Simulation of an Epidemic Process

Seok Hyun Chang, Toorese Lasebikan, Lidia Ortiz Zamora, Anamika Shekhar, Abi Sivatharman

Decision Analytics for Business and Policy

Spring 2022

Table of Contents

Introduction

During the last 2 years, there has been a global effort to address the COVID-19 pandemic. Covid-19 has infected millions of people, and we are still in the midst of this ongoing pandemic. To prevent countless numbers of cases, many leaders in healthcare continuously propose solutions. With its high infection rate, it became clear that controlling the outbreak will depend critically on the successful implementation of policies. This project to understand how the pandemic has progressed, and how R_0 and infectiousness rate changed with policy implications using the SIR model. This approach will help us understand the effectiveness of policies, and if they helped minize the spread of the COVID-19 pandemic.

Our first goal was to look at how the reproduction number, R_0, has changed over time through the pandemic. This was done by developing a SIR model at a state scale to look at the previous patterns of COVID-19. For this model, we covered 5 states: California, Nebraska, Washington, Maine, and Texas. We intentionally decided to include the states that have both high and low population densities to observe whether different policies had consistent effects regardless of population level. After looking at reproduction numbers, we analyzed how the infectiousness of the virus changed over time with respect to policies in such states. We compared variables such as Mask use, Mobility Level, Test per 100,000, and Vaccination Status. After the analysis, we were able to successfully see the effects by visualizing the results. These models highlight the significance of fully implemented policies and limited transmission of disease.

Analytical Framework

The dangers posed by pandemics like COVID-19 compel us to understand the nature of these epidemics - the best we can. Thus, modeling the epidemic process becomes critical to help inform policy response. A standard model, within epidemiology, for predicting how the disease spreads throughout the population and the effect of different interventions is the SIR model. The SIR model divides the population into mutually exclusive groups according to their infection state: susceptible, infected, or recovered.

The susceptible population includes individuals not yet infected but who can become infected. The infected population is those that are infected and can spread the virus. Lastly, the recovered population includes individuals that have recovered from the virus. The evolution over time of these three populations – $S(t)$, $I(t)$, and $R(t)$ – is given by the following system of ordinary differential equations:

$$\frac{dS}{dt} = -\beta(t) \times S(t) \times \frac{I(t)}{N} \tag{1}$$

$$\frac{dI}{dt} = \beta(t) \times S(t) \times \frac{I(t)}{N} - \gamma \times I(t) \tag{2}$$

$$\frac{dR}{dt} = \gamma \times I(t) \tag{3}$$

In these equations, *S(t)* and *I(t)* denote the number of susceptible and infected individuals at time t. *R(t)* includes the number of individuals removed from the system or have recovered at time t. On the other hand, the parameters S_0, I_0, and R_0 describe the initial state of the system (i,e., the initial number of susceptible, infected, and recovered individuals). The parameter beta, β, represents the expected number

of people an infected individual infects over one time period (i.e., day or week) – the infection rate. The parameter gamma, γ, is the probability an infected individual will recover within one time period – the recovery rate.

The rate of change of the susceptible population over time is given by **Equation 1**. We can calculate this rate of change by multiplying the total number of susceptible individuals in time t, *S(t)*, by (1) the number of people a single individual can infect per unit of time, β, and (2) the probability of being in contact with an individual who is infected. This rate of change is negative because the susceptible population decreases as individuals jump from this state to the infected state. The rate of change of the infected population over time (**Equation 2**) subtracts the fraction of recovered individuals in time *t* - the number of infected individuals in time t, *I(t)*, multiplied by the probability of recovery, γ - from the newly infected individuals (obtained in **Equation 1**). The rate of change of the recovered population over time is given by Equation 3, which is simply the fraction of recovered individuals in time *t* (from **Equation 2**).

Another critical parameter of the SIR model – and epidemiology – is the basic reproduction number or R_0. The basic reproduction number gives us the expected number of individuals a single infected individual infects over their entire infectious period. Thus, if $R_0 > 1$, there is an epidemic as the number of infected individuals grows exponentially. On the other hand, if $R_0 < 1$, the disease dies out. R_0 is simply given by the ratio between beta and gamma.

