half of our group continued conducting our initial regression analysis and developing the basic SIR Model, the other half focused on

creating the foundations for a neural net model and our new SIR Model. For the neural nets, we decided to use existing neural net libraries for our project rather than implementing all aspects of a neural net from the ground up to reduce the overhead on our project and to abstract away many of the complicated mathematics and programming requirements for implementing some machine learning framework from scratch. Our library of choice was TensorFlow for python with the Keras API due to their numerous resources and documentation that was openly available to use. We began our neural net implementation by developing a quick binary classification model to confirm that using TensorFlow would be an efficient tool to conduct the rest of our project. We followed some simple tutorials online and found the ease of using TensorFlow for our project was acceptable and that we would proceed in incorporating it into our project. We continued by creating a new neural net model that would take in our master database, and we used the state data inside the database to predict various COVID outcomes such as the number of positive cases and deaths per state.

2.6. Our SIR Model. The basic SIR model does not account for geography, and so predicts smooth, sinusoidal growth. In contrast, as the actual virus spreads to new regions (cities, states, countries, etc.), it causes sinusoidal "jumps" in the total number of cases as a new population is infected. We modeled this as a set of coupled SIR models on a graph, with the coupling being transport of infected and uninfected between regions.

2.6.1. Simulation. To start, initial susceptible, infected, and recovered/dead populations were set for each location, and an infinitesimal transport matrix set such that the total population at each location is constant when transported.

For each timestep, the new population per location was set by a multinomial distribution with probabilities proportional to the exponentiated transport matrix applied to the previous populations, effectively simulating travel according to Poisson distributions. For each location, the number of new infections was chosen according to a binomial distribution, with n equal to the susceptible population, and $p = 1 - e^{-\frac{\beta I}{N}\Delta t}$. While technically a more complicated distribution closer to a hypergeometric

distribution would be more correct, this is correct in the limit of large N, and much easier to sample from. From the infected, the number of newly recovered or dead is then chosen according to another binomial distribution, with $p = 1 - e^{-\gamma \Delta t}$. Finally, the number of deaths is chosen as a fraction of these, according to yet another binomial distribution, this time with a constant probability equal to a given Infection Fatality Rate (IFR).

2.6.2. β Estimation. We took a somewhat Bayesian approach was taken for estimating β using our model.

Firstly, there is the problem of estimating the true data (susceptible, infected, etc) given the reported data. Given the true data, the reported would follow a hypergeometric distribution, and so the distribution of the true given the reported can be expected to follow a negative hypergeometric distribution. For ease of computation this was approximated by negative binomial distributions instead, which is accurate in the limit of a large unexposed population.

For a binomial distribution with $p = 1 - e^{-\lambda}$, the Jeffreys prior can be calculated to be $f(\lambda) = \frac{1}{\pi\sqrt{e^{\lambda}-1}}$, which we used as a very rough estimate of the Jeffreys prior for β, and sampled from. This was done by applying its quantile function $g(u) = -2\log(\cos(\frac{\pi}{2}u))$ to equally-spaced values on $[0, 1]$.

What was then needed was to calculate the likelihood of the data given the different values of β. To do this, for each time except the last, the next was simulated, estimating the appropriate conditional distributions. In order to manage the dimensionality of the problem, the different locations were considered as to be were independent. Likewise with the infected and recovered groups; the susceptible group was disregarded since it is by necessity linearly dependent on the other two.

After normalization this produced a distribution over β, from which the mean and standard deviation could be extracted.

3. Results.

3.1. Multiple Linear Regression. After checking the no multicollinearity assumption and removing the highly correlated variables from the multiple linear re-

gression model, we found that the only statistically significant coefficients were those for the population density and number of airports predictor variables. As expected, the coefficients for population density and number of airports were positive (21.4757 and 12.5072) for the model predicting COVID-19 cases and were positive (0.7089 and 0.3736) for the model predicting COVID-19 deaths. As population density increases, there are likely more cumulative interactions between humans, and as a result, the transmission of the virus from the infected to the susceptible is easier, increasing the number of COVID-19 cases and potential deaths. Similarly, as the number of airports increases, the amount of interstate travel in the U.S. increases, so it is more likely that someone from New York (a state with the most COVID-19 cases and deaths) comes into contact with someone from Alaska (a state with only a few hundred COVID-19 cases). Thus, the transmission of the virus becomes easier, increasing the number of COVID-19 cases and potential deaths. After checking the linearity and normality assumptions, the only variable left was the number of airports. The number of airports was the only predictor variable that satisfied the three assumptions.

