COVID-19 Predictions using SIRD Dynamics and Machine Learning

Kieran Bissessar

Project Code: https://github.com/bissessk/COVID-19-Predictions-using-SIRD-dynamics-and-Machine-Learning

1 INTRODUCTION

COVID-19 (Coronavirus disease 2019) is a disease that is brought on by the virus strain, SARS-Cov-2 (Severe Acute Respiratory Syndrome Coronavirus 2). Coronaviruses are a subcategory of enveloped RNA viruses that tend to effect birds and mammals. For humans, the virus is known to infect the individual's respiratory system. These infections incite bronchitis, pneumonia, and severe respiratory illnesses. SARS (Severe Acute Respiratory Syndrome), MERS (Middle East Respiratory Syndrome), and COVID-19 are all examples of severe respiratory illnesses due to coronaviruses. Also, the virus strain behind the SARS (SARS-Cov) is the predecessor to SARS-Cov-2. Despite the similarities of SARS, MERS, and COVID-19, differences exist. Although COVID-19 is not as lethal as SARS and MERS, it is much more infectious.

A virus can be crudely summarized as RNA that's wrapped with a lipid layer. The only way it can multiply would be to infect a host and enter their cells. Once the organism is exposed to the virus (most likely from inhalation or contact), it works it way down the respiratory tract and infects the lungs among other organs. The virus would then inject its RNA into the cells of those organs and being that RNA contains instructions on how to multiply itself, the infected cell would carry out those instructions. This leads to the production of viruses inside the cell and once enough viruses are made, the cell undergoes apoptosis. Once the cell dies, the virus is free to infect other cells. As cytotoxic T cells, neutrophils, and other immune system related cells arrive to try and attack the virus-infected cells, the virus induces certain cytokines that cause the immune cells attack healthy cells as well as infected ones. Although the majority patients with the syndrome recover with minimal repercussions, some (usually those with underlying immune conditions) exhibit fibrotic scarring in the lungs or become vulnerable to bacterial infections.

The infectious nature of the virus accounts for the fact that there are roughly 4 million confirmed cases of the virus globally at the date this was written (1.3 million of which are in the United States alone). The spread of the virus became a pandemic with many countries carrying out stay-at-home orders and quarantines. Travel bans have been instated across the globe and many airlines (which had made up 1.7 trillion in U.S. economic activity and 10 million American jobs) have become in danger of going out of business. Also, many hospitals have been unable to keep up with the demand and have run out of space and resources such as ventilators.

Something that might help with this would be a model that can provide insight into how the disease would spread. From an economical perspective, this might help businesses (big and small) make certain financial decisions to stay in business. From a political perspective a model is useful to law makers to determine what kind of mandates they should be making to minimize the spread of the disease, help keep their constituents safe, and keep the economy afloat. And from a biological perspective, the spread of the disease can be important in understanding the evolution of the virus. The virus is constantly evolving and if a biologist discovers a specific change in the virus strain, perhaps the effect of that change could be observed in the model. Because of these reasons, many resources are being poured into organizations that could aid in making such a model. Organizations such as the IHME (Institute for Health Metrics and Evaluation) and the CDC (Centers for Disease Control and Prevention) have all been constantly updating their own prediction models.

The goal of this analysis is to design a prediction model that implements machine learning techniques to help formulate the forecasts. Doing so allows one to explore how many of the other

prediction models operate. Also, a lot of machine learning regression techniques are static. So, although it may accurately make short term predictions, the goal is to design a machine learning algorithm that can consider a dynamical system as an attempt to make long term predictions. And although there are some dynamical machine learning algorithms that exist, most notably the Kalman Filter (used primarily in engineering settings), designing a model for a specific problem may be a better fit for the problem at hand.

A compartment model was used to construct the model. The SIR model is a common compartment model used in the field of epidemiology. In this analysis, a variant of the SIR model was used – the SIRD model. The SIRD model is composed of 4 components that make up a population. The compartments are the susceptible compartment, the infectious compartment, the recovered compartment, and the dead compartment. Susceptible is defined as the population who can still catch the virus. Infectious is defined as the compartment that has the virus and can infect others. Recovered is defined as the population who have had the disease, but don't anymore. The assumption that recovered people couldn't get re-infected was made. Figure 1 below demonstrates how people can move from one compartment to another. And the Death compartment is defined as the population that were unable to recover and died.

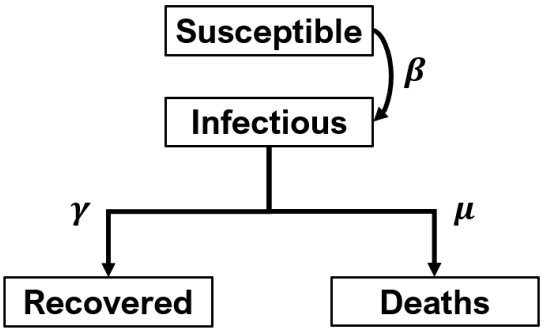


Figure 1: This figure depicts the compartments that make up the SIRD model and how people can move from one compartment to another. People can move from the susceptible compartment to the infectious one. The rate in which this happens is defined as $\boldsymbol{\beta}$. From the infectious compartment people can go to either the recovered compartment or the death compartment at a rate defined as $\boldsymbol{\gamma}$ and $\boldsymbol{\mu}$ respectively.

