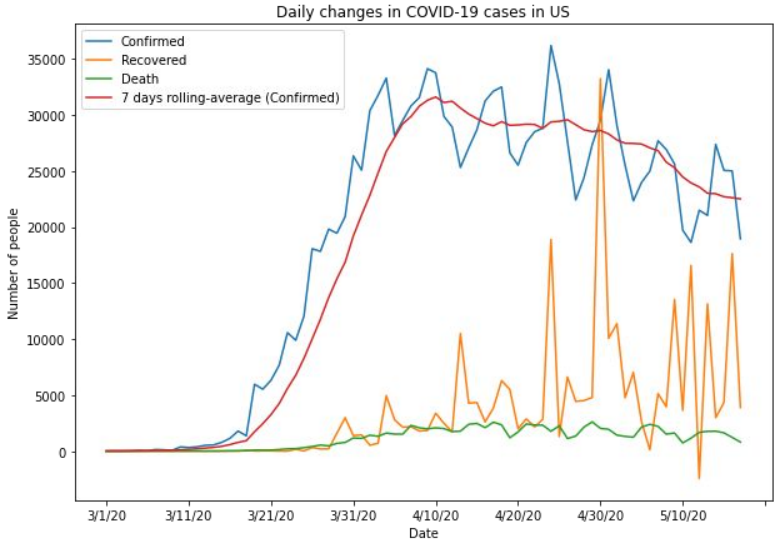
Data Mining Semester Project

Predicting the Impact of COVID-19



Prakash Dhimal
William Austin
CS 584 GMU
May 19th, 2020

Motivation

Problem and Motivation:

- The World Health Organization (WHO) on March 11, 2020 declared COVID-19 a pandemic.
- 4.8 million people infected globally, with more than 300,000 deaths.
- There is significant uncertainty about what the effects the pandemic will be going forward.
- Large amounts of data have been collected as the spread of COVID-19 has progressed.
- Using data mining techniques to attempt to predict likely outcomes will support the task of building an appropriate and effective response strategy.

Project goal:

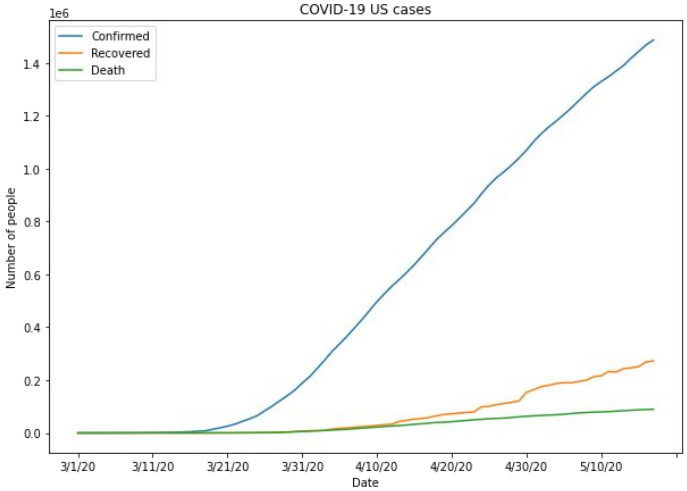
Use data mining techniques to forecast the spread of COVID-19

- Estimating the transmission rate (***β***) based on historical data.

Data source:

The Johns Hopkins University COVID-19 Data Repository

- Contains daily values for confirmed, recovered and deaths
- Broken down by country and region
- Time series data
- Cumulative number



Modeling infectious disease

Compartmental models:

- SIR model
- Divide the population (N) into compartments:
  - **S** - Susceptible
  - **I** - Infectectious
  - **R** - Removed (Recovered + Dead)
  - **N = S + I + R**

