Modeling COVID-19 Spread in a Linear Dynamical System

Mario Rodriguez-Montoya
ESE 105
Washington University in Saint Louis
St. Louis, Missouri
mario.rodriguez-montoya@wustl.edu

Finn Doornweerd
ESE 105
Washington University in Saint Louis
St. Louis, Missouri
dfinn@wustl.edu

Owen Cromly
ESE 105
Washington University in Saint Louis
St. Louis, Missouri
c.owen@wustl.edu

***Abstract*—To analyze and predict trends in COVID data, we created a generalized model for epidemics, representing the changes over time of the people susceptible to the disease. We then applied this generalization to COVID data in the Saint Louis Metropolitan area, and preparing this data to better suit the model led to the final result. This process consisted of three sets of procedures, and so, to allow for more readability, this document is composed of three parts, each of which allowed for the completion of the part proceeding it.**

I. Introduction

This project is important in the sense that, ideally, it would create a generalized model of the spread of epidemic diseases. More relevantly, during this aftermath of COVID-19, it is more important than ever to examine the spread of diseases to facilitate the creation of prophylactic measures. So then the more focused goal of this project is to, by adapting our generalized model to fit COIVD-19's spread, further examine the causes, consequences and efficacy of measures taken during the periods of COVID outbursts.

This is no simple feat, and, realistically, there is no singular cause nor is there any truly finite number of causes for any of these outbreaks in COVID cases. As such, we make no assumptions as to the root cause of any one wave or of the disease as a whole, and instead come to conclusions as trends which pertain to one or more waves. For example, our data records cases that were documented in Saint Louis and by officials, so it is more than likely people who were never tested nor went to the hospital were not recorded. Another factor outside of the control lies in the idea that people who catch the virus outside of the Saint Louis Metropolitan area (where our data is sourced) then move into the area were not recorded if they did not visit the hospital. This means as, depending on the phase of COVID, they might not have been quarantined and so should still count in the infected population. These ideas reveal imperfection in our data, and so the same weaknesses apply to the models that come from the data, as compensation of these faults would simply come down to further assumption-making.

By creating a generalized mathematical model for the spread of COVID-19 over time, we aim to create a model with the ability to frame any change in the condition of the epidemic (e.g. new policies, a potential vaccination, cure, etc.) mathematically. This is done to model certain theoretical changes in factors at play during the epidemic, and to show how hypothetical differences in real world scenarios could affect the outcomes of the epidemic. Using this generalization, we created models which had the capability to consider any outside factor systematically, assuming the other factors are dealt with accordingly.

II. Methods

Before elucidating the means by which we met our goals, we felt it important to mention we made certain design choices specific to COVID data, as the modeling of any other disease through our generalization would require a change in the method to account for the behavior of certain diseases. For example, as will be covered in $Part\ two$, we chose to take data over two week periods as opposed to daily, which works only after making certain assumptions about the behavior of the virus. This means our generalized model would have to be adjusted for it to be accurate to other virus behavior.

A. Part One

Using the framework of a linear dynamical system, we created a model which represented the progression of an arbitrary epidemic by using a matrix A containing flow rates between Susceptible, Infected, Recovered, and Dead population proportions to produce a set of vectors $\{x_1...x_T\} \in \mathbf{R}^4$ representing daily instance values for each proportion S, I, R, D over a period of T days. This is a SIRD (Susceptible, Infected, Recovered, Diseased) model and follows the formula

$$x_n = A^n x_0$$

where x_n represents the instances of SIRD after n days and x_0 represents their initial values.

To create this model, we first designed a matrix A based upon expected spread dynamics:

$$A = \begin{bmatrix} 1 - I' & F' & 0 & 0 \\ I' & 1 - R' - D' - F' & 0 & 0 \\ 0 & R' & 1 & 0 \\ 0 & D' & 0 & 1 \end{bmatrix}$$

Here, I', D', R' and F' represent 'infection', 'death', 'recovery', and 'failed recovery' (recovery without gaining immunity) rates, respectively. These words are in quotation because

they carry a particular definition. Specifically, they represent flow rates in a system in which constant proportions of susceptible people become infected, and of infected people die, and of infected people recover, each day. It is important to note that this matrix is set up in such a way that values of SIRD always sum to the same value, regardless of day. In this way, we represent a closed population.