The relationship between the four compartments can be modeled mathematically into a set of differential equations as shown in Figure 2 below.

$$\frac{dS}{dt} = -\frac{\beta SI}{N}$$
$$\frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I - \mu I$$
$$\frac{dR}{dt} = \gamma I$$
$$\frac{dD}{dt} = \mu I$$

Figure 2: This figure depicts the dynamics between the susceptible, infectious, recovered, and dead compartments. $\boldsymbol{S}$ is defined as the amount of susceptible people, $\boldsymbol{I}$ is defined as the number of people in the infectious compartment $\boldsymbol{R}$ is defined as the number of people in the recovered compartment, $\boldsymbol{D}$ is defined as the number of people in the dead compartment, and is $\boldsymbol{N}$ defined as the total population.

The data used in this analysis came from the CSSE (Center for Systems Science and Engineering at John Hopkins University). Provided in that dataset is the confirmed counts for the number of cases, recoveries, and deaths for a collection of countries every day since January 22, 2020. Although this is a public dataset used to track the spread and posted on GitHub instead of with an actual paper, several studies were done with that data. The methods used in this analysis were inspired from such studies.

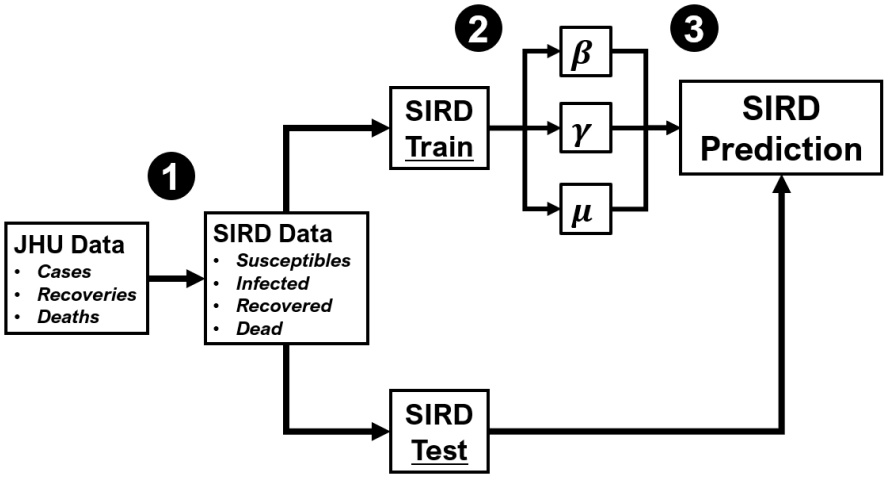
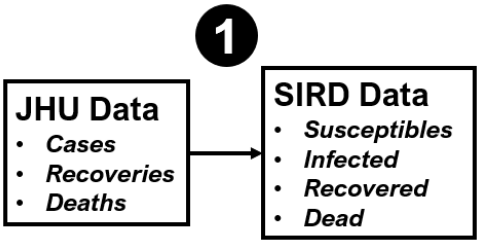


Figure 3: This figure depicts the pipeline used to analyze the JHU data. The numbers represent the main parts of the workflow. The workflow can be organized into three main parts. The first part involves translating the raw data collected by the CSSE into the SIRD data. The second part involves using the training set that was subsetted from the SIRD data and training it to create a model that would allow us to construct forecasts. The third part involves using the model created in the second part to make predictions about the susceptible, infectious, recovered, and dead counts.



$$Suceptibles = Population - Cases$$
$$Infected = Cases - Recoveries - Deaths$$
$$Recovered = Recoveries$$
$$Dead = Deaths$$

Figure 4: This figure highlights the first part of the workflow which is the conversion of the raw confirmed cases, recoveries, and deaths to the susceptible, infected recovered, and dead counts. The equations above represent how exactly they were calculated. The recoveries and deaths were taken directly from the CSSE data. The suceptible counts were defined as the population subtracted by the confirmed cases. The infected counts were calculated by subtracting the confirmed cases by the recoveries and deaths. As previously mentioned, the compartments must sum to the population. After doing this the suceptible, infected, recovered, and dead counts were obtained for all the days of all the countries from the initial dataset.