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

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

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

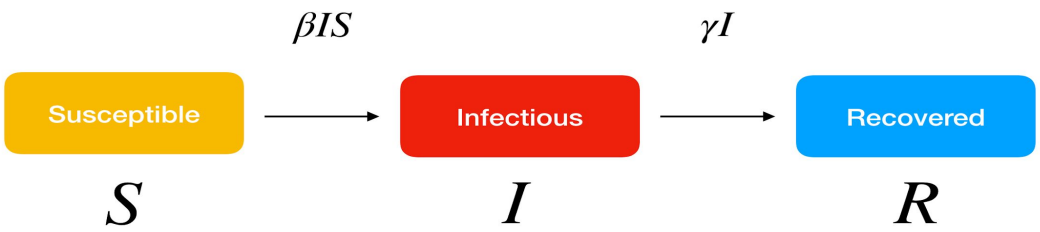


Fig: Differential equations to represent the SIR model

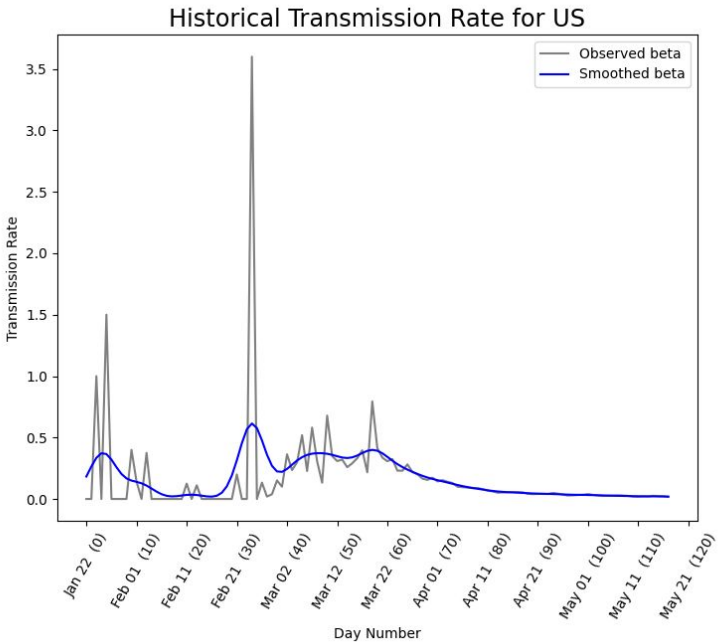
Standard SIR Model Parameters

- Transmission Rate (***β***): the rate at which those in the Susceptible group move into the Infectious group
  - Use curve fitting to extrapolate the transmission rate.
- Recovery rate (**γ**): the rate in which those in the Infectious group move into the Recovered/Removed group
  - D = the number of days an infected person can spread the disease
  - **γ** = 1/D
- Maximum time to run the model (60 days)
- Note that we hear about R_0, which is the “basic reproduction number” for the virus and represents how many healthy people the average infected individual transmits the virus to. Sometimes, this is also called “R-nought”. The formula is $R_0 = \beta/\gamma = \beta * D$.
  - Note that we sometimes use the convention that ***β*** value is multiplied, by **N**, or the population size. Each different population is modeled separately, so is just a constant that is simply rolled into the value of ***β***, and the projections are not changed.
- 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 PPE for health care workers.

Curve fitting the parameters

Transmission rate (beta):

- Treat it like a regression problem
- β at time t can be represented as a function of time : $\beta(t)$
- We tried a number of extrapolation techniques for this function
  - Linear Regression
  - Quadratic Regression
  - Simple and Weighted Averaging
  - Extrapolation based on slope
  - Neural Network Regression
- Most of these methods are parametric techniques, so we had to experiment with different values for all of these models to see what works best.
- Note that future numbers of infected and recovered individuals are very sensitive to $\beta(t)$. However, we still prefer this method because β and γ are causal variables and changes in their values can be linked to real world phenomenon, whereas the changes in S, I, and R may seem removed from human intuition because they change exponentially.



Model Selection Process

In order 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 models against.
2. Create as many possible combinations of models using different model types and fixed parameters. We generated 1715 different models from 8 basic types, as we explain below.
3. For each of the 188 countries, run the model and score the results, using RMSE values, 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 scores for each country. Larger countries with a higher population, or countries further along in the outbreak will have larger RMSE values than other countries.
5. Choose the model that has the best average case performance for all the countries.
6. Aggregate model data along various dimensions to explore trends and parameter values.

Model Types and Parameters

Model Types

1. **Linear Regression:** Fit the set of preceding daily observations in the learning point to a linear function and extrapolate
2. **Quadratic Regression:** Use at least 3 points, and use built-in numpy functionality
3. **Value Based:** Look at the previous $\boldsymbol{\beta}(t)$ values from the learning period sample, and combine them to form a weighted sum for the 14 day estimate.
4. **Slope Based:** Calculate a slope estimate based on a weighted average of the previous $\boldsymbol{\beta}'(t)$ values, and use it to extrapolate a 14 day estimate.