Like any model, the SIR model makes several assumptions. The model assumes that the population is fixed and homogeneous. With a homogeneous population, every individual has the same age, health conditions, duration of infection, and is in contact with the same number of people. With a fixed population, the total population at any given time is determined by **Equation 4**.

$$N(t) \; = \; S(t) \; + \; I(t) \; + \; R(t) \tag{4}$$

The model also assumes that the incubation period is zero (i.e., once an individual becomes infected, they can immediately spread it to others) and once recovered, individuals are immune and no longer able to infect others.

Data Summary

The data used for the implementation of the SIR models was from the New York Times COVID data repository. The data consisted of cumulative (daily) counts of U.S. COVID-19 cases and death at the state level over time. We also used data from the Institute for Health Metric and Evaluation (IHME) which contained daily data on vaccine coverage, mask use, and mobility at the state level over time. A full data dictionary is included in the appendix.

Exploratory Data Analysis

When we compare the infected cases in the five selected states, we see each state experienced two spikes in infections (**Figure 1**, in appendix). We can also observe that Nebraska consistently had the lowest proportion of cases. On the other hand, California and Texas experienced high levels of infections. Not

surprisingly, when exploring COVID-19 deaths among these five states (**Figure 2**, in appendix), we see Texas and California at the top - while Nebraska has the lowest number of deaths. These visible differences in both infected cases and deaths raise questions about policy responses by the local government in these states. We aim to answer some of these questions in Part 2 of our analysis. In the next section, Part 1 of our analysis, we implement the SIR model to help us understand how the basic reproduction number changed throughout the epidemic.

Implementation Details

The code for the SIR model was written in Jupyter Notebook using Python3. The Python libraries used for this project were *itertools*, *numpy*, *math*, *pandas*, *gurobipy*, *matplotlib.pyplot*, and *scipy.integrate*. For this project, we assumed it took fourteen days, or two weeks, to recover from the COVID-19 infection. Therefore, we calculated the number of COVID-19 infections per day by first inserting fourteen zeros at the beginning of the array that contains the cumulative (daily) counts of COVID-19 cases. Then, we deleted the last fourteen elements of this newly created array, so that it would be the same size as the array it originated from. Finally, to create an array containing the number of COVID-19 infections per day, we subtracted the newly created array (containing fourteen zeros at the beginning) from the original array (the array containing the cumulative (daily) counts of COVID-19 cases). This calculates the number of COVID-19 infections per day, as it removes people who have been infected after fourteen days, assuming that they have recovered. After creating new arrays that contain the number of COVID-19 infections per day for each state, we then plotted the number of infections per day for each state. We used these infection graphs that represent the actual infection curve to help answer our two research questions. Moreover, for our SIR model, since we assumed that it would take people infected with COVID-19 fourteen days to recover, we calculated gamma, γ, to be 0.071 ($1 \div 14$). We assumed beta, β, to be gamma multiplied by R_0.

Research Question 1

For our first research question, we wanted to know how R_0 changed throughout the pandemic. We had to consider multiple R_0 values for the pandemic, as R_0 can change due to factors such as the variant of the virus and social distancing behavior. First, we divided each infection curve into time periods (or epochs) by analyzing the increasing, decreasing, and flat trends of the curve itself. For example, if a part of the curve was increasing, we would assign that increasing trend to be a time period. As seen by the actual infection curve for Texas below, a time period (or epoch) was assigned for day 100 through day 150, as the curve is increasing for this time range.

Number of Infected in Texas
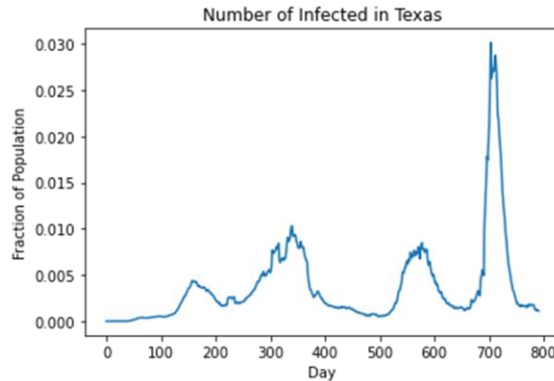
Then, for each time period, we manually picked an R_0 value. We then ran the SIR model using the R_0 values manually chosen, plotting the infection curve the SIR model simulated over the actual infection curve. We repeated the process, manually adjusting the R_0 values until the infection curve the SIR model produced (the simulated infection curve) closely matched the actual infection curve. The simulated curve is plotted with the actual curve for Texas below. The other simulated curves are included in **Figure 3** of the appendix.