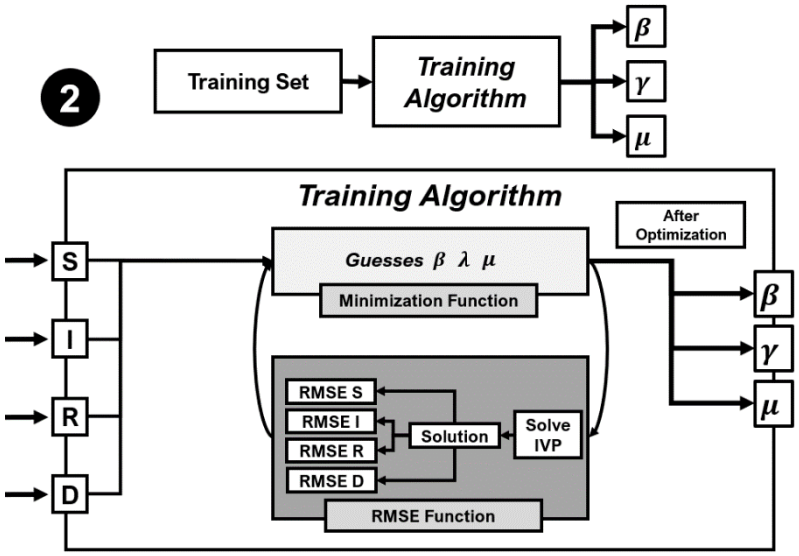


Figure 5: This figure highlights the second part of the workflow. This portion of the pipeline focuses on training the data. The SIRD dataset assembled in the previous part is passed into the minimization function where it estimates $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\mu}$ values that could then be plugged into the set of differential equations in figure 2 to make predictions. The way the training method was designed was that it will initially guess initial $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\mu}$ values. Then it will be passed into another function that creates a solution set of SIRD values based on those guessed parameters. Following this, a weighted sum of the root mean square error (RMSE) values will be calculated between the estimated SIRD values and the actual SIRD values acquired from the training set. The function will then return an average of the RMSE values. The RMSE function is the cost function and the minimization function minimizes the RMSE function using `scipy.optimize.minimize`. This will allow the minimization function to return $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\mu}$ estimates that have the lowest RMSE values when solved and compared to the training data. The method used to acquire the local minimum is known as `L-BFGS-B`. L-BFGS-B is a variant of the limited memory Broyden–Fletcher–Goldfarb–Shanno algorithm. It is known to be efficient at finding local minimums for dynamical systems.

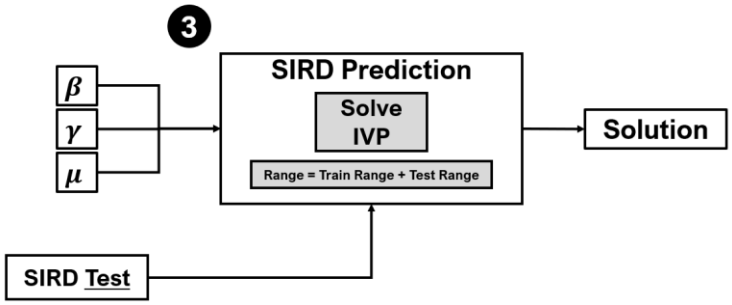


Figure 6: This figure highlights the last portion of the of the workflow. Using the $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\mu}$ estimates obtained after training, predictions are made by solving the initial value problem using the dynamics in Figure 2 and the initial S, I, R, and D values obtained in the first part of the workflow (Figure 4). The SIRD Test set is used to get the total range of the prediction as well as test how the prediction fares with what was observed.

In this analysis, the results of the model will be compared to the testing dataset and its accuracy will be compared to another machine learning prediction model known as Polynomial Ridge Regression.

