Predicting the Impact of COVID-19

William Austin
waustin3@gmu.edu

Prakash Dhimal
pdhimal@gmu.edu

May 20, 2020

Abstract

The coronavirus (COVID-19) outbreak caused by the 2019 Novel Coronavirus (2019-nCov) has become a global pandemic, with more than 4 million people infected, and hundreds of thousands of deaths around the globe. Countries across the globe are taking action to control the spread, but there is significant uncertainty about what the future holds. Predicting when the spread of the virus will peak using compartmental modeling in epidemiology has been a huge focus among government, academic institutions, and the private sector. In this paper, we discuss the basic compartmental model in epidemiology, the SIR model, and use data mining techniques to fit the SIR model curve with COVID-19 data. Our goal is to forecast the spread of COVID-19 confirmed cases, fatalities, and recoveries in different countries around the world.

0.1 Summary of the report

The structure of this paper is as follows. In section 1, we introduce COVID-19 and the basic compartmental model in epidemiology, the SIR model. In Section 2 we briefly discuss related work. In section 3, we present the solution to our problem. In Section 4.1, we discuss the data that we used for this project. Section 4.2 dives into the set-up of the SIR model. In section 5, we discuss our results and give our analysis on those results. In section 6, we discuss the evaluation metrics used to evaluate the performance of our model. Section 7, concludes the paper.

1 Introduction

The coronavirus (COVID-19) outbreak, also known as COVID-19 was first reported by Wuhan Municipal Health Commission, China on 31st December 2019 and has since spread to the vast majority of countries. The World Health Organization (WHO) declared COVID-19 as a Public Health Emergency of International Concern (PHEIC) on January, 30th 2020 [4]. Since then, the outbreak has become a global pandemic, with more than 4 million people infected, and hundreds of thousands of deaths around the globe. COVID-19 is not only causing mortality but is also putting considerable stress on the health systems of countries across the globe. Countries across the globe are taking action to control the spread, but there is significant uncertainty about what the future holds. As shown in figure 1, COVID-19 cases in the United States have been hovering around all-time highs, while the country and specific states are getting ready loosen social distancing measures put in place to control the spread of the disease. This project's goal is to use the principles of data mining and mathematical models of epidemic spread to forecast the spread of COVID-19 cases, deaths, and recoveries. This paper aims to deliver an overview of the SIR model and the outcome of our simulation by using the dataset of COVID-19.

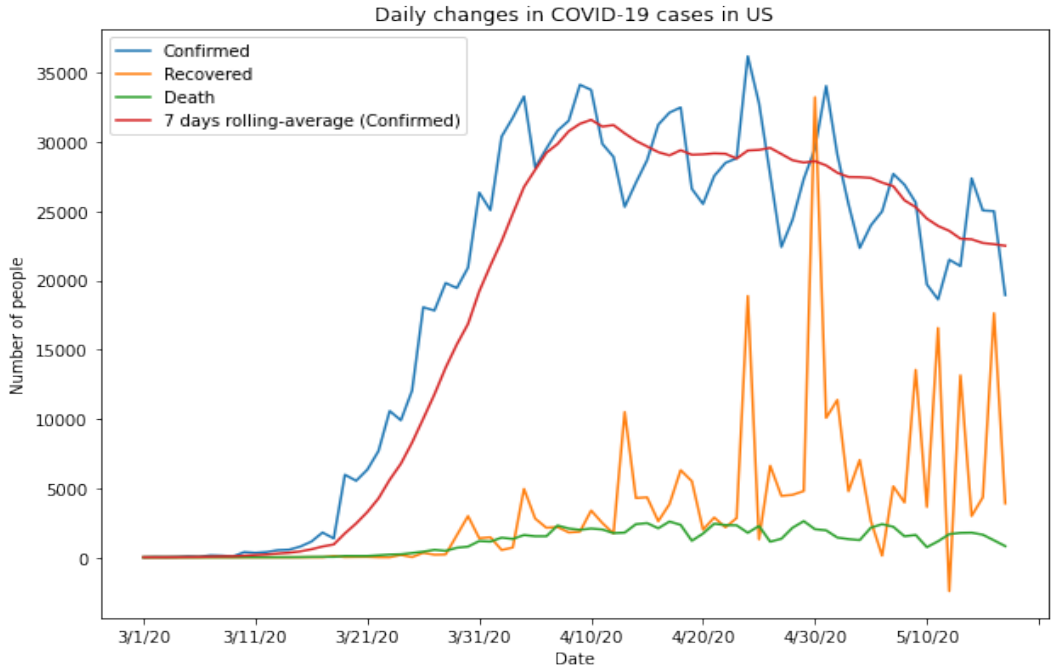

Figure 1: Daily changes in COVID-19 cases in the US (03/01/2020 - 05/17/2020)

2 Related work

In recent weeks, academic institutions, government, and industry researchers have put in a lot of effort into using Artificial Intelligence, Machine learning, and Data mining techniques to model the spread of the Novel Coronavirus. There is a large body of research and data around Covid-19 posted on Kaggle and other platforms to help scientists collaborate on this problem at unprecedented levels. Governments and organizations involved in COVID-19 response across the globe are adopting data mining techniques to assist in analyzing the spread of the virus.

The Institute for Health Metrics and Evaluation (IHME) [1], developed a curve-fitting model that looks at how the disease progressed in other geographies, like China, Italy, and Spain, and tries to extrapolate a prediction from there. IHME's model mainly uses statistical model estimation for population death rate and health service utilization.