Texas
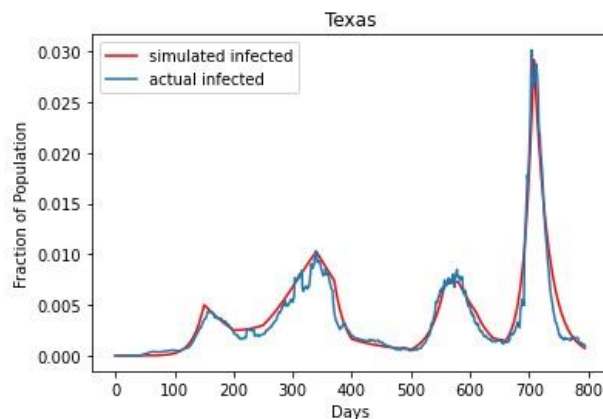
Research Question 2

For our second research question we wanted to know the change in R_0 and infected cases with testing, mobility, masking, and vaccination keeping in mind the variants of COVID. We will first look into testing patterns that changed with R_0 values. Looking at the plots below, we can see that testing kept on increasing till 1 year after the first cases, and then there is a huge drop which was around August when things were going back to normal and it spiked up to around the time of omicron. We can also see that whenever there is a spike in basic reproductive numbers there is also a spike in testing and this is consistent for the 5 states.
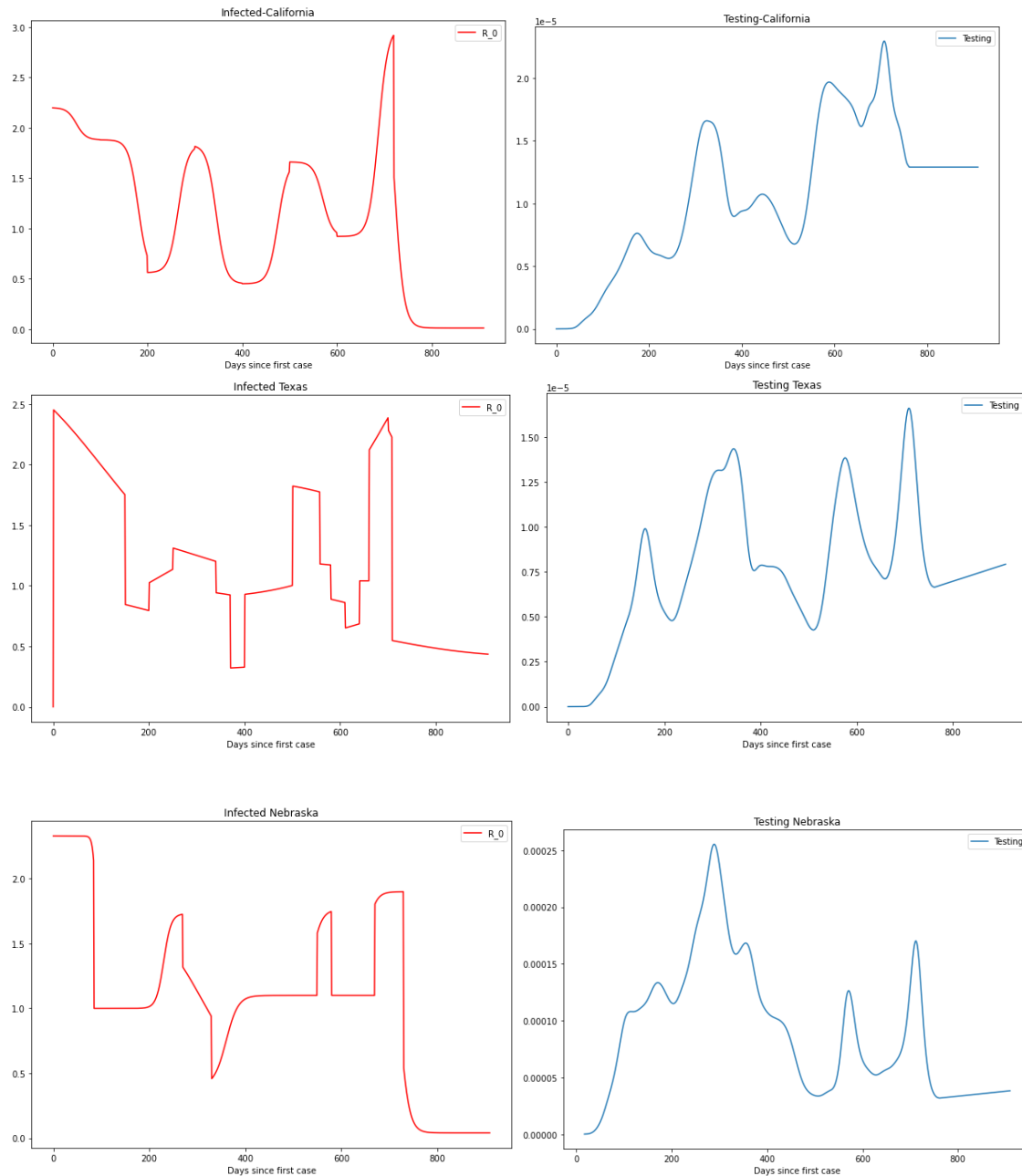
Figure :Testing vs R0