2 Results

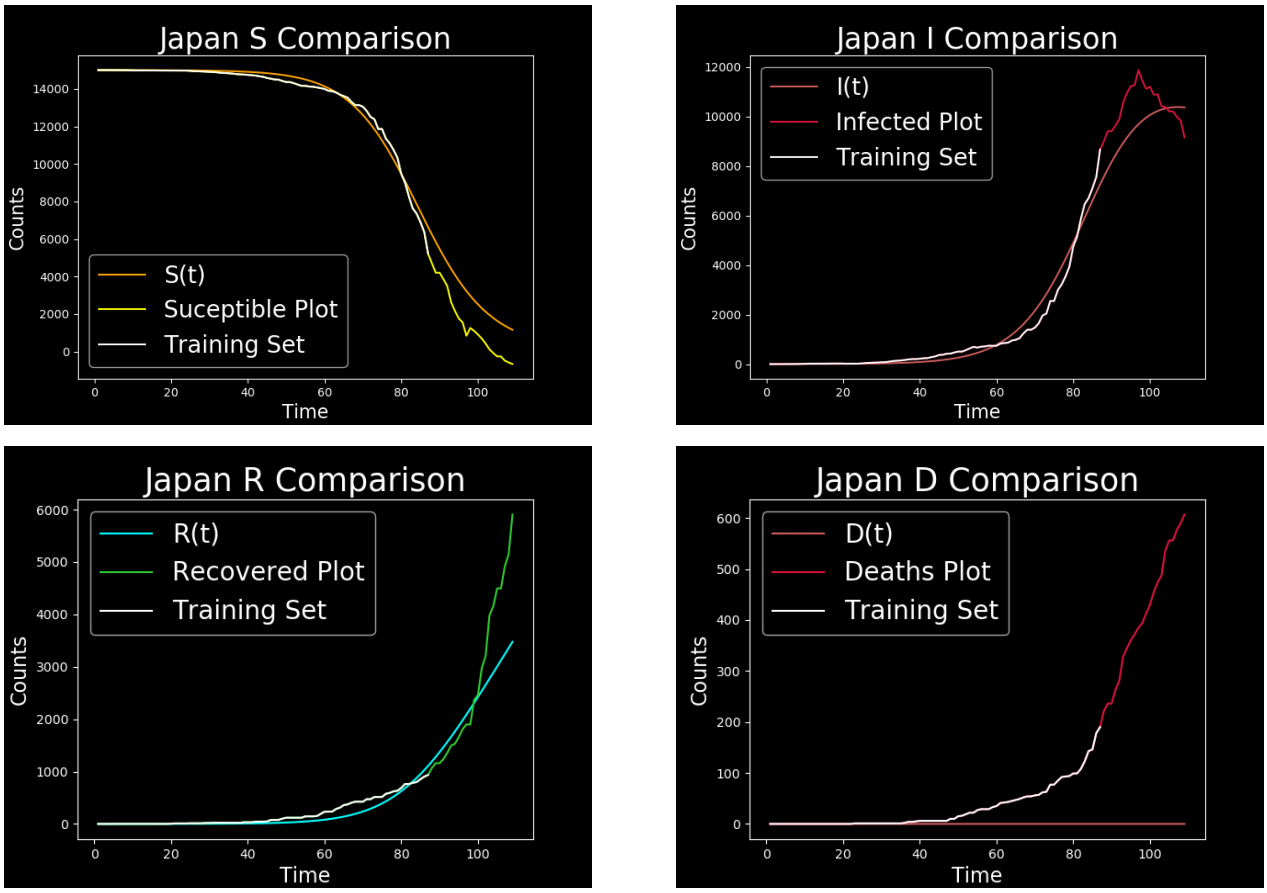


Figure 7: This figure shows the results that the SIRD prediction model had for the country of Japan. Each graph plots the model and the SIRD set. The training cases are colored white. The training size for all these graphs was chosen to be 80%.

- **S Comparison (Upper Left):** The results show that the model closely resembles the test cases. The training accuracy was high and the test accuracy was slightly lower as expected. Despite this the general curve of the function was accurately determined.
- **I Comparison (Upper Right):** The results show that the model closely resembles the training cases. Note how the training cases (white) is only increasing. But, despite this the model was still able to see that a the Infected counts were going to decrease soon. If a regression curve were to have seen only the training cases, it would never know that the infected counts were going to lower soon.
- **R Comparison (Lower Left):** The results show that the model was underestimating the recovered counts. Although the training accuracy was relatively high, the testing accuracy wasn't as high.
- **D Comparison (Lower Right):** The model doesn't seem be making predictions for the death counts.

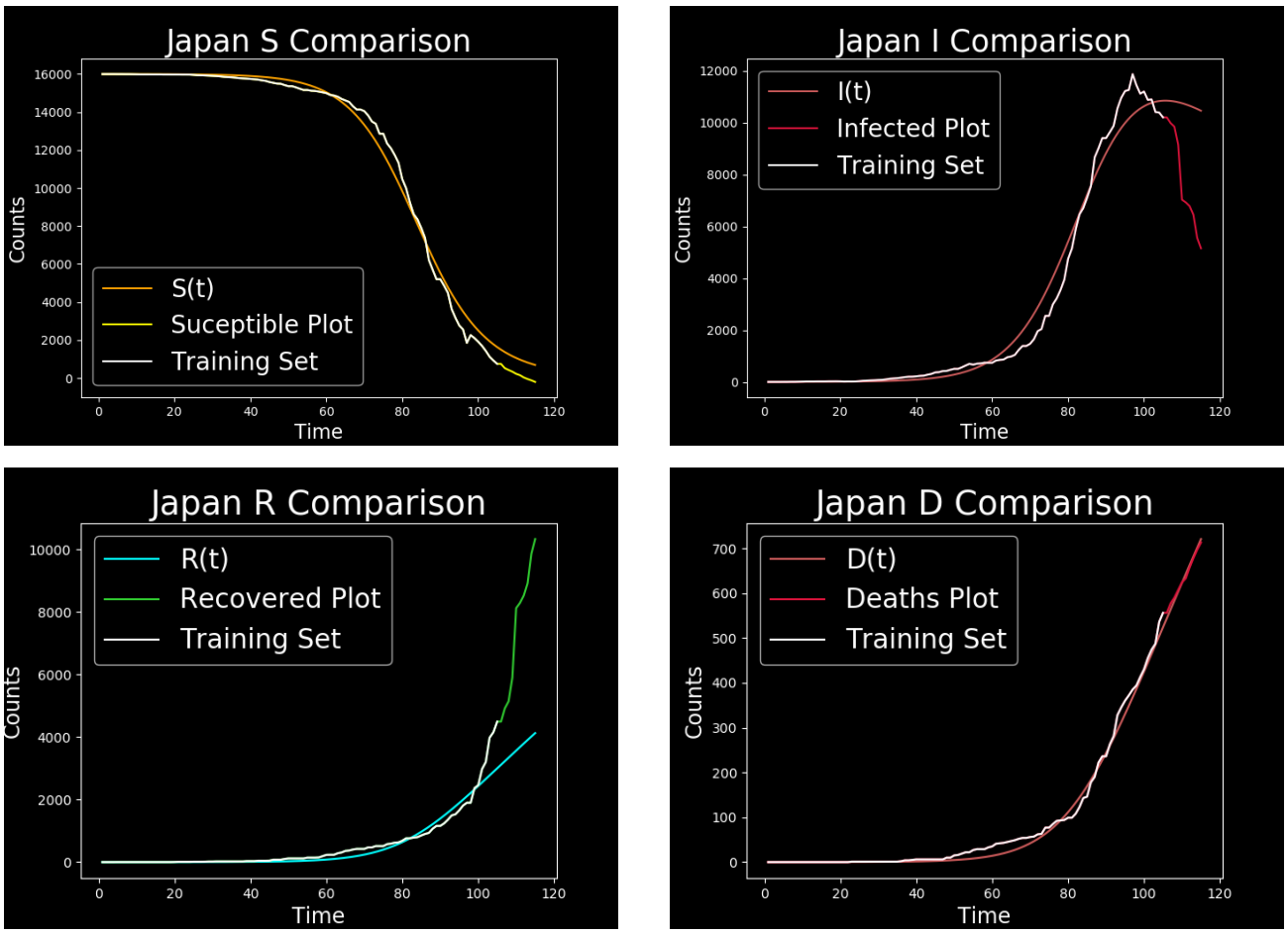


Figure 8: This figure shows the results that the SIRD prediction model had for the country of Japan. Each graph plots the model and the SIRD set. The training cases are colored white. The training size for all these graphs was chosen to be 90%.