We second, we populated these constants with arbitrary values in order to simulate an example epidemic over some number of days. We used the values 0.05, 0.01, 0.10, and 0.04 for I', D', R' and F' and simulated 200 days of activity.

We third, modified A with an additional constant L' (loss of immunity) to allow for some fraction of the recovered population to become susceptible again each day:

$$A = \begin{bmatrix} 1-I' & F' & L' & 0 \\ I' & 1-R'-D'-F' & 0 & 0 \\ 0 & R' & 1-L' & 0 \\ 0 & D' & 0 & 1 \end{bmatrix}$$

and simulated 1000 days of activity.

We lastly modified A again to remove the possibility of failed recovery F'. We did this in preparation to model actual COVID-19 data for two reasons. First, F' and L' can be considered a redundancy, as both represent loss of immunity, but over different time-scales. It is better to represent loss of immunity with a single constant. Second, we expected that in the case of actual COVID-19 data, few people will recover without gaining at least temporary immunity. The value F' is therefore unnecessary. The final A was represented with the following entries:

$$A = \begin{bmatrix} 1-I' & 0 & L' & 0 \\ I' & 1-R'-D' & 0 & 0 \\ 0 & R' & 1-L' & 0 \\ 0 & D' & 0 & 1 \end{bmatrix}$$

B. Part Two

The next phase of our experiment pertained to the particularization of our SIRD model to actual COVID-19 spread in the St. Louis Metropolitan area, finding the best possible values for I', D', R' and L' given the actual recorded trend of infections and deaths. Towards this effort, we used MATLAB functions *ss* and *lsim* to help us create the linear dynamical system, and the MATLAB function *fmincon* to optimize I', D', R' and L' based on a cost function of our design. This process was accomplished in several parts.

1) Data Preparation: Our COVID-19 data was prepared in order to enhance the effectiveness of our cost function. Specifically, it was converted from cumulative cases and deaths into SIRD time-series. To this end, certain assumptions had to be implemented into the prepared data. The current proportion dead at each day was determined to be equal to cumulative deaths at each day, normalized to proportion of overall population. The current proportion infected at each day was determined to be equal to the increase in cumulative cases over the previous two weeks, normalized to proportion of overall population. Current proportion recovered (and immune) at each day was determined by subtracting cumulative deaths and current infections at each day from cumulative cases at each day, taking the change in the result at each day over the previous six months, and normalizing to proportion of overall population. Current proportion susceptible at each day was determined by subtracting proportion dead, proportion infected, and proportion recovered at each day from one.

These preparations take our initial cumulative data of two variables and modify it into SIRD data of four variables directly comparable to model SIRD output. This direct comparability, while made at the sacrifice of the purity of our data, greatly enhances our cost function at comparatively little expense.

2) Cost Function: Our cost function is straightforward given our prepared data. It was determined with this modified percent error formula:

$$cost = \sum_{S,I,R,D} |\frac{||m-a||}{||a||}| \tag{1}$$

Where m and a represent model and actual S, I, R, and D time-series vectors respectively. Percent error is used in lieu of absolute error because S, I, R, and D accuracy should be weighted equally, and the result need not be multiplied by 100 to attain the percentage values because the constant scalar has no effect on relative costs.

3) Time Ranges: Our final SIRD model is a composite of several submodels representing subranges of the overall time series, where in each has its own constants. This is the case because the dynamics of COVID-19 spread in the St. Louis Metropolitan area changed over time–one set of constants is not enough to describe the full trends. These ranges were determined through graphical analysis of our prepared SIRD data. As we can expect rates of death, recovery, and loss of immunity to remain relatively consistent throughout our time series, we looked specifically at changes in the proportion infected in order to divide our time series into various 'waves.' Ultimately, this was an imprecise measure and a potential source of error.

4) Optimization Specifics: The function *fmincon* was used to optimize values for I', D', R' and L' as well as S_0, I_0, R_0, and D_0 in a linear dynamical system of matrix A for each range in order to best match the prepared SIRD data for that range. These optimizations were subject to the following constraints:

1) All optimized values must be positive.
2) S_0, I_0, R_0, and D_0 for each submodel must equal the final S, I, R, and D values of the previous submodel, or must equal 1, 0, 0, and 0 respectively for the first submodel. This ensures that our model is continuous and starts from the correct initial values.
3) I', D', R' and L' are bounded to between 0 and 0.1, 0.1, 0.1, or 1, respectively. This ensures that our constants are not 'cooperatively unfeasable.'
4) Optimization for each submodel must return the I', D', R' and L' constants (at a local minimum of cost) which

are closest to those of the previous submodel, or 0.05, 0.01, 0.10, and 0.04 respectively for the first submodel. In other words, they must optimize with minimum necessary change.