Prediction Curves, for Types 3 & 4

1. **Constant:** Use the estimate of $\boldsymbol{\beta}(t)$ for all 14 days in the prediction interval
2. **Linear:** Interpolate from the last observed $\boldsymbol{\beta}(t)$ value in the learning period to the predicted value after 14 days
3. **Quadratic:** Connect the observed $\boldsymbol{\beta}(t)$ curve with a parabola that matches the slope and value of the last observed $\boldsymbol{\beta}$ value at the end of the learning period and is equal to the predicted beta value after 14 days.

Other parameters

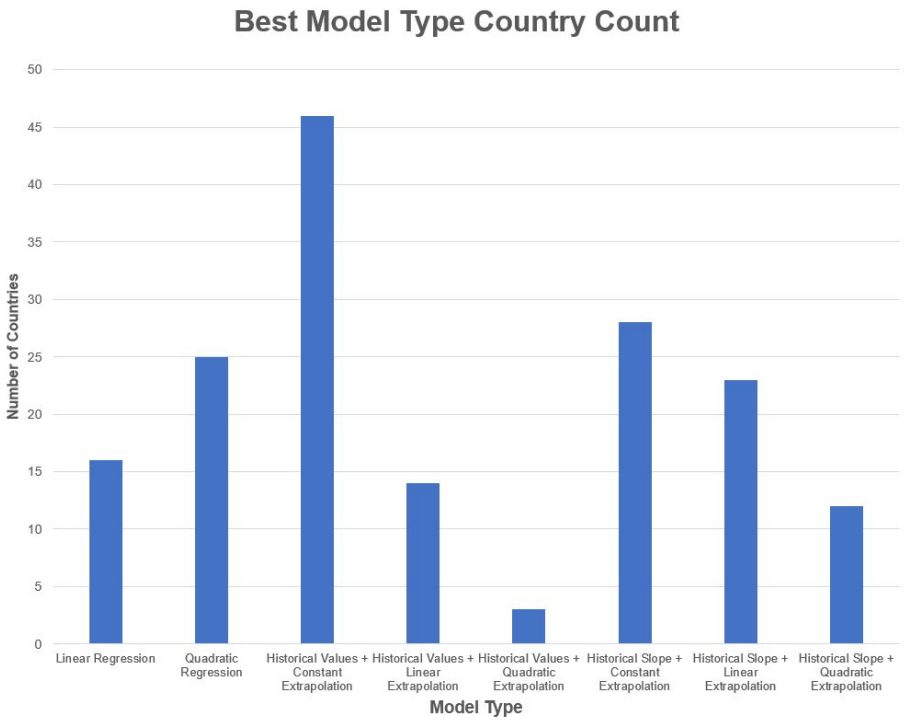
1. **Sample Size:** How many points do we look at for the prediction? We always look at the most recent set of points, working our way backwards from the end of the learning period. Ranges between 1 and 60.
2. **Memory Factor:** For the Value and Slope based methods (3 & 4), we use an exponential decay function applied to the points to impact how much weight they carry in the prediction. Ranges between 0.9 and 1.0 for the decay α value.
3. **Smoothing Value:** This is a constant, σ, used for Gaussian smoothing of the observed $\boldsymbol{\beta}(t)$. The behavior of different values is shown for the Italy slide. Note that we are always using these smoothed $\boldsymbol{\beta}(t)$ values for our predictions.

Results and Observations from Model Selection

Throughout this process, we noted a few things about the best models:

- The distribution of top models is extremely close together. There is less than an 11% average RMSE drop across the top ranked 100 models.
- The best model for the United States was based on a quadratic regression with 10 samples.
- In our case, we found that the overall best method was to compute an exponential moving average (EMA) with the previous 2 weeks of smoothed data and generate a linear interpolation to this predicted future point.
- While these models do well in the *average* case, they typically perform about 8 times worse than the best fit model for each country.
    - More specifically, quadratic and slope-based models tend to fit some of the data fairly well, but the average case is very bad.
- As noted, models selecting an exponential moving average of points or slopes tend to do well. Linear extrapolation and quadratic extrapolation work well, but quadratic extrapolation can also lead to large errors because it also tries to match slope at $\boldsymbol{\beta}(t_{max})$.
- Constant value extrapolation tends to be bad. As expected, most of the $\boldsymbol{\beta}(t)$ values are stabilizing, but for countries still experiencing a spike, predicting a flat $\boldsymbol{\beta}(t)$ leads to large errors.
    - It is difficult to create a model that predicts well in both the early and later 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 σ value (1.5-2) for smoothing $\boldsymbol{\beta}(t)$, leading to an incorrect spike at the end of the learning period.