- **S Comparison (Upper Left):** Increasing training size by 10 percent made the model slightly more accurate.
- **I Comparison (Upper Right):** Increasing training size by 10 percent also slightly increased increases the accuracy in the infected count predictions.
- **R Comparison (Lower Left):** Although the training size was increased by 10 percent, the model was still underestimating the recovered counts.
- **D Comparison (Lower Right):** After increasing the training size by 10 percent, the model predicts the death count rather accurately. This is a dramatic different from the results when the training size was only 80%.

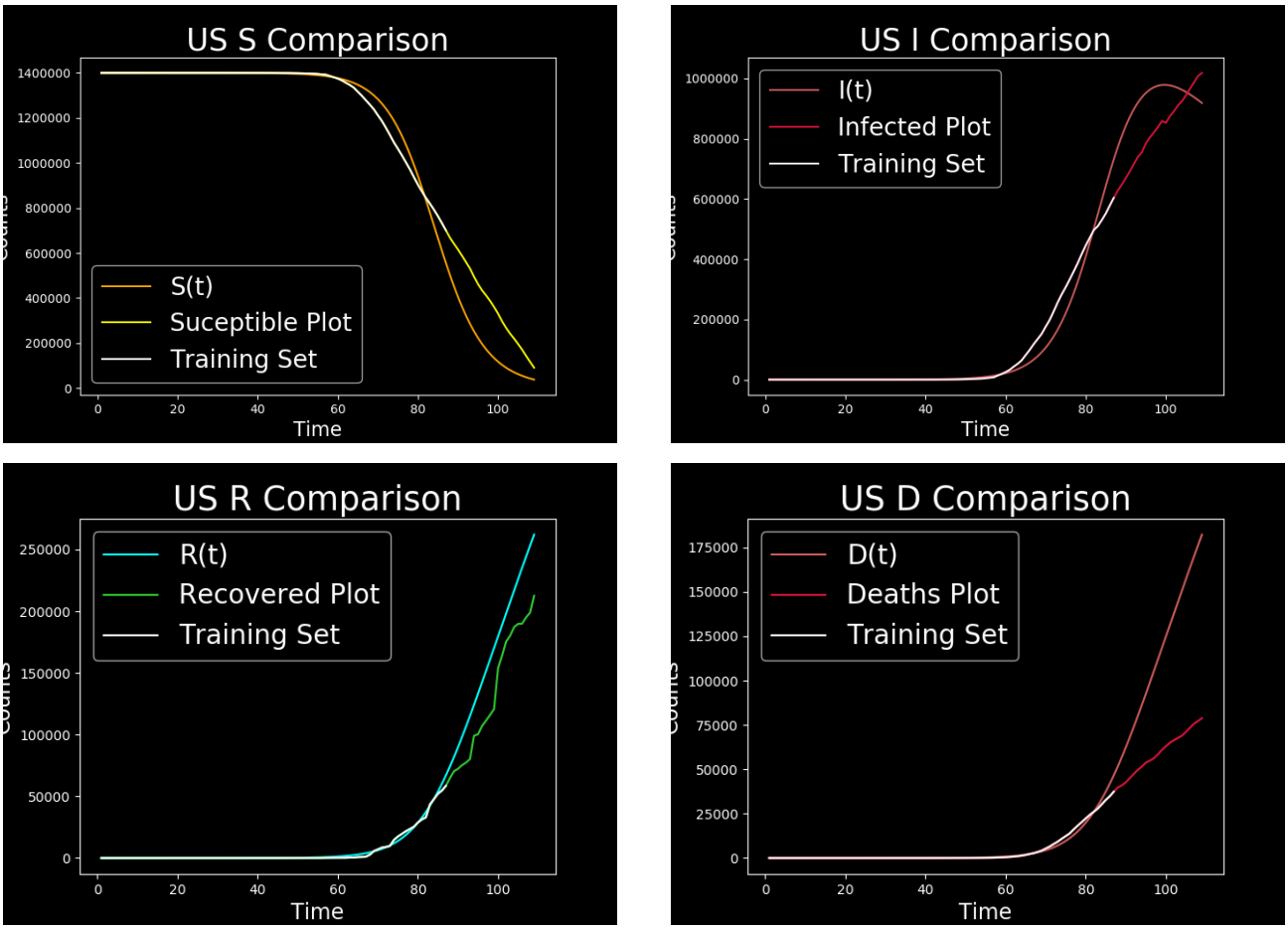


Figure 9: This figure shows the results that the SIRD prediction model had for the United States. Each graph plots the model and the SIRD set. The training cases are colored white. The training size for all these graphs was chosen to be 80%.

- **S Comparison (Upper Left):** Although the test accuracy doesn't seem to be remarkably different, it does seem like the model is seeing a curve when there isn't one yet.
- **I Comparison (Upper Right):** The model seems to think that the amount of people infected will drop but according to the dataset this has not happened yet.
- **R Comparison (Lower Left):** The model resembles the recovered plot.
- **D Comparison (Lower Right):** The model is overestimating the death counts. There is a lower training error and a high test error, suggesting overfitting.