It may be noted that, due to constraint two, only I', D', R' and L' constants are being optimized, as S_0, I_0, R_0, and D_0 are always predetermined.

Running this model on our initial conditions, we produced a model SIRD time series that is most optimized for our time ranges. Cumulative cases were then calculated by multiplying S element-wise with I' to produce a time series of new infections at each day and producing from it a 'running tally' time series.

5) Policy Model: We went on to use this model to test a policy that would reduce new cases and deaths during the delta surge by 25%. The policy we tested was the enforcement of 24/7 mask wearing for 95% of the population. According to the Arizona Department of Health Services, wearing masks can reduce spread by 56% compared to wearing none [2]. Reference [1] states 35% of people already wore masks around the time of the delta surge. Therefore, the reduction in spread from our policy is $(95\% - 35\%) \times 56\% = 34\%$.

With this number, we ran our model from May 1, 2021 to November 1, 2021 with a new matrix A_{policy} wherein

$$I'_{policy} = I' \times (100\% - 34\%)$$

and recorded the change in infections and deaths across that time interval alongside the change in infections and deaths from the unaltered model across the same time interval. While this policy would reduce cases and deaths, this would ultimately not be a feasible policy because the government would not be able to force 95% of the population into wearing a mask all of the time without the policy being ruled as unconstitutional.

C. Part Three

We concluded our exploration of epidemic modeling with linear dynamical systems by attempting to draw insight from unrelated data through another specific instance of our general model. We sought to determine from data labeled mockdata the vaccinated proportion of a closed population across time and the proportion of a closed population experiencing breakthrough infection across time. In order to do so, we modified our matrix to account for vaccinated population and introduce constants for rate of vaccination of population and rate of breakthrough infections. The same Matlab functions were used again to optimize constants. The new entries of matrix A are as follows:

$$\begin{bmatrix} 1-I'-V' & F' & L' & 0 & 0 \\ I' & 1-R'-D'-F' & 0 & 0 & B' \\ 0 & R' & 1-L'-V' & 0 & 0 \\ 0 & D' & 0 & 1 & 0 \\ V' & 0 & V' & 0 & 1-B' \end{bmatrix}$$

where V' represents vaccination rate of non-infected, non-dead population and B' represents breakthrough infection rate. In the linear dynamical system that uses this matrix, $\{x_0 ... x_T\} \in \mathbf{R}^5$ and represent SIRDV proportions (Susceptible, Infected, Recovered, Dead, Vaccinated). Note that in this model, recovered individuals may lose their immunity, but they also are receiving vaccinations.

1) Data Preparation: No SIRD preparation was performed on this data, as reasonable assumptions could not be made about how long immunity lasts.

2) Cost Function: Our cost function for the mock epidemic follows the same principle as that of our COVID-19 cost function, but it does not have access to reasonable approximations of susceptible and recovered values.

$$cost = \sum_{I,D} |\frac{||m-a||}{||a||}| \tag{2}$$

Parameters m and a are model and actual time-series vectors (for infected and dead proportions).

3) Time Ranges: Our model is a composite of two submodels representing subranges of the time series. Specifically, they represent days 1-99 and 100-365. This is the case because at day 100, vaccinations began to roll out, and rates of vaccination and breakthrough infection changed from zero to some positive number.

4) Optimization Specifics: As in $Part\ Two$, the function *fmincon* was used to optimize constants I', D', R', B', V', and L' (infection, death, recovery, breakthrough infection, vaccination, and loss of immunity rates) and initial values S, I, R, D, and V in a linear dynamical system of matrix A for each range in order to best match the cases and cumulative deaths for that range. These optimizations were subject to the following constraints:

1) All optimized values must be positive.
2) S_0, I_0, R_0, D_0 and V_0 for the second submodel must equal the final S, I, R, and D values of the first submodel. Those for the first model must equal 1, 0, 0, 0, and 0, respectively. This ensures that our model is continuous and starts from the correct initial values.
3) Constants B' and V' for the first submodel are bound to equal 0.
4) All other constants for both submodels are bound to between 0 and 1. These values can not be assumed because the mock epidemic is not necessarily a normal or known disease.