An excellent introduction to the most prominent models being currently developed for projections in the US is given by FiveThirtyEight [6]. Current projections and a discussion of the models from IMHE, MIT, University of Massachusetts, Columbia University, University of Texas, and others are included. While most organizations are utilizing a compartmental model to generate predictions, there are major differences in the design of the models, parameters, and data sets being incorporated. A few of the highlights are:

- The University of Texas model makes an effort to incorporate the effect of social distancing and uses different assumptions than IMHE to fit the model to the data. Social distancing metrics based on information reported by mobile devices are used.
- The MIT model is based on an SEIR model and tries to quantify the effects of government

interventions.

- The University of Massachusetts model is based on a Bayesian compartmental model and makes predictions based on the prior distribution of the data.
- The Columbia University model is based on county-level data from across the US and hospital bed counts. Its projections are focused on generating several future scenarios at varying confidence levels.

As we see, the vast majority of the COVID-19 scenarios and forecasts have been based on mathematical compartmental models that capture the probability of moving between Susceptible, Infected, and Recovered (SIR) states. While any model is only as good as the assumptions we make, these models, in particular, are sensitive to assumptions and the results differ considerably. In our work, we use a curve-fitting model to uncover the parameters that are used in this model and have used official sources like the WHO and the CDC to make assumptions where needed. We call this the **curve fitted SIR model**. Later in this report, we will give an overview of compartmental models and in particular, we will show the equations and intuition describing the behavior of the SIR model.

3 Solution

Our analysis of this problem begins by looking at the existing work done in this area, as described in the previous section. Additionally, we considered the data sources that were available to us for experimentation. The Johns Hopkins University COVID-19 Data Repository contains daily values for confirmed cases, recovered individuals, and deaths, broken down by country and region. With access to this data, we were able to plot it and see how the data resembles the output of a Suspected-Infected-Removed (**SIR**) model in many cases. We describe additional details about the data set later in this report.

The SIR model is one of the simpler approaches to looking at the spread of infectious diseases, so it became a very useful starting point for us. One of the main reasons for this simplicity is that there are only two parameters (β and γ) for the model, describing the rate of flow of individuals between the three states. This was advantageous for us because we can determine exact historical values for these parameters, based on the data set.

Our next activity was to graph the historical values for β and γ. After doing this, it became obvious that the true values will vary over time. However, this makes sense, because the actual values of these parameters are a reflection of what happens in the world, which is a result of human behavior, and can change over time. An example graph of the changes in the value of beta across the historical data range is shown below.

Based on these results, the remainder of our project work was aimed at generating predictions for what the future could look like by estimating the value of β and γ and applying the SIR model equations. Although these equations can be applied to any future time, the SIR model is extremely sensitive to these parameters, so we typically only generate 30 days of predictions.

By focusing on estimating the parameter values for the SIR model, we can try to attack it as a regression problem. In our case, we imagine that the true value of β on day t can be given by the

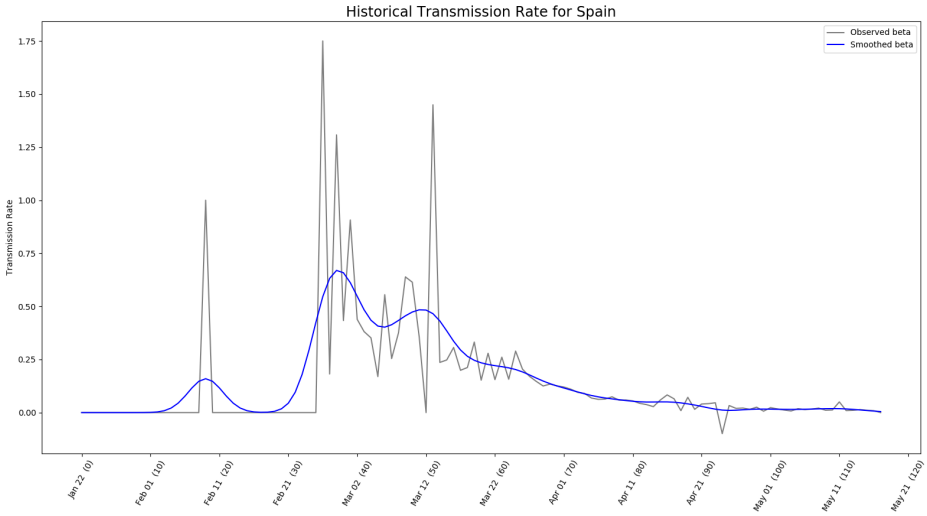

Figure 2: Historical Beta Values for Spain, with Smoothing

function $\beta(t)$. Likewise, we can also imagine that the true value of γ is represented by $\gamma(t)$ on day t. However, as the graph above shows, these functions do not fall into a clear category. Therefore, we tried several extrapolation techniques. This includes:

- Linear Regression
- Quadratic Regression
- Weighted Average of $\beta(t)$ values with Exponential Decay
- Weighted Average of $\beta'(t)$ values (Slope) with Exponential Decay
- Neural Network Regression

With this set of models, our next step was to incorporate the historical data to validate how accurate the predictions are for a given prediction range. We can then compare the errors for all candidate models, choose the one with the lowest error, and use it to make predictions for future time frames. More information about the details of this process is included in the rest of the report.