Next we moved on to look into changes in mobility. In order to plot the mobility and R_0 on the same axis, we scaled our R_0 value. Looking at the plots below we can observe that with lockdown in the picture there is a sudden drop in mobility which takes a time of 1 year to go back to the mean. Similar to testing, the mobility improves when R_0 decreases, and as R_0 increases the mobility drops. In Maine, the improvement is maximum as R_0 is stabilized.
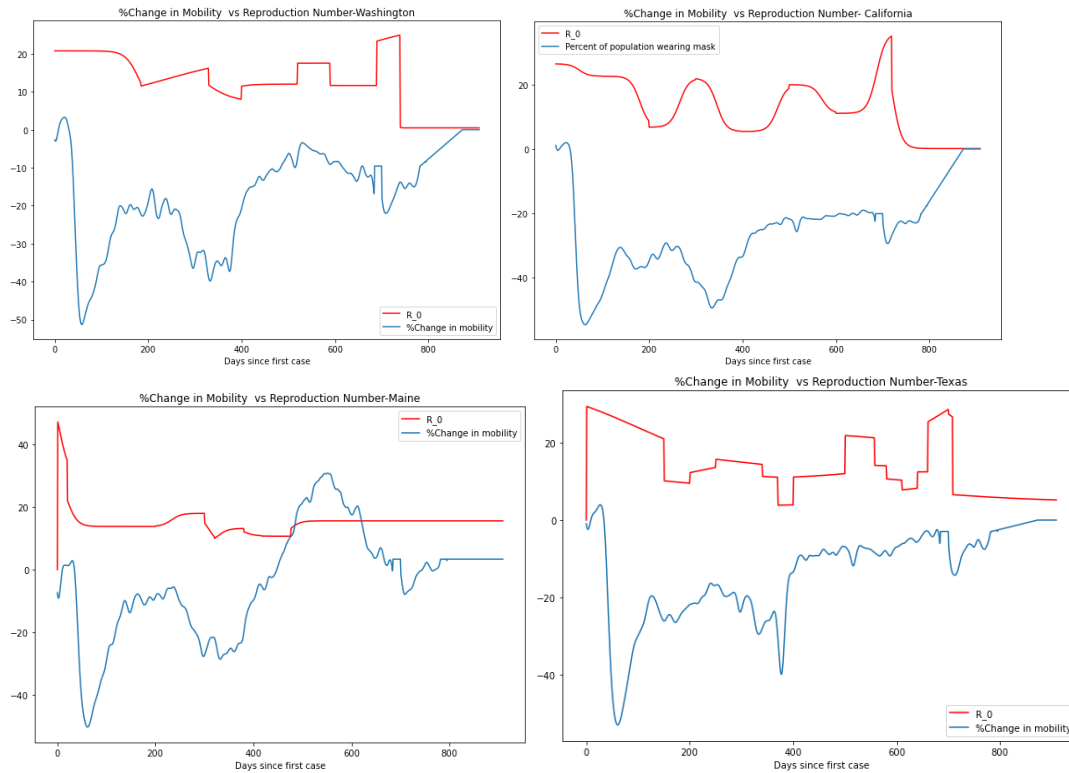
Figure: Mobility vs R0