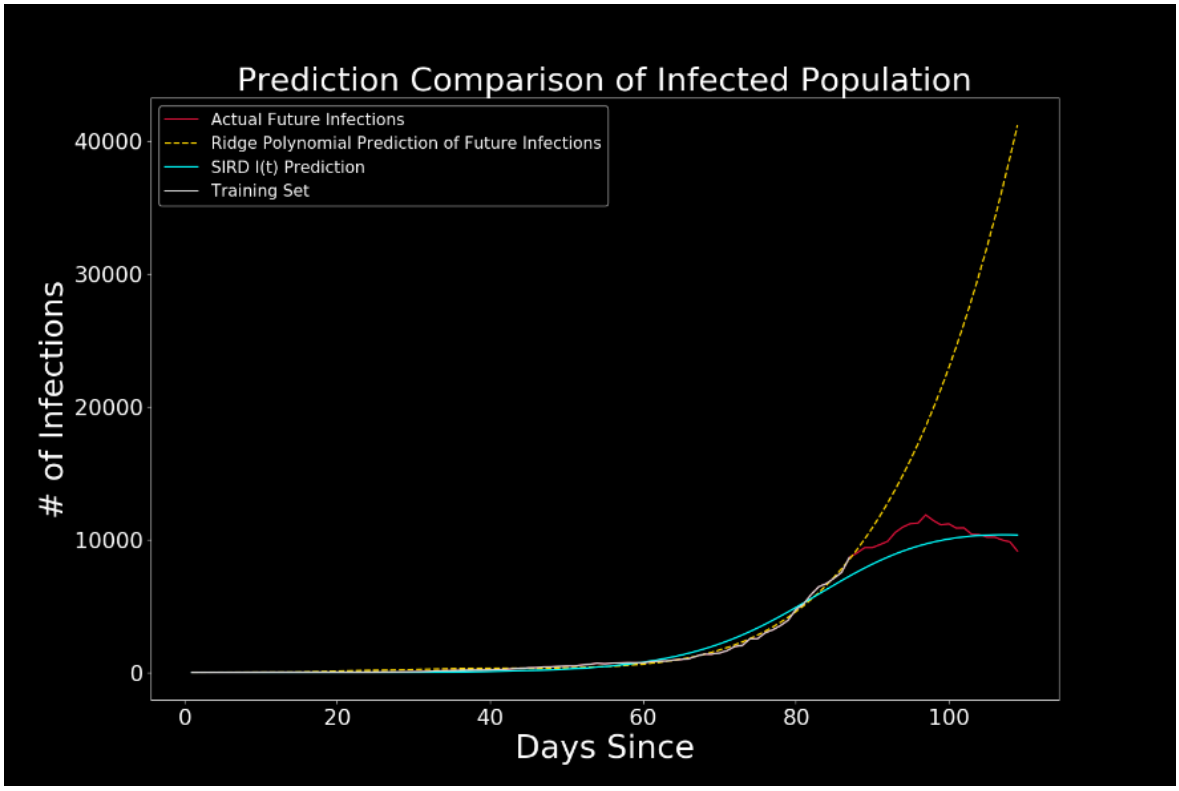


Figure 10: This figure compares Sklearn's ridge polynomial ridge regression on the training data (80%) for Japan to the created SIRD prediction model. The prediction done by the SIRD outperforms the prediction done by the regression. This is because the regression model has no reason to believe there would be a decrease based on the training data as opposed to the SIRD model which was able to factor in the dynamical relationships in Figure 2 to help with that.

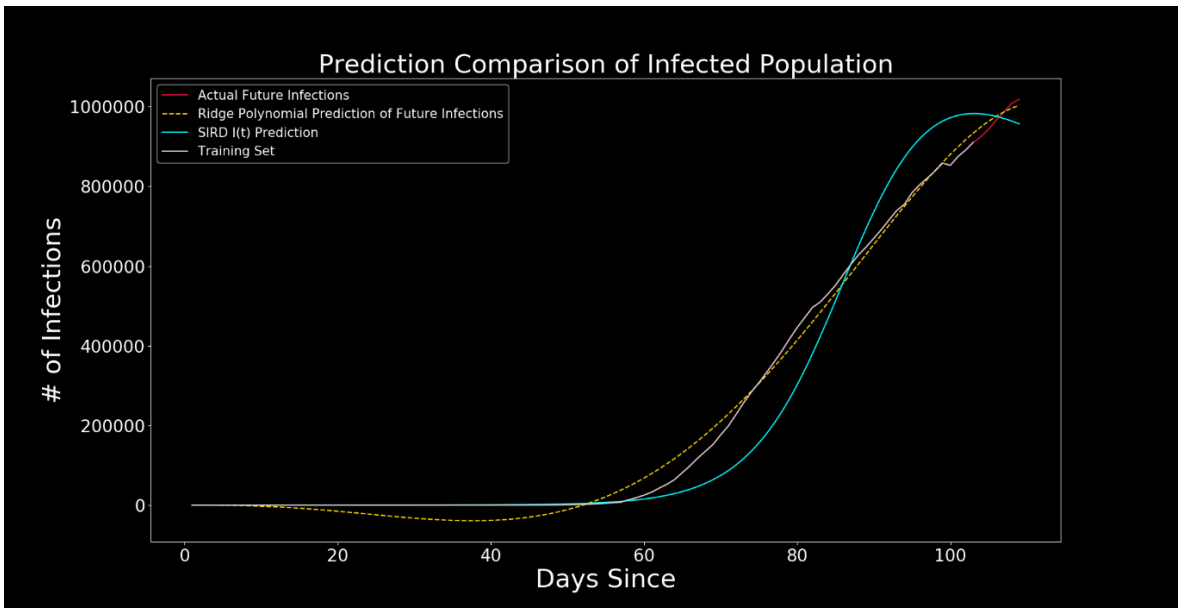


Figure 11: This figure compares Sklearn's ridge polynomial ridge regression on the training data (95%) for the United States to the created SIRD prediction model. It can be seen the regression closely fits the