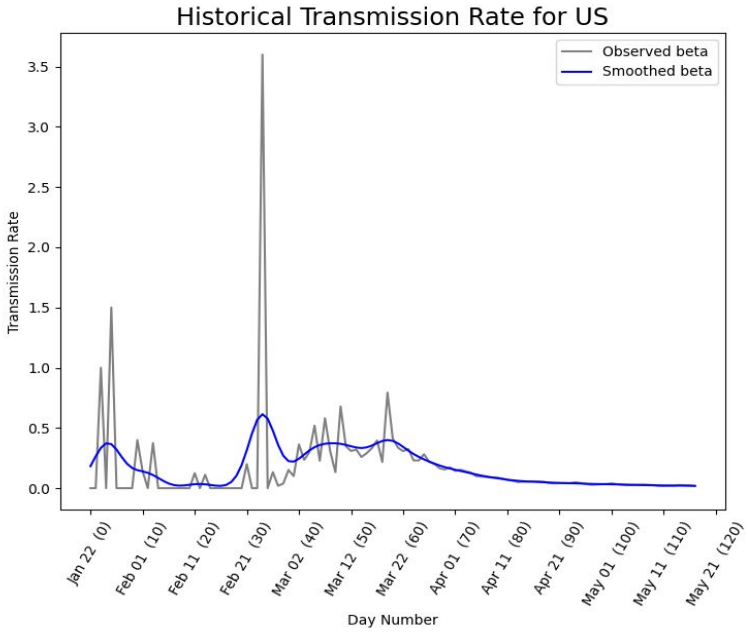
Model Results: Best Models for Each Country