We then studied the percent of the population reporting always wearing a mask and its change with R_0 values. Unlike testing, there are barely any abrupt changes in the masking pattern shown below. We can see that the highest masking percentage was reached when R_0 was still low, this corresponds to the time when people were coming out of their houses but the mask mandate was still in action. Over time we can see a dip but it again picks up during omicron.
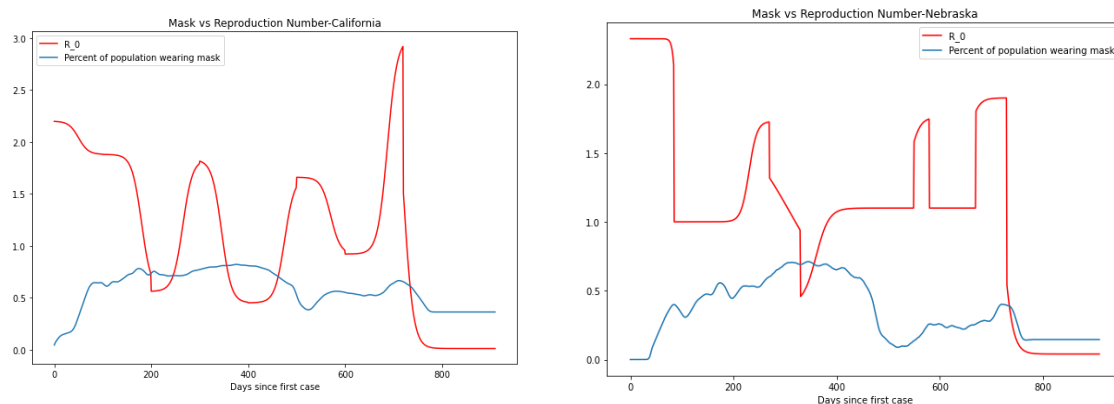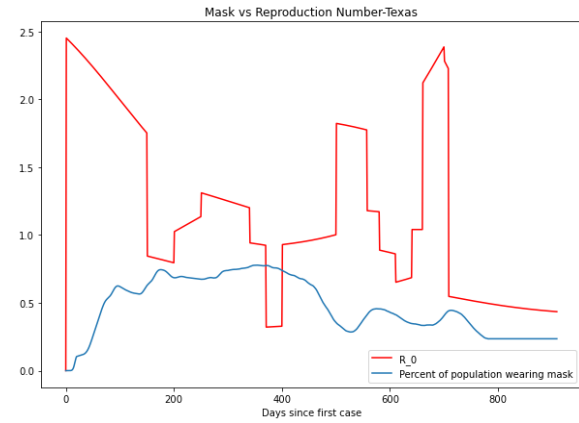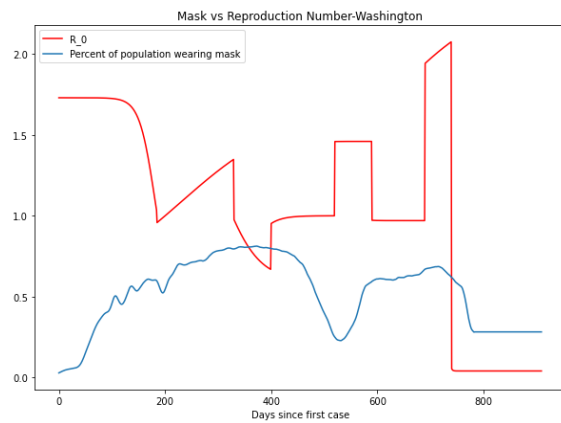
Lastly, we looked into vaccinations, we can clearly see post delta there is a spike in people getting vaccinated post the delta variant breakout. In Maine, even though cases are not that high but vaccination is more but in Texas, the cases are higher than vaccination.
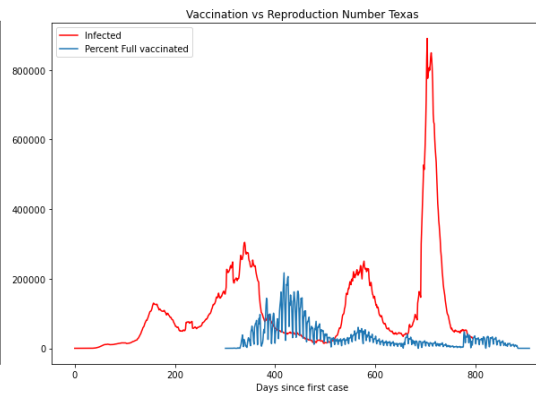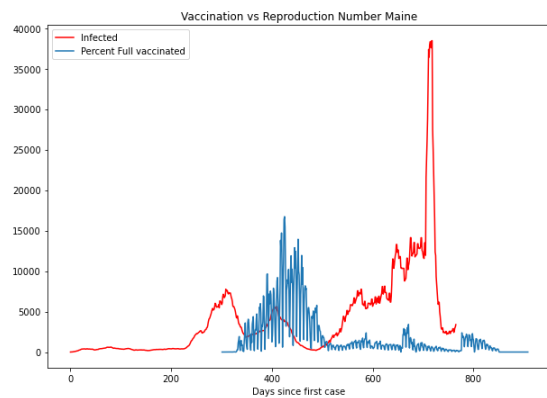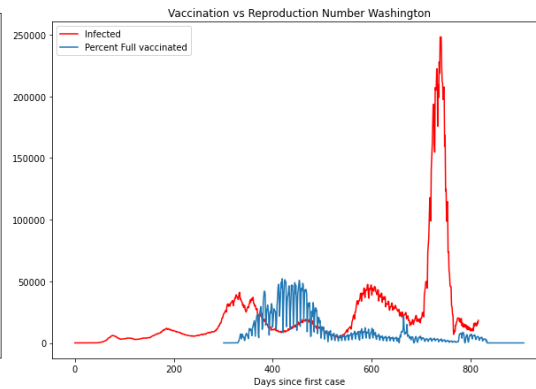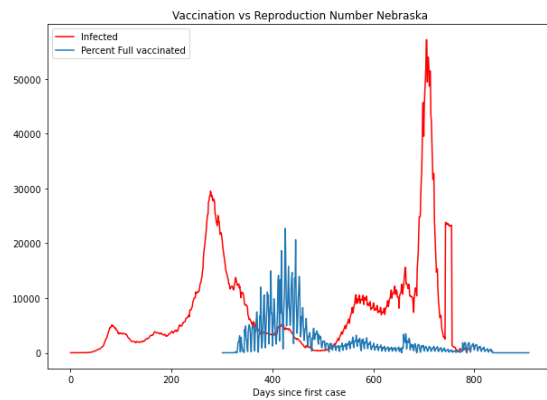
Conclusion