5) Vaccinated and Breakthrough Time Series: The ultimate predictions for proportion vaccinated and proportion experiencing breakthrough infection were taken by pulling the time series V from the time series $SIRDV$, recording it, and multiplying it by the vaccination rate to produce a breakthrough time series. There are some immediate limitations to this model. For example, infection is modeled as a simple proportional function of susceptible population; however, in reality, infection is a more complicated function of susceptible population, infected population, and behavior. It is for this reason that the SIRD time series takes on a simple shape.

III. Results and Discussion

We found that our optimized linear dynamical systems were able to reasonably approximate both COVID-19 and mock epidemic data such that we were able to draw conclusions about the disease spread.

A. Part One

The implementation of the linear dynamical systems with the two versions of matrix A produced the two SIRD time series in Figures 1 and 2. These time series follow the expected flow rates, wherein all populations eventually flow to recovered and dead or dead, respectively. This demonstrates that the basic framework of A works in the modeling of an arbitrary disease-spread time series.

In the second model, everyone eventually dies. However, this is expected for our arbitrary values, and merely indicates that they are much too high. However, this phenomenon highlights an important motivation in our design of the COVID-19 model: the fact that for a linear dynamical system to properly model COVID-19, its rates of infection, death, recovery, and failed retention of immunity must be allowed to change over time, as deaths have been allowed to increase without bound for constant values thereof. This further justifies the need to use submodels.

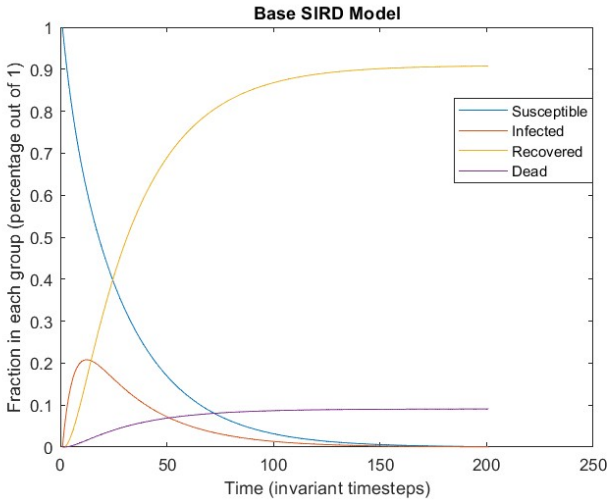

Fig. 1. SIRD model where it is not possible for people to be reinfected after recovering from the disease. Eventually, because the only categories where people must stay according to our model are recovered or dead, everyone will end up in one of the two.

These models showed us the extent to which each and every one of the choices we made had on our model, and how different mathematical interpretations of a real world scenario can promote different trends in data. This showed us although more than one approach is justifiable, assuming there is no lapse between the mathematical interpretation and the intention, there are differences in the final outcome.

B. Part Two

An important part of our modeling process was the transfer of our actual COVID-19 data to an approximate SIRD model. There are several key justifications as to why this can improve our model:

1) It allows us to incorporate additional knowledge about the disease into our data. Modeling COVID-19 infections as lasting two weeks and immunity as lasting six months encourages our SIRD model to conform better to actual dynamics of COVID-19 spread, where it might otherwise optimize for a model COVID-19 that does not infect individuals for very long, or that can reinfect more often than it should.
2) It also allows for a more direct cost function. A cost function with unmodified data would only have access to I and D information, of the total SIRD information available.
3) It may also allow for a more adaptable model. Because S, I, R, and D values are all optimized for, they are more likely to be individually accurate, while a system that only optimizes I and D values may only have SIRD values that are accurate in conjunction with each other under the specific circumstances in which the optimization is determined.