actual cases. Although the SIRD prediction model doesn't seem to be too off from the actual cases, it seems to think that there should be a dip around that time which isn't seen yet from the data. In this case, the regression model seems to be much more accurate when the training data is a larger percentage, or predictions when the curve isn't changing much. In this scenario the infected count in the United States seem to be increasing in a way that it is easy for the regression model to predict.

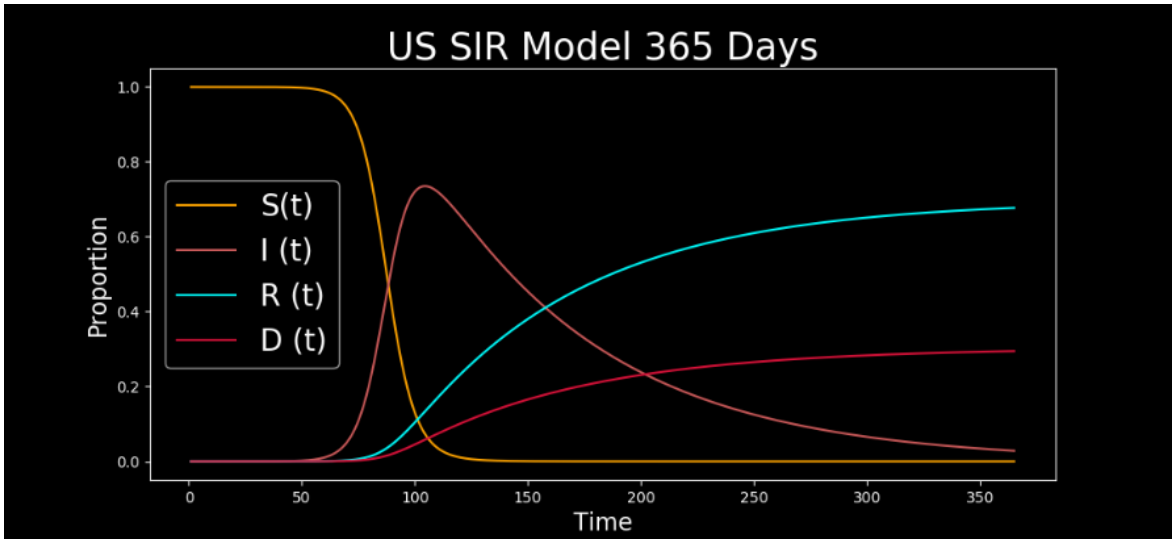


Figure 12: This figure uses the data provided to predict the suceptible, infected, recovered, and death counts in the United States for the next year. This was done to observe what a much longer predication of the model would look like. The model thinks that the United States is around its peak of infected people and should decrease slower over the next few months.

3 Discussion

Based on the results, it seems as if the SIRD prediction model performed better of the data that came from Japan than the one that came from the US. This can be seen after comparing Figure 7 and 9. Generally, the model's predictions seem to somewhat resemble the testing set of the data. Especially seen in Figure 7 in the infected cases, the training data did not give any indications that it was going to decrease, but because the model looks at the training that and considers the dynamic properties in Figure 2, it was able to make that prediction.

In some of the cases, the training data seemed to fit very close, but the prediction was not nearly as close when compared to the testing set. In machine learning, this is usually an indicator of overfitting. As shown in Figure 5, the training method optimizes the weighted sum of the root mean square between the predicted and actual S I R, and D values. So, I way to fix this would be to lower the weight of value that overfitting.

Figure 10 and 11 compares the predictions made by the SIRD prediction model and Sklearn's Ridge Polynomial Regression. Based on the results, it seems that the ridge polynomial regression was better at making short term predictions as opposed to longer ones. Shown in Figure 10, the ridge repression especially didn't fare as well when there were sharp maxima or minimas in the test set. The purpose of the SIRD prediction model though was to use machine learning techniques to estimate answers to questions like when will the curve start to flatten or when will the number of infections start to go down. So, a way to judge the model would be to see how it handles predicting general trends and finding a peak as opposed to the specific number since there's room for error.

As mentioned before, the SIRD prediction mode is far from perfect. For example, in Figure 7, the model had a difficult time modeling Japan's deaths when the training data was at 80%. Also, in the examples that used the United States' data (Figure 9), the actual data does show what the model is showing. Specifically, in the suceptible plot, the model is trying to see a curve when the data doesn't suggest one. Also, the infected plot prediction is showing a dip when the data doesn't show that yet.