4 Experiments

4.1 Data

We use the Novel Coronavirus COVID-19 Data Repository from Johns Hopkins University's Center for Systems Science and Engineering (CSSE) [2] as our main data source. This data set provides total counts of how many cases have been confirmed, broken down by country and region/province. This is a time-series data and the number of cases on any given day is the cumulative number. These time-series data are for global confirmed cases, recovered cases, and deaths. Australia,

Canada, and China are reported at the province/state level. The US and other countries are at the country level. In addition, it is an aggregated data source, so it is based on contributions from organizations around the world, including the WHO (World Health Organization) and various other public organizations around the world.

This data is updated every day since COVID-19 cases are evolving daily in different parts of the world. We have collected the data up to the date of this report. Therefore, we are using partial data sets. However, we were able to build and test models based on data collected up to this point.

After acquiring the data from the data source, we represented the data as a matrix (table). We aggregated the data on the country level. We then removed the features that we didn't need for this analysis and created features that we needed. We implemented stratified sampling to split the time series data collected so far into the training set and testing set.

In addition to the daily data provided by JHU, for accurate predictions, we also need to know what the current population of each country or region in the data set is. To get these figures, we downloaded a simple data set from Kaggle [5] and modified it slightly so that the country labels match between the two data sources.

For each country, before any COVID-19 cases have been reported, all individuals are considered susceptible, so this population value is used in our calculations for the initial condition, $S(0)$. In addition, we also sometimes choose to scale the β variable by the population count to give a more literal interpretation of the constant.

4.2 Modeling infectious disease

4.2.1 SIR model

One of the best ways to model infectious diseases like COVID-19 is to use a compartmental model. A compartmental model separates the population into several compartments. The **SIR** model is one of the compartmental models used for modeling epidemics. In this model we divide the population (of a country, state, city, etc) into three groups:

1. **S**usceptible: This is a group of individuals that have not been infected with the disease at time (t). This means that this group of people are susceptible to the disease.

2. **I**nfectious: This group of individuals are currently infected with the disease at time (t) and are capable of spreading the disease to those in the susceptible category. Those Infected are expected to recover at a certain time in the future and gain "immunity".

3. **R**emoved: This group of individuals have already caught the virus, and cannot be infected again (Recovered), or unfortunately lost their lives (Dead). Here we assume that once you get the virus and recover from it, you gain some sort of immunity against the virus. In most SIR modeling, this group is referred to as Recovered. Since we are combining the Recovered and Death into the same group, we decided to call it Removed.

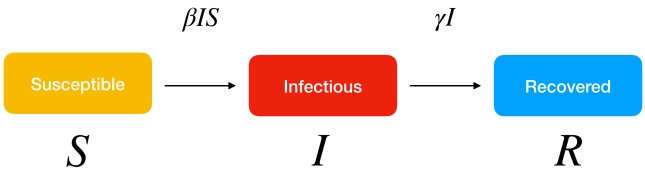

Figure 3: SIR Flow model [7]

4.2.2 Model parameters

To understand the SIR model, we need to define a few other parameters.

- **Transmission rate (β)**: This is the rate at which people are getting infected. In other words, it is the rate at which those in the Susceptible group turn move into the Infected group. This is also called this **infection rate**.
- **Recovery rate (γ)**: This is the rate at which people are recovering from the disease. In other words, it is the rate at which those in the Infected group move into the Removed group. The recovery rate depends on how many days the infection lasts. According to the World Health Organization, For COVID-19, 14 days is the maximum number of days for the incubation period. 5 days is the average incubation period [4]. Based on the number of days an infected person can spread the disease **D**, we can derive the recovery rate as:

$$\gamma = \frac{1}{D} \tag{1}$$

Given transmission rate (β) and D, we can derive the basic **reproduction number** R_0 as:

$$R_0 = \beta * D = \frac{\beta}{\gamma} \tag{2}$$

R_0 represents how many healthy people the average infected individual transmits the virus to. If it is high, the probability of pandemic is also higher.

We would also like to note that the values for β and γ and how they change are very relatable to the measures being taken to fight the pandemic. In particular:

- β can be thought of as a measure of how quickly the infected individuals spread the virus to susceptible individuals. We may not have any effect on some variables, like the virulence of the disease, but the effectiveness of human behaviors like social distancing and good hygiene will have a direct impact on the value of β.
- Factors that may influence γ include things like the availability of a vaccine or other treatments, hospital bed counts, ventilators, and Personal Protective Equipment (PPE) for health care workers.

At any given time (t) during the pandemic, we want to know the number of infected, susceptible, recovered, and fatalities. For this, we use ordinary differential equations to describe the rate of

change of each group in the SIR model. The SIR model can be expressed by the following set of ordinary differential equations.

$$\frac{dS}{dt} = -\frac{\beta * I(t) * S(t)}{N} \tag{3}$$

$$\frac{dI}{dt} = \frac{\beta * I(t) * S(t)}{N} - \gamma * I(t) \tag{4}$$

$$\frac{dR}{dt} = \gamma * I(t) \tag{5}$$

$$N = S + I + R \tag{6}$$

where **N** is the total population.

4.2.3 Modeling COVID-19

As we mentioned in our Solution section, we begin our analysis by starting with a basic SIR model and some historical data that can help us make predictions for the number of individuals in the S, I, and R groups at some point in the future. This gives us an idea of how the virus will spread in the coming months.