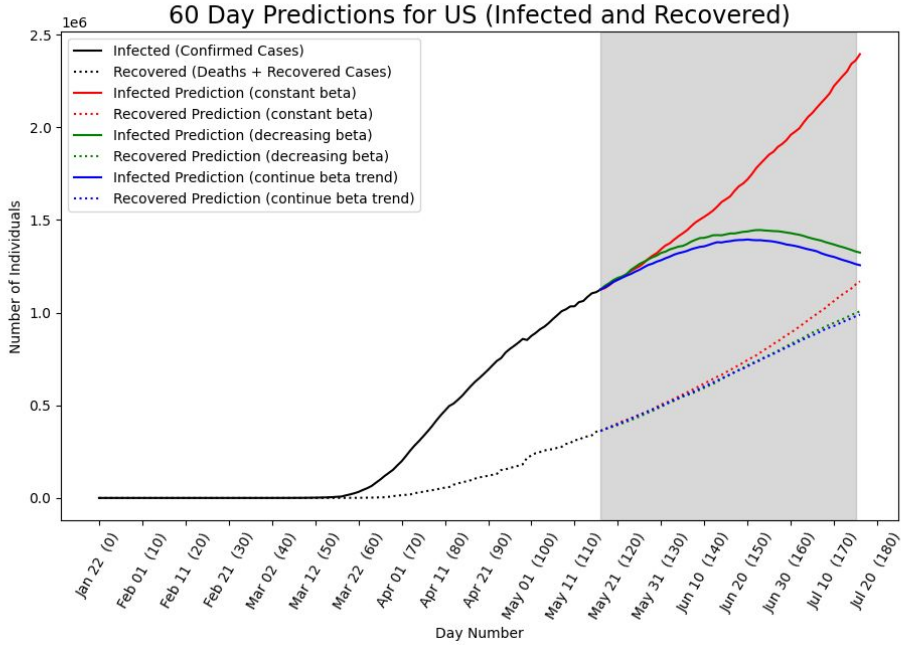


Some Observations

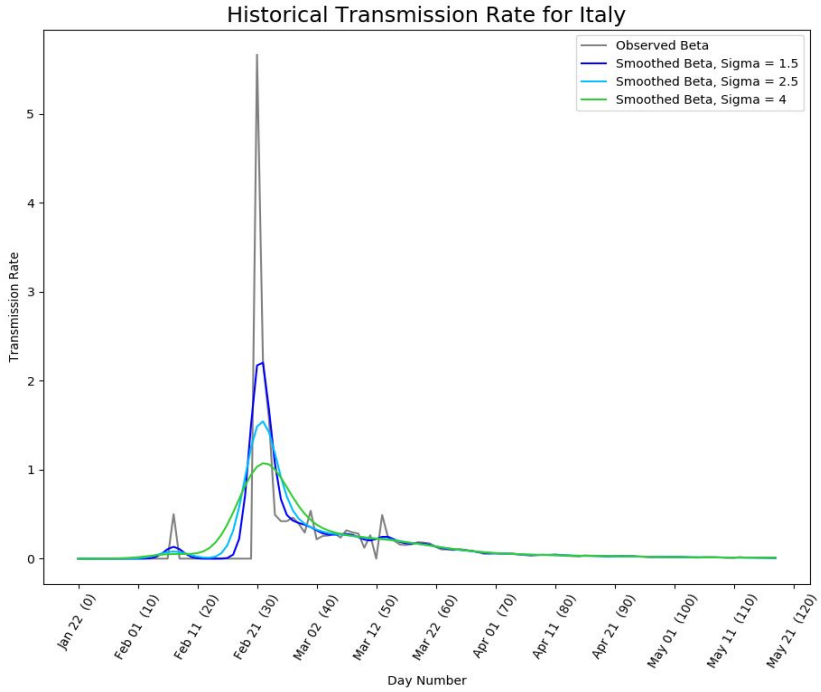
1. All model types are represented, which shows that the $\beta(t)$ function is difficult to model and make predictions for.
2. Many countries have entered a period where the $\beta(t)$ curve has significantly flattened, which is why the Value-based method with constant extrapolation was quite successful.
3. The countries with higher volatility (like the US are more successfully modeled with a slope-based or quadratic approach.