3.2. Naïve β Estimation. The time-series data made available by Johns Hopkins at [8] was used for this estimation. The ratio of actual deaths to reported, $1 + \frac{5126}{13156} = 1.39$, was estimated using the data available at [5]; the estimate of Case Fatality Rate, or ratio of deaths to confirmed cases, was taken from [6]; the estimate of Infection Fatality Rate, or ratio of deaths to actual infections, was taken from [7].

The true number of cumulative infections was estimated by $Confirmed \times \frac{CFR}{IFR}$; the number of current infectious was estimated as $Cumulative - \frac{Deaths}{IFR}$.

For each day for which there was a nonzero infected population, $\frac{dC}{dt}$ was estimated using second-order central differences, or first-order one-sided on the boundaries. Excluding any negative estimates as unphysical, the results were found to be best plotted on a log scale. Thus the mean and sample standard deviation of $\log(\beta)$ were computed, with values $-0.957(30)$ and 1.146 respectively. The estimates and a normal curve with the same mean and variance are plotted in figure 1.

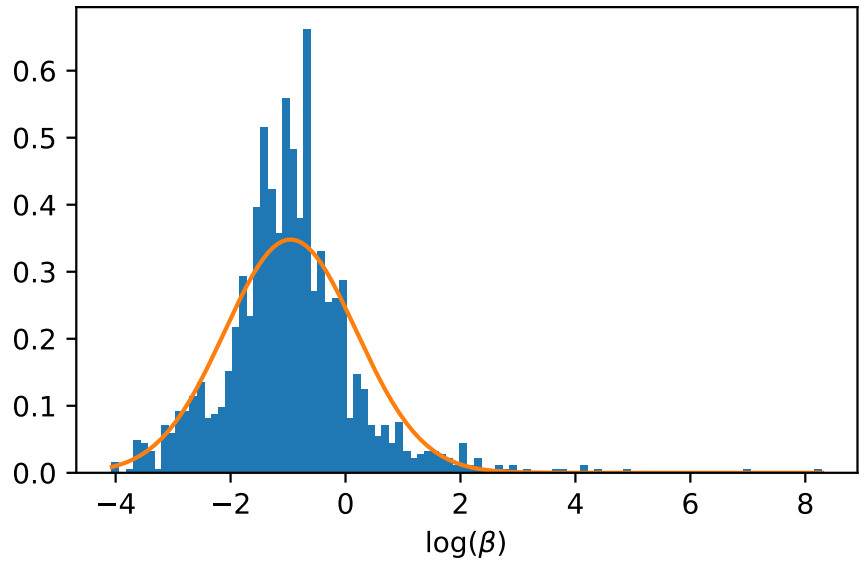

FIG. 1. *Aggregate estimated values of β for the 50 states from 01/22 to 04/29.*

3.3. Basic SIR Model. After omitting the dynamics of birth and death from our basic SIR model, we found that the basic SIR model was able to predict the parameters for the United States more accurately. Our model predicted a beta (the infection rate of the disease) value of 0.1496 and a gamma (the recovery rate) value of 0.002533. Our gamma value seems to be an underestimate, which is most likely because of the sparsity of recovery rate data. Due to our gamma value being lower than expected, the model produced a reproduction rate (R_0) of approximately 59.078, which is approximately ten times higher than it should be.