Our mechanism for creating these projections are based on the fact that the SIR model has two parameters (β, and γ), and we can use their historical values to anticipate how they will change in the future. Focusing on this method is advantageous for several reasons:

- Predicting future values of $S(t)$, $I(t)$, and $R(t)$ is complicated by the fact that these functions have exponential behavior and generally are not easily modeled by common techniques like polynomial regression. In addition, they may seem removed from human intuition as well, because of their same exponential rates of change. On the other hand, given the appropriate machinery, estimating $\beta(t)$ and $\gamma(t)$ may be more realistic because they are not exponential.
- Given that β and γ are the actual model parameters, and have a causal relationship to the S, I, and R curves, it makes more sense to estimate the model parameters and run the model in the forward direction to make projections, rather than estimating the model output and solving for the parameters in the reverse direction.
- As we noted above, there is some human intuition behind these parameters because their values can be linked to real world phenomenon. Therefore, modeling β and γ accurately would give us a better understanding of how factors like social distancing, contact tracing, and others have a direct relationship with the model, allowing us to separate and analyze each of these causal forces independently.

4.2.4 Historical transmission rate (beta)

We treat the Transmission rate (beta) as a regression problem. β at time t can be represented as a function of time: $\beta(t)$. Therefore our investigation is reduced to a regression problem. We tried several extrapolation and regression techniques to achieve this goal:

- Linear Regression
- Quadratic Regression
- Weighted Average of $\beta(t)$ values with Exponential Decay
- Weighted Average of $\beta'(t)$ values (Slope) with Exponential Decay
- Neural Network Regression

While the regression methods are self-explanatory, the weighted average techniques are a custom extrapolation technique that we created for this project. In these cases, we work backward through the samples in the sample period, and assign a weight to each sample based on an exponentially decaying memory factor, $\alpha^i, \alpha < 1$. This means that samples closer to the end of the sample time series (and closer to the prediction time frame) will be more influential in the future $\beta(t)$ estimate.

We have implemented two varieties of this summarization technique:

- In the first variant, we summarize the **values** of $\beta(t)$ directly. Obviously, this prevents us from making predictions outside of the historical range of $\beta(t)$, but in many cases, the $\beta(t)$ function tends to be a decreasing function after an initial spike, so, as we will see, it gives reasonable results for countries that are further along in their progression, and are currently in the flattening phase.
- The second variant summarizes weighted slope values (which can also be thought of as $\beta'(t)$ values) at a sample. Then, this aggregated slope estimate is combined with the end of the $\beta(t)$ value at the end of the observation period to extrapolate the data and generate a future estimate.

Note that we investigated the use of a neural network to predict future values of $\beta(t)$, but decided against using it for three main reasons, listed below:

1. The neural network has a large number of parameters, and our data size is not quite large enough to train it sufficiently, so our variance values were very large.
2. The training time is slow, so evaluating different variations of models is inefficient.
3. In terms of flexibility, the input size is fixed, so we need to create and train a new network model for every sample size, which is cumbersome.

These methods are all parametric techniques, so we had to experiment with different values for all of these models to see what works best. Figure 4 shows the historical transmission rate for the United States.