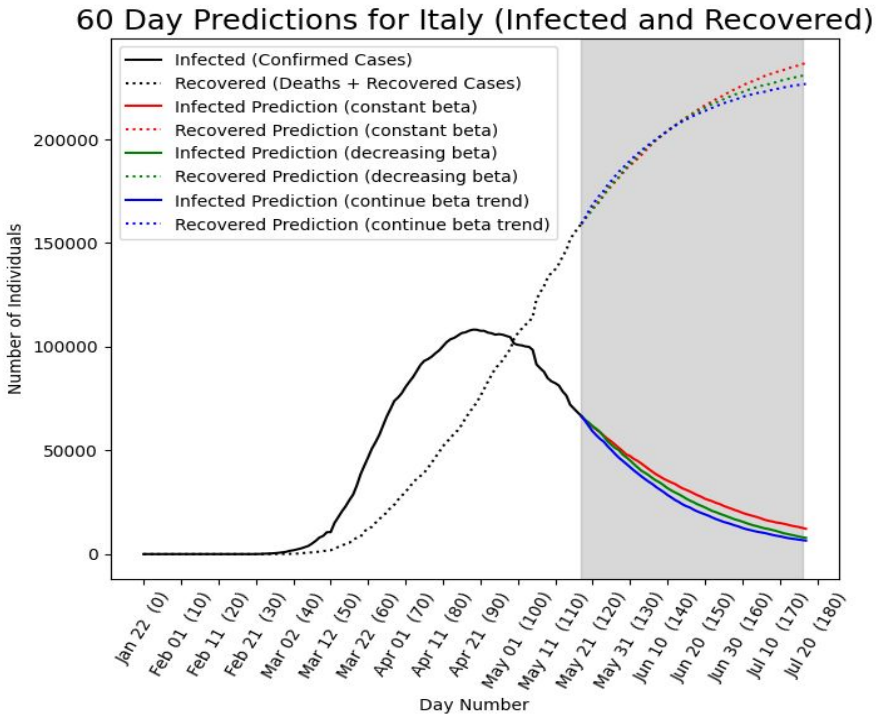
Predictions for United States



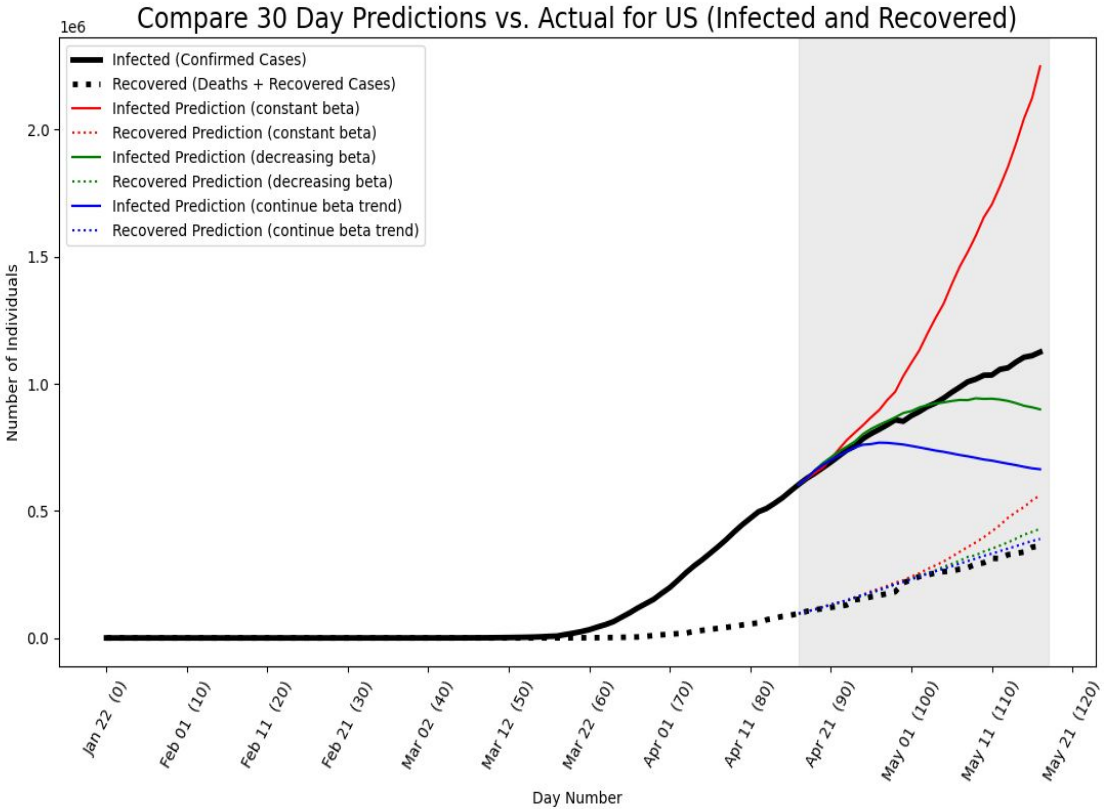


Predictions for Italy





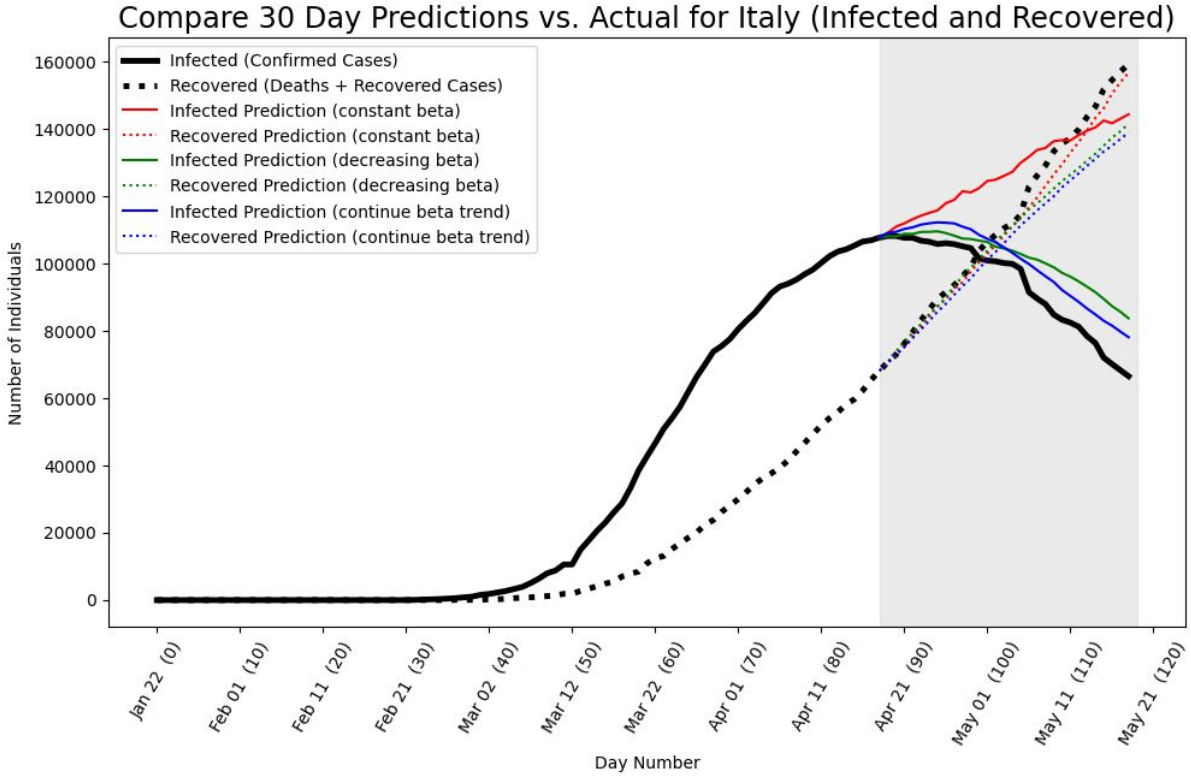
Validation (US Data Set)