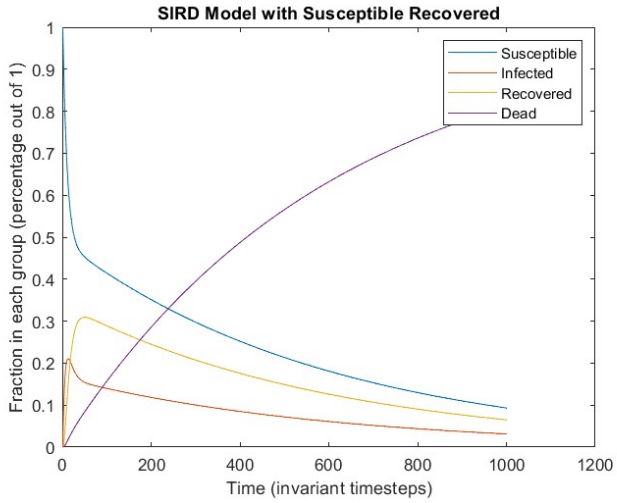

Fig. 2. SIRD model where it is possible for people to be reinfected after recovering from the disease. Eventually, because it is impossible for anyone to be cured forever, everyone will die.

Our optimized model SIRD time series is compared to our prepared SIRD time series in Figures 3 and 4. Our model cumulative infections and deaths are compared to real cumulative infections and deaths in Figure 5. It can be seen visually in both cases that our model creates a reasonable approximation of the actual data. Therefore, we can consider examinations using this model to be valid.

In our examination of a hypothetical policy change (requiring masks for 95% of the population) to have been enacted over the time period associated with the delta variant (May 1, 2021 to November 1, 2021), we projected a new model SIRD time series with the proper reduction in infectiousness, of which the I and D values are compared with those of our

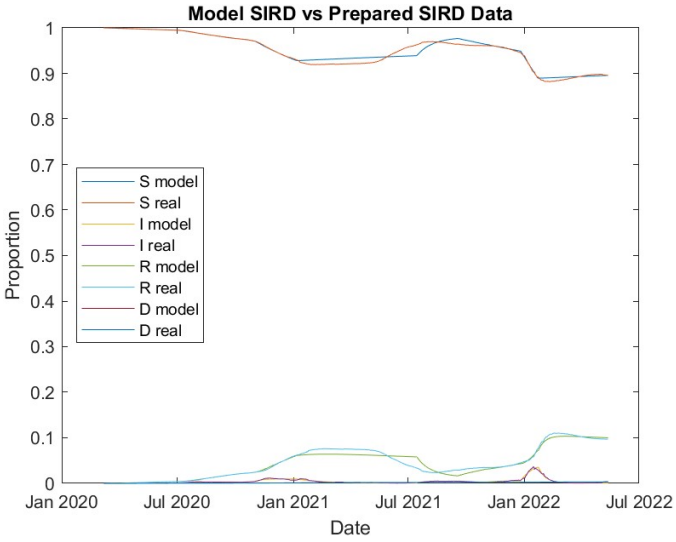

Fig. 3. Comparison between our prepared SIRD data and our modelled SIRD data. S and R can be seen to be predicted relatively well by our model.

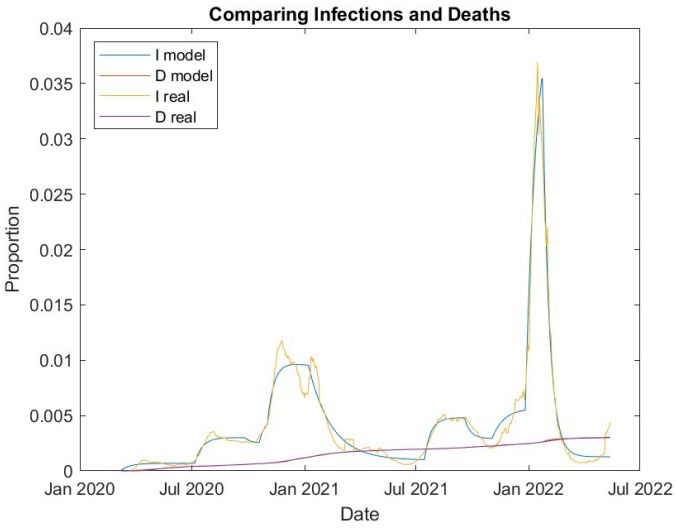

Fig. 4. A closer look at comparison between our prepared _I_D data and our modelled _I_D data. It can be seen that these, too, are predicted well by our model.

unaltered model in Figure 6.

C. Part Three

Part Three was an application of the generalized model, and created opportunity for analysis of trends dealing with vaccinations. The created model allowed for the comparison of how vaccinations affected COVID case counts in the real world versus through a mathematical interpretation. This model shows how a higher vaccination rate drastically reduces the number of people getting infected and dying. This largely supports the idea that mass vaccination efforts can help stop the spread of highly infectious diseases. The modeled data plotted against the real data can be seen in Fig. 7.