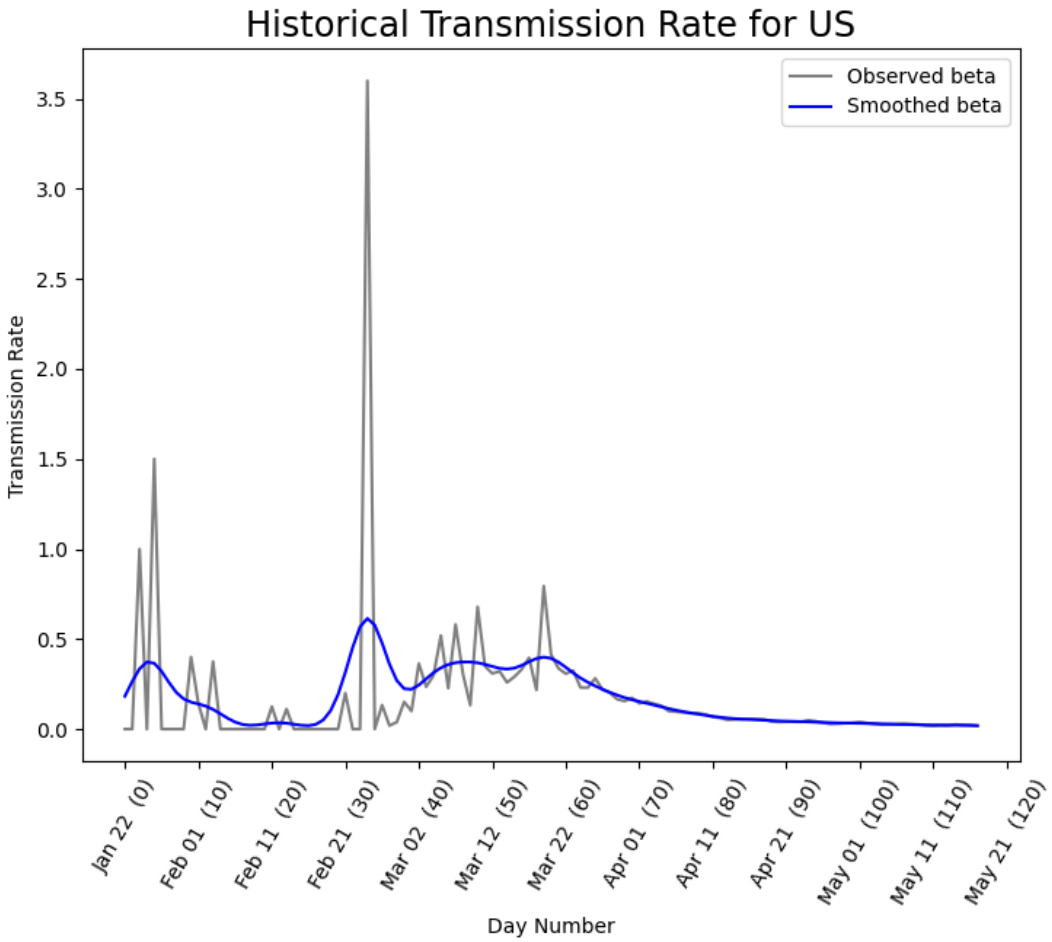

Figure 4: Historical transmission rate for US

4.2.5 Model Selection

To determine the best method for estimating future values of $\beta(t)$, we performed the following steps:

1. Reserved the last two weeks of the data to run our candidate models against. We refer to all data preceding this time frame as the *learning period*, and the final two weeks as the *prediction period*. We chose a two week prediction period because that duration is consistent with the validation time frames for others in the field of COVID-19 model generation [6].

2. Create as combinations of models using different model types and a range of fixed parameters. We generated 1715 different models from the basic types mentioned in the previous list.

3. For each of the 188 countries in the data set, run the model to create future values of $\underline{(t)}$, and score the results, using RMSE (root mean square error) values. These values are computed

by comparing the projected $I(t)$ values from the model with the actual $I(t)$ values. This equates to approximately 322,000 individual SIR simulations for a 14 day period.

4. Aggregate the results and normalize RMSE values by country. We do this by simply dividing all RMSE values based on a country's data set by the *minimal* RMSE value for that country. Larger countries with a higher population, or countries further along in the outbreak will have larger raw RMSE values than other countries, so this method gives us *relative errors*, allowing us to compare the various models more directly.

5. Choose the model that has the best average case performance for all the countries.

6. Analyze model results along various dimensions and parameter ranges to explore trends and identify potential improvements.

4.2.6 Model Parameters

The characteristics of the most important model parameters are described below.

1. **Sample Size**. This value answers the question of how many preceding points we look at while generating the prediction. We always look at the most recent set of points, working our way backward from the end of the learning period. Values for this parameter range between 1 and 60.

2. **Memory Factor**. As described earlier, for the value and slope based models, we use an exponential decay function applied to the points to impact how much weight they carry in the prediction. Values for the memory factor range between 0.9 and 1.0. This functions equivalently to an α decay parameter used in many other algorithms.

3. **Smoothing Value**. This is a constant, typically represented by σ that roughly determines the standard deviation Gaussian filter that we use for smoothing of the observed $\beta(t)$ values. When generating our models, we tried values for this parameter ranging from 1.5 to 4. The behavior of different values in this range is shown below. Note that we are always using these smoothed $\beta(t)$ values for our predictions because the actual $\beta(t)$ values calculated from the historical data are far too volatile to make meaningful predictions.

In addition to these main parameters, we also support 3 different interpolation techniques to fill in $\beta(t)$ predictions based off of a single extrapolation, typically two weeks in the future. The methods we support are below

- **Constant**. Use the estimate of $\beta(t)$for all 14 days in the prediction interval

- **Linear**. Interpolate from the last observed $\beta(t)$value in the learning period to the predicted value after 14 days

- **Quadratic**. Connect the observed $\beta(t)$ curve with a parabola that matches the slope and value of the last observed β value at the end of the learning period and is equal to the predicted beta value after 14 days.

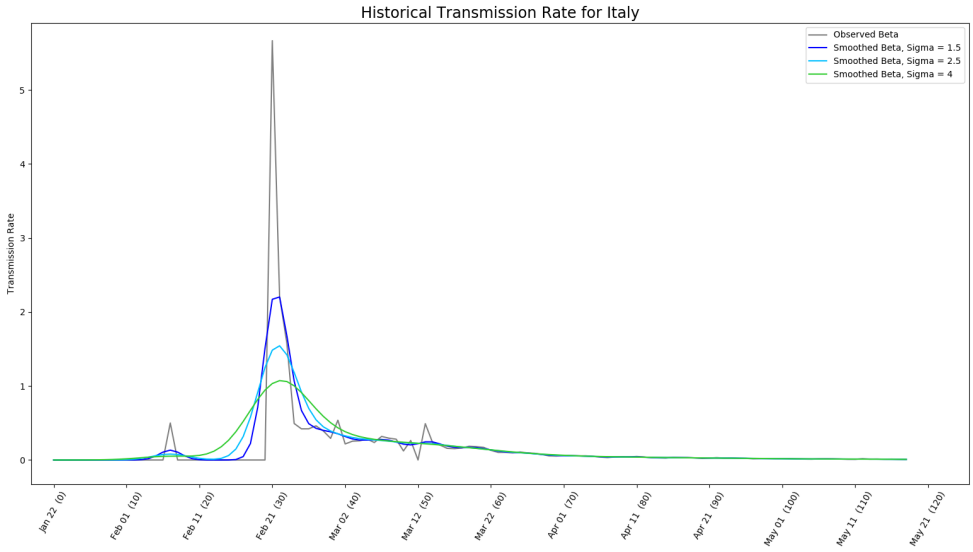

Figure 5: Impact of Changing the Smoothing Factor, σ, on the Historical $\beta(t)$ Values for Italy

5 Experimental results and analysis

After performing the above experimentation, we were able to closely analyze the performance characteristics of the various candidate models. One of the first things that we computed was to find the best model for each country's data. The relative performance of each model category is shown below.