This model is based on the SIR framework. It builds on the assumption that transfer between people is done in a deterministic way. This simply means that people move from one state of the system to the next state. However, the study of the COVID virus has shown us that transfer can also be done in an indirect manner. The best way to solve this is to incorporate a stochastic process into the model. This helps to ensure that the probability of each person in any state in the system can be continuously measured. Building in such a fail-safe ensures that your model is able to more accurately understand epidemic patterns. A second limitation of the model is that recovery dates are constants. In a practical situation, this is not the case. Recovery dates differ significantly among the members of the population. For example, covid has shown us that a not insignificant part of the population suffers from "long covid". As a result, the actual recovery rates vary significantly. Finding a manner to reflect the different days of recovery might help to simulate a more accurate model.

This model does take into account the impact certain policy "interventions" can have on the rate of covid transmission. However, there is no way to quantitatively measure the impact of some of these policies on the population. For example, the model showed that lockdown and lack of mobility led to a reduction of transmission rates in certain states. On the other hand, there is no way to see the trade-off that occurred in communities where this lack of mobility was detrimental to their mental health.

In conclusion, a model by its very nature is an approximation of reality. Despite this, the model in this case study did a very significant job in simulating the covid case. It is very obvious and necessary for a model to incorporate more factors to accurately depict the complex realities of the world. A model that does this might be useful, but it runs the risk of being unnecessarily complex and leading to overfitting of your data. While this leads to low bias, there is also the possibility of introducing a high amount of variance into your model. This increases the difficulty of analysis and implementation. Hence, all the advantages of a simple model end up being lost.