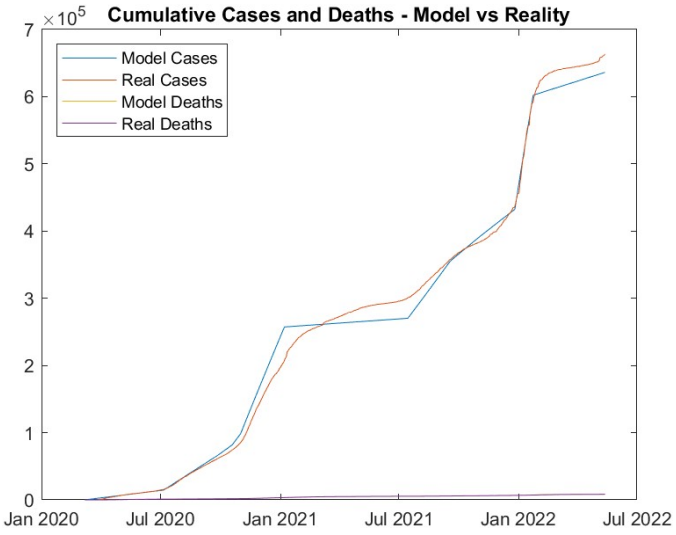

Fig. 5. Comparison between modelled and actual cumulative cases and deaths. This is the most direct measure of our model's accuracy.

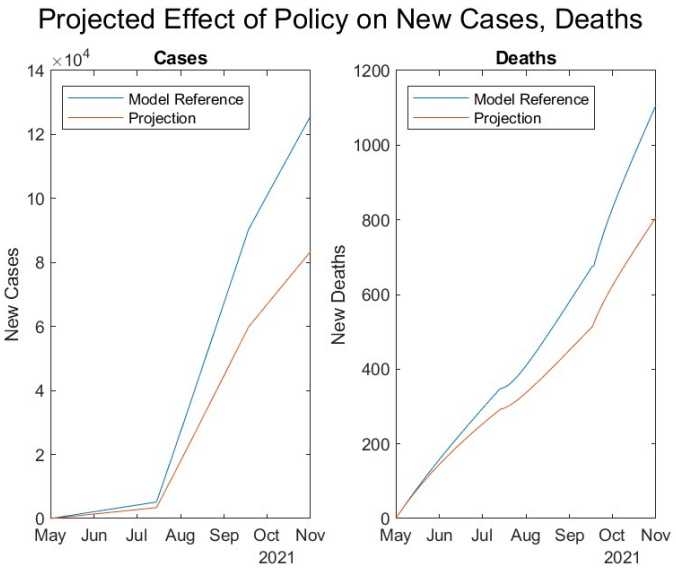

Fig. 6. Projected changes in cases and deaths between May 1 and November 1 2021 with our policy change, compared to model with no change. New cases are reduced by 33.6%, and new deaths are reduced by 27.2%.

IV. Conclusion

A. Part One

Part One served its purpose of creating a base SIRD model to then generalize in *Part Two* and apply in *Part Three* through recreating the effects of the model in MATLAB, which allowed for a better understanding of the SIRD model as well as creating an idealized progression of a virus given certain assumptions. By then making different assumptions, the graph resulted in different final outcome after t days, which showed the magnitude of the effect certain assumptions make on a purely mathematical model of a real life event. This was tantamount in deciding how to weigh certain design choices over others, and allowed for better

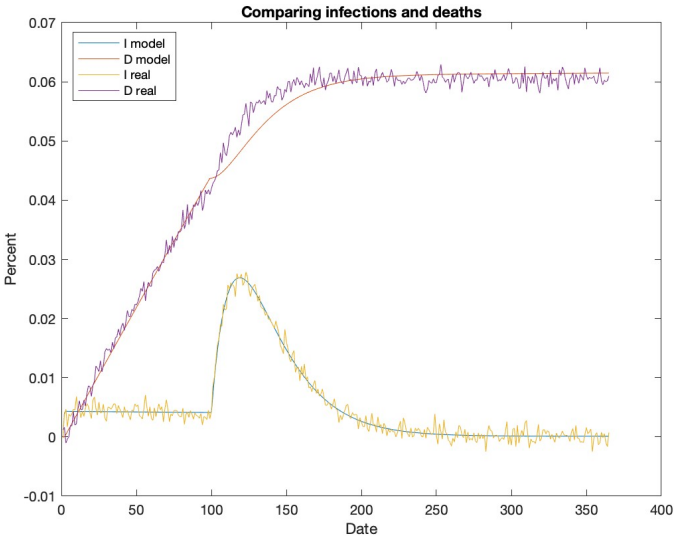

Fig. 7. The actual data plotted against the models projected data over time with vaccinations.