One of the important takeaways from this result is that all model categories are represented, which means that there is not a single model type that strongly outperforms the other candidate models. This fact is an indicator of how difficult the regression problem for $\beta(t)$ is.

We also note that the best model for the United States is a Quadratic Regression model with 10 samples. In general, the countries that fit to quadratic regression or interpolation well are those that are still experiencing large (nonlinear) swings in their numbers. In a qualitative sense, this is an indication that the $\beta(t)$ values United States have not stabilized yet. On the other hand, countries that are modeled by linear or constant models typically are far enough along in the recovery period that their $\beta(t)$ function has flattened out.

We also reviewed the relationship between average model error and the number of sample points, separated by model categories. This output is shown below.

Note that the y-scale is logarithmic. This shows that the quadratic models (orange line) generally have higher error rates, and in fact, the values for the sample size $= 3$ and sample size $= 5$ points are removed for the quadratic category because of some extremely bad data that would have stretched the graph too much to be useful. However, the general conclusion is that too few sample points or too many sample points will both have a negative impact on model performance.

Our next experiment was to determine the models that performed the best overall. To do this we run each model variant on the data for all of the countries and scale aggregate the normalized error

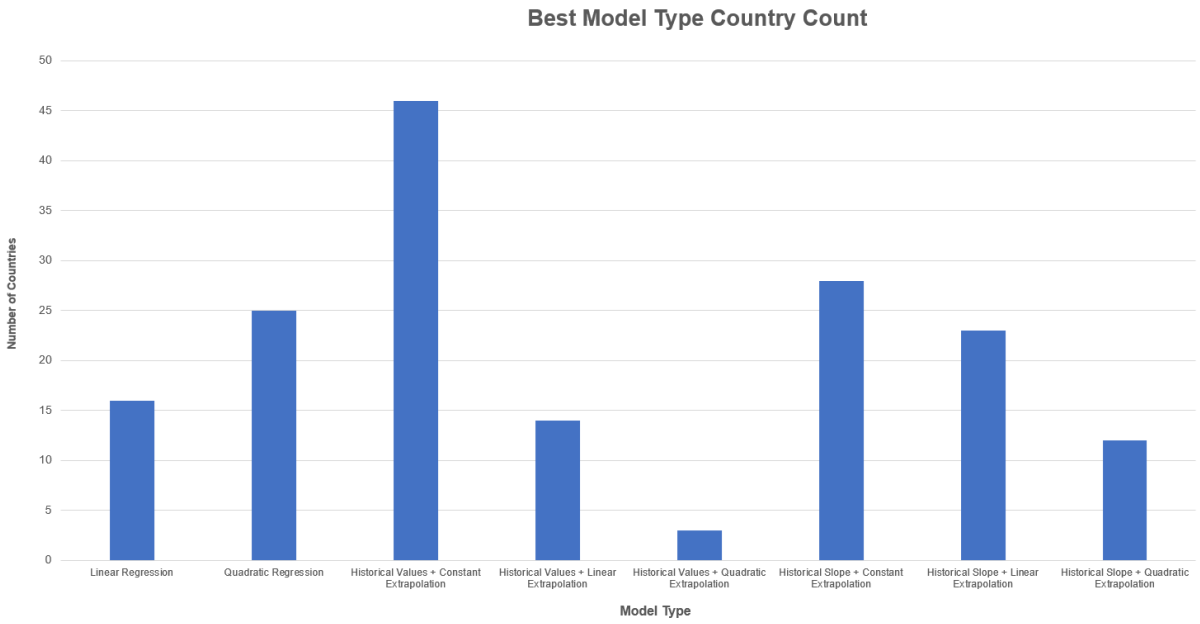

Figure 6: Best Models for Each Country, by Category

values. Therefore, the model that performs the best for a single country will have a relative error of 1 for that data set.

After collecting the results we see that the best performing overall model is our custom weighted value model with an exponential decay of $\sigma = 0.01$ and a sample size of two weeks. Some other observations about these results are:

- The average relative error of the top model was 7.72, which means that there is typically another model that performs 8 times better on a random data set. Therefore, the average model is typically not good compared to the best model for a given data set or country.
- The distribution of top models is extremely close. The top-ranking 100 models all had average relative error values of less than 12.
- This weighted averaging method with exponential decay tends to work well in a variety of cases. However, in general, different models tend to work better for countries in different phases of the pandemic.
- The absolute worst models were the quadratic and slope-based approaches with a small sample size (3-7 previous days), and a small smoothing value of 1.5 or 2. This can lead to an incorrect spike at the end of the learning period, drastically throwing off predictions.

Based on these results, we can use the above methods to generate longer-term predictions and graph the model results against actual data. The image below shows a 30-day validation scenario for the United States. The green, red, and blue lines for our estimates are our optimistic, pessimistic, and trending estimate (based on the past 7 days). All of these projections are based on how β will change, as compared to the historical data. As expected, the true value for the $I(t)$ curve is between our estimates. This is true for most of the countries we looked at.