Appendix

[Institute for Health Metric and Evaluation (IHME)](#) The IHME website contains three data sets of interest (1) Reference scenario 2020, (2) Reference Scenario 2021, and (3) Reference Scenario 2022. These data sets contain multiple reported COVID-19 related measures and projections - including deaths, hospital resource use, infections and testing, mask use, and mobility. The data sets include daily data at the state level. We found this data after some research on policy interventions at the state level. For our second analysis - the relationship between policy interventions and R_0 - we focused on three specific measures:

- Mask Use: percent of the population reporting always wearing a mask when leaving home
- Vaccine Coverage: counts of initially vaccinated (one or two doses)
- Mobility: the percent change in mobility from baseline

[New York Times COVID Data](#)

The New York Times data repository contains cumulative (daily) counts of U.S. COVID-19 cases and death at the state level over time. We incorporated this data in our first analysis (using SIR models to determine the change of R0 over time). This data source was a suggestion by Peter Zhang, Assistant Professor of Operation Research at Carnegie Mellon University's Heinz College of Information Systems and Public Policy.

Cases of Infections and COVID-19 Deaths

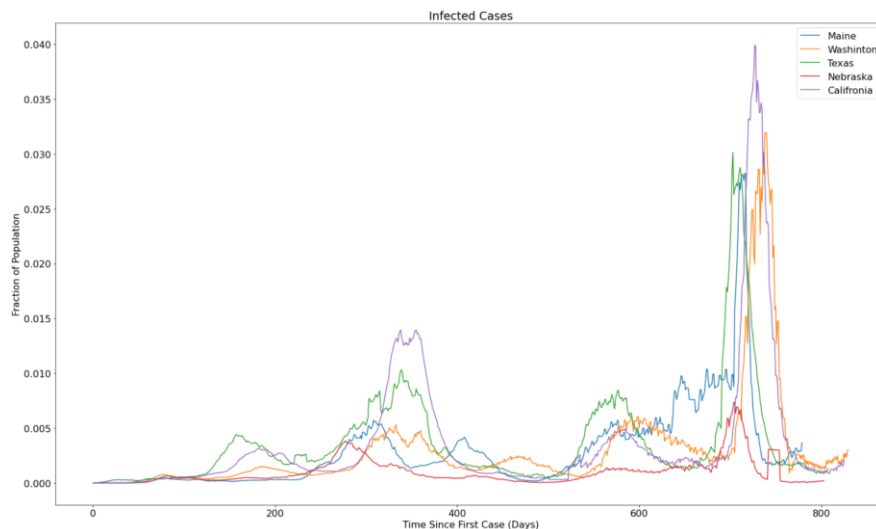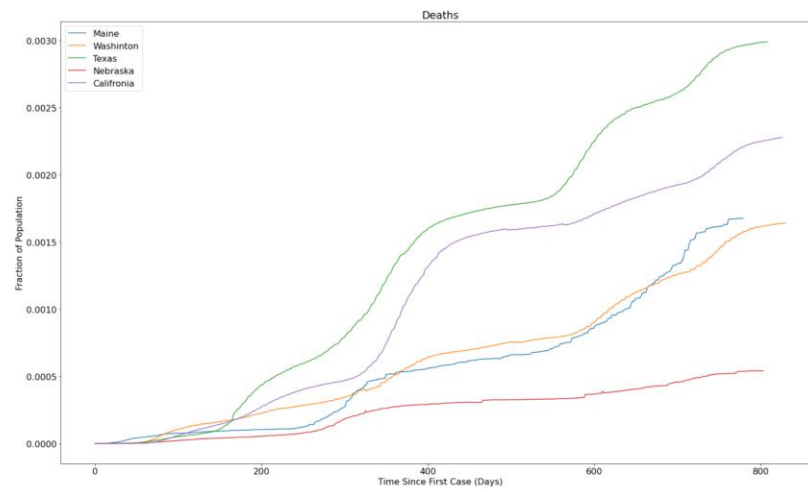**Figure 1**. Cases of Infected Over Time for Selected States

Figure 2. Covid-19 Deaths Over Time for Selected States

SIR Model Simulations

Figure 3: Simulated Infection Curves