There are a few reasons for the error that could be found in the model. The two largest things that effect it is the minimization done in the training algorithm used to estimate the parameters. The next is the initial estimate of the susceptible population. There is a possibility that the minimization done during training is finding a local minimum instead of something close to the global minimum. This could effect the parameter values which could explain why the model couldn't model Japan's infected at 80% training. Also the initial suceptible count was guessed for each country. According to some scientific literature about SIR modeling, the initial suceptible count depends on a lot of things and is different for each country. Most choose a number between 10,000 and 60,000. So the way it was determined as by trying different ones until the model sort of fit. For Japan this number was 16,000 and for the US it was 1.4 million. These numbers are drastically different and since the US seems abnormally high, this could provide a reason for why the model performed better for Japan than it did on the US. A minimization algorithm was originally written to find better initial suceptible estimates, but the run time was not practical.

Another was to show the inaccuracies of the model would be to calculate the R0 value. R0 is calculated by dividing beta by gamma. R0 represents the reproduction number and is defined as the average number of people infected from one other person. This value was 13 and 18 in Japan and the US respectively. This is drastically larger than the R0 values found in most scientific literature which puts the value between 2 and 3 for most countries with Covid-19.

This model was not made to do anything as well or more than prediction models done by the IHME and the CDC. It was just meant to apply concepts done in machine learning to dynamical systems. It was also done to try and make more long-term predictions than the ones done by most regression models. There are so many more things to consider. Some of the best compartment models can have dozens of different compartments and differentials. How the virus is being treated and how people are treating social distancing are all things that effect it. Also the model is only as good as the data that is fed in. There are many sources that suggest the difficulty od determining if a person actually died from Covid19 etc.

Machine learning is defined as being able to use computer algorithms that improve with experience by feeding it information. This model does exactly that. The model is able to combine the information taken in (case, infect, and death counts) with knowledge about how the dynamical system should work to create a prediction that gets better when more information is added to it.

4 References

Airlines For America. 2020. The Airline Industry | Airlines For America. [online] Available at: <https://www.airlines.org/industry/> [Accessed 19 May 2020].

Amira Binti Hamzah, F., 2018. Coronatracker: World-Wide COVID-19 Outbreak Data Analysis And Prediction. [online] Who.int. Available at: <https://www.who.int/bulletin/online_first/20-255695.pdf> [Accessed 19 May 2020].

Billiau, A., 1980. Virus Replication- An Introduction. [online] Nature. Available at: <https://www.nature.com/articles/pr198040> [Accessed 19 May 2020].

Eletreby, R., 2020. The Effects Of Evolutionary Adaptations On Spreading Processes In Complex Networks. [online] Available at: <https://www.pnas.org/content/117/11/5664> [Accessed 19 May 2020].

Hewings-Martin, Y., 2020. How Do SARS And MERS Compare With COVID-19?. [online] Medicalnewstoday.com. Available at: <https://www.medicalnewstoday.com/articles/how-do-sars-and-mers-compare-with-covid-19> [Accessed 19 May 2020].

Hou, C., Chen, J., Zhou, Y., Hua, L., Yuan, J., He, S., . . . Jia, E. (2020). The effectiveness of the quarantine of wuhan city against the corona virus disease 2019 (COVID-19): Well-mixed SEIR model analysis. Journal of Medical Virology, doi:10.1002/jmv.25827

IA, R., 1997. Cytokines And Immunity To Viral Infections. - Pubmed - NCBI. [online] Ncbi.nlm.nih.gov. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/9416507> [Accessed 19 May 2020].

Johns Hopkins Coronavirus Resource Center. 2020. COVID-19 Map. [online] Available at: <https://coronavirus.jhu.edu/map.html> [Accessed 19 May 2020].

Karako, K., Song, P., Chen, Y., & Tang, W. (2020). Analysis of COVID-19 infection spread in japan based on stochastic transition model. Bioscience Trends, doi:10.5582/bst.2020.01482

Roda, W. C., Varughese, M. B., Han, D., & Li, M. Y. (2020). Why is it difficult to accurately predict the COVID-19 epidemic? Infectious Disease Modelling, 5, 271-281. doi:10.1016/j.idm.2020.03.001

Sanger, D., 2020. Under Intense Criticism, Trump Says Government Will Buy More Ventilators. [online] Nytimes.com. Available at: <https://www.nytimes.com/2020/03/27/us/politics/coronavirus-trump-ventilators-gm-ventec.html> [Accessed 19 May 2020].

Sasaki, K., 2020. COVID-19 Dynamics With SIR Model. [online] The First Cry of Atom. Available at: <https://www.lewuathe.com/covid-19-dynamics-with-sir-model.html> [Accessed 19 May 2020].Cao, H., Wu, H., & Wang, X. (2020). Bifurcation analysis of a discrete SIR epidemic model with constant recovery. Advances in Difference Equations, 2020(1) doi:10.1186/s13662-020-2510-9

Segal, A., 2005. How Neutrophils Kill Microbes. [online] NCBI. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2092448/> [Accessed 19 May 2020].

Wan, K., Chen, J., Lu, C., Dong, L., Wu, Z., & Zhang, L. (2020). When will the battle against novel coronavirus end in wuhan: A SEIR modeling analysis. Journal of Global Health, 10(1), 011002. doi:10.7189/jogh.10.011002

Wang, H., Wang, Z., Dong, Y., Chang, R., Xu, C., Yu, X., . . . Cai, Y. (2020). Phase-adjusted estimation of the number of coronavirus disease 2019 cases in wuhan, china. Cell Discovery, 6(1) doi:10.1038/s41421-020-0148-0

Datasets Used:

CSSE Confirmed Cases

https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv

CSSE Confirmed Deaths

https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_global.csv

CSSE Confirmed Recovered

https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_recovered_global.csv