reasoning behind our assumptions. Later projects relating to this method of modeling epidemics would mean weighing the SIRD model against other epidemic models, and seeing how they compare to one another in their predicted results after t timesteps.

B. Part Two

This section has, undoubtedly, the most utility in terms analyzing epidemics based on the generalization of a mathematical model. The creation of this model spurred certain assumptions to solve problems with a purely mathematical interpretation of real world data. This idea is represented again in $Part\ Three$, as to account for new variables, new assumptions must be made. These assumptions, however, were certainly a weakness of mathematical modeling as a whole, and limited the scope of predictive accuracy our model could hope to achieve. These are the simplifying assumptions made in the model:

1) Average COVID-19 infections last exactly two weeks, and average immunity lasts exactly six months.
2) Everyone who recovers from COVID-19 has at the very least temporary immunity.
3) The population of St. Louis is closed to migration, birth, and non-COVID-19 death.
4) The dynamics of COVID-19 spread actually change from one time range to the next, rather than there being an unchanging dynamic between time ranges that our system is too simple to model.

In the future, new projects could weigh the benefits of this SIRD model over others to see the differences between the assumptions made to accurately model both to given data. This could, in theory, spawn a hybrid of the two models, or facilitate a deeper interpretation of the data.

In terms of improving our model, by using the standard deviation of the data over different periods of time, an algorithm could find the best mathematically justifiable time bounds given a set of waves.

C. Part Three

Through the process of creating a model for COVID in the Saint Louis Metropolitan area that accounts for vaccinations, several conclusions were made about the data in terms of its viability in creating an accurate representation of the epidemic using the given data and the accuracy of expansion of the SIRD model as a whole. Firstly, re-purposing the SIRD model for new variables showed the flexibility of linear dynamical systems as well as the limitations of the system, as the system created an near-accurate model of the given data, but necessitated human interaction to choose time bounds. This also means that the model could suffer in accuracy when trends are less apparent, as there are more necessitated breakages in the waves of the disease and the model will only decide on values of A based on the associated bounds of time, so there is more opportunity for human error.

Again, assumptions are a major factor in the weaknesses of the created model, as because of the nature of a mathematical model, it will always be necessary to bridge certain gaps between a near infinite number of variables in the real world and the mathematical one where, in the context of a linear dynamical system, all variables must have definition. This means the less relevant variables, such as people coming into the Saint Louis metropolitan area while being unvaccinated and people leaving Saint Louis while vaccinated, are not accounted for in the model, in this specific case setting the population as a static number.

Improvements in this section include approaching noise in the data through probabilistic means, meaning using methods like standard deviation to account for potential outliers in the data and smooth it. This, in theory, should improve the performance of the SIRD model, as without a large amount of data, it would be difficult to create accurate models without accounting for factors like outliers.

In terms of further exploration into this topic, experiments could be preformed to model vaccination rate with the SIRD model for other areas and comparing the assumed constant traits of one area to one another. For example, one could create the same model with New York city's COVID data and compare the population density of the Saint Louis Metropolitan area to New York's, or the rate of people coming in and leaving the cities, as New York is bound to have many more tourists and people not counted as part of the general New York population.

References

[1] Durkee, Alison."Here's Who's Still Wearing Masks the Most (and Least)." Forbes, Forbes Magazine, 21 Apr. 2022, https://www.forbes.com/sites/alisondurkee/2021/11/30/heres-whos-still-wearing-masks-the-most-and-least/?sh=3b6157dd5698.

[2] Rigler, Jessica. "Study Finds Mask Use Associated with Reduced Risk of Contracting COVID-19." AZ Dept. of Health Services Director's Blog, 7 Feb. 2022, https://directorsblog.health.azdhs.gov/study-finds-mask-use-associated-with-reduced-risk-of-contracting-covid-19/.