This graph shows the result of our initial modeling efforts. For the data from the United States, tt shows:

- The solid black line and dotted lines are are actual value from the data set for I(t) and R(t).
- The red, green, and blue lines are based on our initial set of estimates.
    - The red line shows the outcome if the β(t) value at the end of the learning period (white) stays constant throughout the prediction period (gray).
    - The green line shows the scenario where β(t) approaches 0 in a linear fashion, over the prediction period.
    - Blue shows the effect of a continuation of the 7 day linear trend line for β(t).
- This shows that the actual future value lies between our pessimistic and optimistic predictions, and this case is true for almost all countries.
- The sensitivity of the I(t) curve to β(t) is also shown because the trend at the end of the learning period had somewhat decreased, and it had a significant effect on the projections.

Validation (Italy Data Set)



Unlike the US, Italy noticed a severe drop in the number of infected individuals during the prediction period.

- This was one of the few countries where the real numbers do not fall between the green (optimistic) and red (pessimistic) lines.
- Note that our prediction lines are not completely smooth. This is because we add some random noise to the projections for $\boldsymbol{\beta}(t)$.
    - A side benefit of this procedure is that we can simulate the same $\boldsymbol{\beta}(t)$ predictions with different noise values. Then, when we run the SIR model with the noisy $\boldsymbol{\beta}$, we get slightly different numbers each time, so we can generate a project “cone” with confidence intervals, instead of just a single prediction.

Exploring a Neural Network to Predict $\beta(t)$

For this project, after investigating the standard regression models and our custom models based on previous values of $\beta(t)$ and $\beta'(t)$, we also created functionality for generating a training set to be used for training a neural network based on the `sklearn.neural_network.MLPRegressor` class. However, we did not investigate this further for several reasons:

- Although neural networks can model complicated functions, they require copious amounts of data for training. The total size of our data set is: 188 countries x 115 days of observations x 3 observations per day = 64860 data points, which is fairly small.
    - To train the test neural network, we can generated a training set based on a sliding window for each country with 5 observations spread across a 3 week sampling period intended to predict the value of $\beta(t)$ two weeks in the future. The best performing NN shape had hidden layers of size 5, 4, and 2. This implies that the total number of parameters to learn is:
        - (5 x 5 + 1) + (5 x 4 + 1) + (4 x 2 + 1) + (2 x 1 + 1) = 59 which is fairly high for a small network and not enough data. As a result, our weights did not consistently converged and the result was very high variance.
- The neural network has a very slow training time and is not particularly flexible for easily allowing varying sample windows, which corresponds to the input layer size. This problem may be a candidate for an RNN, but the data size is probably still too small.

Conclusion

- We build a curve fitted SIR model to forecast the trajectory of COVID-19 cases for the next 60 days.
- 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.
- 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.
- Other countries like Brazil are on an exponential growth phase.
- This model can be applied to any other country.
- COVID-19 is still an ongoing pandemic and the spread this disease largely depends on each country's policies and social distancing measures.

Questions?