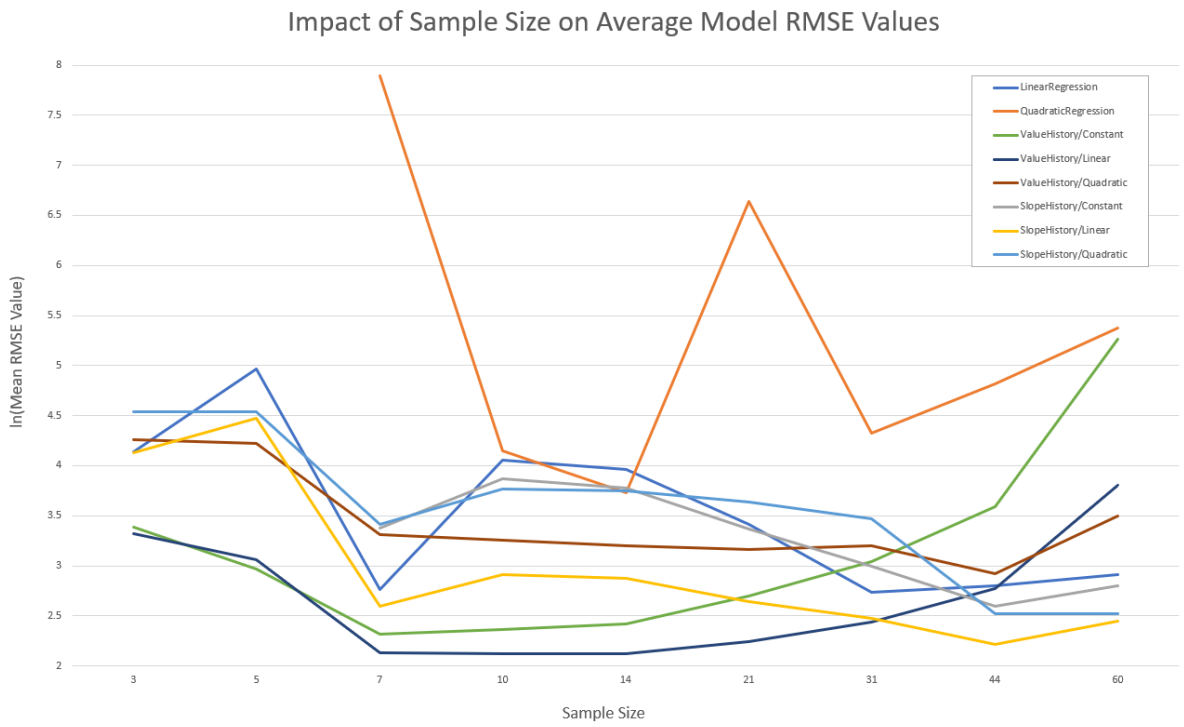

Figure 7: Impact of Sample Size on Model Error

In addition, our 60 day scenario for the US data is shown in figure 9 below.

6 Evaluation

Given that we are undertaking a regression problem rather than a classification problem, we cannot use traditional metrics based on the confusion matrix generated by the test set. Rather, we used R2 and mean square error (MSE) scores. We computed these values for each model that we generate.

In addition to these scores, we also used our models to generate estimates for future counts and then evaluate how close our predictions are once these counts become available. This is a fortunate consequence of analyzing time-series data for an event that is currently in progress.

We ran a ton of different variations of models against the data for all of the countries to do a 14-day prediction, calculate the RMSE score for the prediction, and then aggregates the results from all the models across all of the countries to pick the best one.

6.0.1 Discussion

Any model is only as good as the assumptions that we make. Unfortunately, the SIR model makes many overly simplistic assumptions, as a trade-off for simplicity. Many of these assumptions prevent a challenge for us when trying to improve the accuracy of our estimates because the data reflects activities that are happening in reality, but not accounted for in any way by the model.

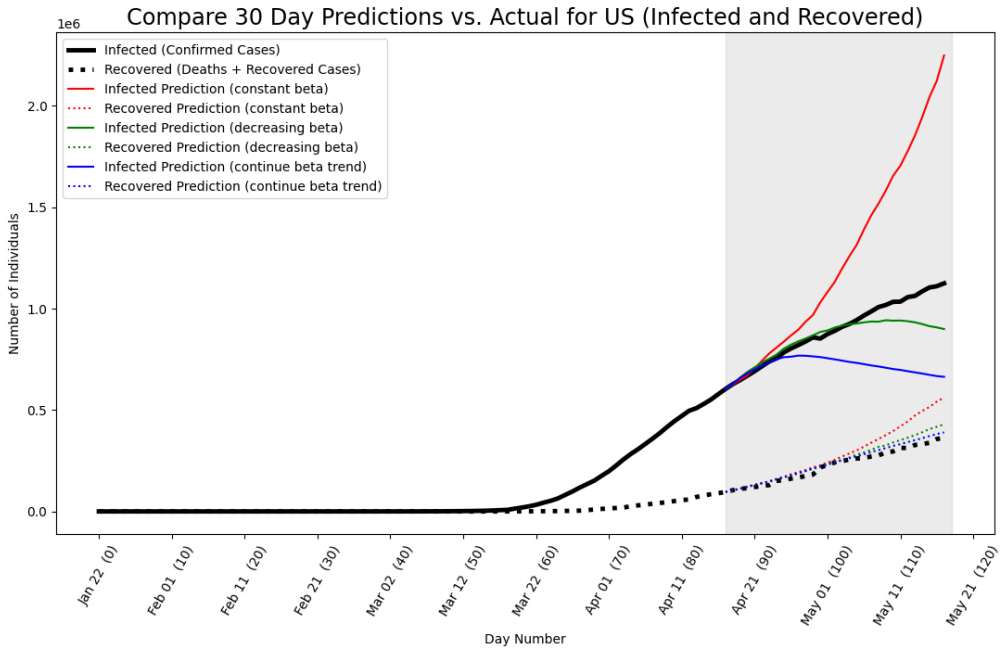

Figure 8: 30 Days prediction Vs. Actual values for US

The sensitivity of the β and γ parameters and the large impact on the SIR model output, even in cases of exceedingly small discrepancies was also challenging. We found that initial assumptions and the results differ considerably between models.