3.4. Neural Nets. Our initial plan was to perform predictions for several time frames to get an idea of the rate of infection and rate of deaths per state to then feed into our SIR model. Unfortunately, after some initial training runs, we encountered a major roadblock: the predicted values in our trained neural nets had extremely high residuals. We encountered residuals as high as 18,000 for the mean absolute error for estimated positive cases, and these high residuals prevented us from proceeding further onto our final goal of estimated infection and death rates. We tried different approaches to optimizing our neural nets, most of them by trial and error and cycling

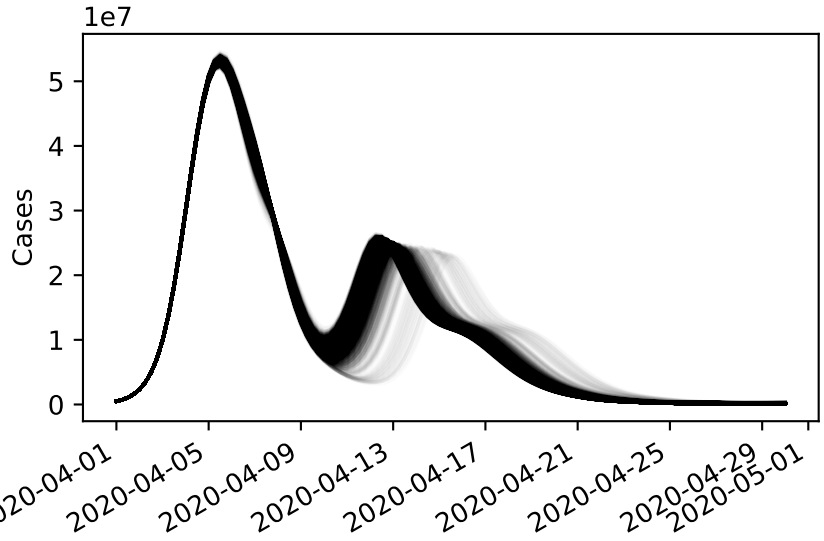

FIG. 2. *An example set of simulated trajectories of the total infected population, using placeholder parameters.*

from different iterations ranging from models with very few nodes with only one hidden neural net cluster to hundreds of nodes with multiple hidden clusters. Eventually, after these iterations were performed, we found a neural net with one hidden cluster and a moderate number of nodes, 64 per cluster, to be acceptable because it allowed us to reduce our residuals to around 5,900 with a testing training split of 80/20. We found this residual to be acceptable because we realized that many states in the U.S. had been under-testing their populations, so the true number of actual cases was likely much higher than reported. Unfortunately, by the time we realized this, our presentation deadline was approaching, so instead of refining our model even further to eventually predict the infection rate and deaths, we instead proceeded to focus our attention in completing our SIR model.

3.5. Our SIR Model.

3.5.1. Simulation. Unlike the basic SIR model, our model is able to exhibit "jumps" in the infection curve as a new location becomes exposed, as shown for example in figure 2; in this respect the model is successful.

In contrast, using the naïve estimates of the parameters produces a more tame

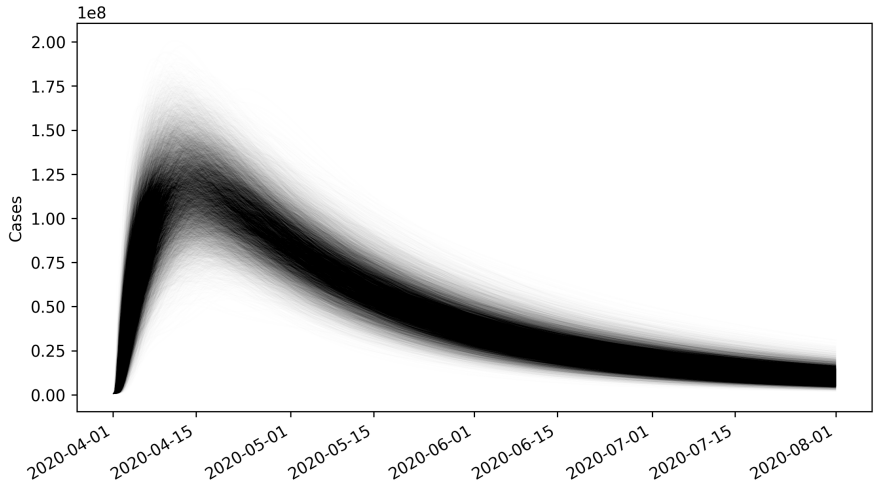

FIG. 3. *A set of simulated trajectories of the total infected population, using the naïvely-estimated parameters.*

curve, as seen in figure 3. This is likely due to the placeholder data having only 4 locations instead of 50.

3.5.2. β **Estimation.** Unfortunately this calculation is yet in progress, and so cannot be reported on.

4. Issues and Summary. We will outline a few potential issues that need to be resolved so that our model can be more accurate. The first issue we will discuss is that our model does not account for many of the factors that contribute to the spread of COVID-19, such as the effects of social distancing. Secondly, we found that multiple linear regression does not work well for COVID-19 forecasting. A potential solution would be to use logistic regression with a binary variable for survival/death from COVID-19. Lastly, in order to have an accurate model, we need reliable data. The data we collected were from various reliable sources. A background check on the data and how it was collected could help improve our model. While our model does have some issues, it was still helpful in making some significant findings. Through our regression model, we were able to find that the number of airports and population density supposedly play an important role in predicting the number of deaths and total cases due to COVID-19. We also found that neural nets have a lot of potential to be useful in forecasting COVID-19. Through our own SIR model, we found that

compared with a basic SIR model, our model shows jumps in the infection curve as a new population is infected. The main limitation for our SIR model was the inaccuracy of our neural nets. For future models, we will have to reconfigure our neural nets in order to fully explore the SIR model. Moving forward, we will estimate the infection and recovery rates for the model with our current neural nets and compare to real world outcomes. We will also create a secondary model that uses actual data provided by state databases to create another SIR model based on real world values instead of predicted values. As the world continues to fight against the COVID-19 pandemic, reliable statistical models are needed to accurately predict the spread of infection. Accurate predictions will help inform world leaders to make the proper decisions for their respective country. With the improvements listed, our hope is that the model we created can be useful in aiding in the fight against COVID-19.

REFERENCES

[1] Frank, A., & Grady, C. (2020, March 22). Phone booths, parades, and 10-minute test kits: How countries worldwide are fighting Covid-19. Retrieved from https://www.vox.com/science-and-health/2020/3/22/21189889/coronavirus-covid-19-pandemic-response-south-korea-phillipines-italy-nicaragua-senegal-hong-kong.

[2] Lutton, L. (2020, April 10). Coronavirus case numbers in the United States: April 10. Retrieved from https://www.medicaleconomics.com/news/coronavirus-case-numbers-united-states-april-10.

[3] Modes of transmission of virus causing COVID-19: implications for IPC precaution recommendations. (n.d.). Retrieved from https://www.who.int/news-room/commentaries/detail/modes-of-transmission-of-virus-causing-covid-19-implications-for-ipc-precaution-recommendations.

[4] Sasaki, K. (2020, March 11). COVID-19 dynamics with SIR model. Retrieved from https://www.lewuathe.com/covid-19-dynamics-with-sir-model.html

[5] COVID-19 Data. (n.d.). Retrieved May 04, 2020, from https://www1.nyc.gov/site/doh/covid/covid-19-data.page

[6] Coronavirus Mortality Rate (COVID-19) - Worldometer. (2020, May 5). Retrieved from https://www.worldometers.info/coronavirus/coronavirus-death-rate/

[7] Russell, T. W., Hellewell , J., Jarvis , C. I., van Zandvoort , K., Abbott , S., Ratnayake , R., ... Kucharski , A. J. (2020, March 26). Estimating the infection and case fatality ratio for coronavirus disease (COVID-19) using age-adjusted data from

the outbreak on the Diamond Princess cruise ship, February 2020. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7118348/

[8] COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. Retrieved April 29, 2020, from https://github.com/CSSEGISandData/COVID-19

[9] 2020 World Population by Country. (2020, May 10). Retrieved from https://worldpopulationreview.com/

[10] C. (n.d.). COVID19Tracking/covid-tracking-data. Retrieved May 10, 2020, from https://github.com/COVID19Tracking/covid-tracking-data/

[11] Population Distribution by Age. (2019, December 4). Retrieved from https://www.kff.org/other/state-indicator/distribution-by-age/?currentTimeframe=0sortModel=

[12] Professionally Active Physicians. (2019, March 8). Retrieved from https://www.kff.org/other/state-indicator/total-active-physicians/?currentTimeframe=0sortModel=

[13] United States Airports. (n.d.). Retrieved May 10, 2020, from https://www.globalair.com/airport/state.aspx