Another lesson learned from this project is that the general regression problem is extremely challenging and additional work will be needed to continue to improve the future predictions for the $\beta(t)$ function, as it is generally not a simple function. Our work could probably be extended by adding logic to detect conditions when some models are more applicable than others.

7 Conclusion

We build a curve fitted SIR model to forecast the trajectory of COVID-19 cases for the next 60 days. This model is based on official sources like the WHO and the CDC estimates for case, death, and recovery counts. At the time of writing this paper, the number of cases in the United States is projected to peak around the first week of July with 1.6 million active cases (Confirmed - Recovered - Death). Italy, one of the hardest-hit countries in the European Union will continue its downward trend reach 13000 active cases by the first week of July. Brazil, on the other hand, is on an exponential growth phase and is continuing to rise exponential reaching 2.5 million active cases by the first week of July. This model can be applied to any other country. COVID-19 is still an ongoing pandemic and the spread of this disease largely depends on each country's policies and social distancing measures.

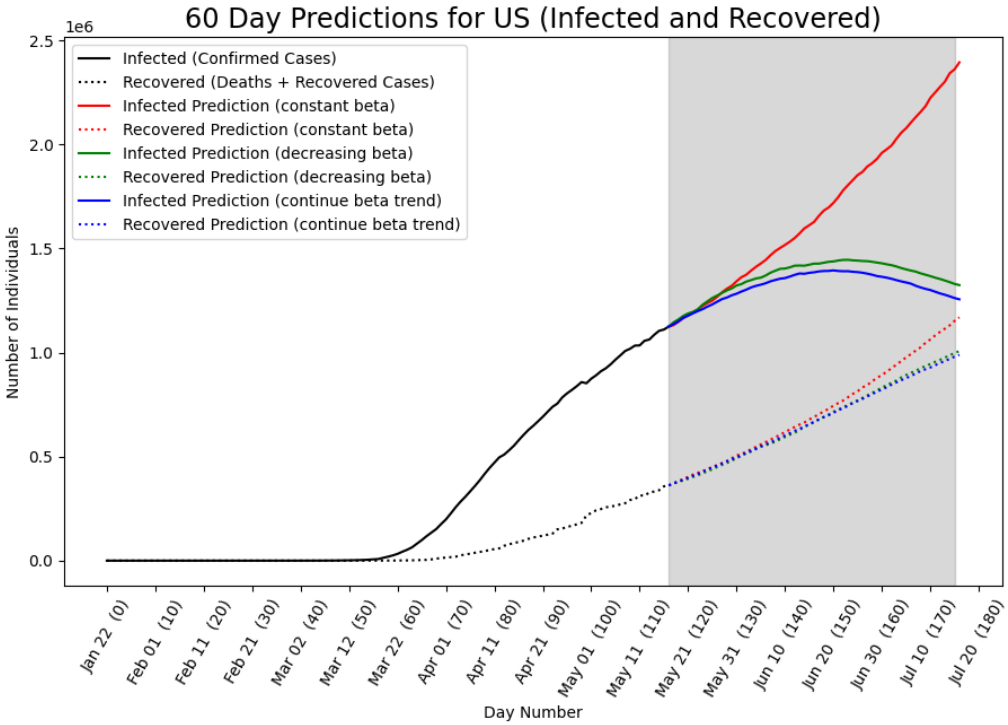

Figure 9: 60 Days prediction for US

7.1 Authors and Contributions

William Austin and Prakash Dhimal jointly investigated the problem, analyzed and interpreted the data, developed the SIR model, implemented regression analysis to project the transmission rate, and evaluated the performance of the models for this paper.

References

[1] and Christopher JL Murray. Forecasting the impact of the first wave of the covid-19 pandemic on hospital demand and deaths for the usa and european economic area countries. *medRxiv*, 2020. 2

[2] Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). Covid-19 data repository by the center for systems science and engineering (csse) at johns hopkins university, 2020. 4.1

[3] Gabriel Goh. Epidemic calculator. http://gabgoh.github.io/COVID/index.html.

[4] World Health Organization. Who timeline - covid-19. 1, 4.2.2

[5] Tanu N Prabu. Population by country - 2020. https://www.kaggle.com/tanuprabhu/population-by-country-2020. 4.1

[6] Jay Boice Ryan Best. Where the latest covid-19 models think we're headed - and why they disagree. 2020. [Online; accessed 17-May-2020]. 2, 1

[7] K. Sasaki. Covid-19 dynamics with sir model. 2020. 3, 7.1

List of Figures

1 Daily changes in COVID-19 cases in the US (03/01/2020 - 05/17/2020) 2

2 Historical Beta Values for Spain, with Smoothing 4

3 SIR Flow model [7] 6

4 Historical transmission rate for US 9

5 Impact of Changing the Smoothing Factor, σ, on the Historical $\beta(t)$ Values for Italy 11

6 Best Models for Each Country, by Category 12

7 Impact of Sample Size on Model Error 13

8 30 Days prediction Vs. Actual values for US 14

9 60 Days prediction for US 15