CSE 246 Project: COVID-19 Cultural and Intervention Exploration and Spread Forecasts

Niharika Srivastav
nsrivast@ucsc.edu
UC Santa Cruz

Samaa Gazzaz
sgazzaz@ucsc.edu
UC Santa Cruz

Priya Padmanaban
ppadman2@ucsc.edu
UC Santa Cruz

Katelyn Suhr
ksuhr@ucsc.edu
UC Santa Cruz

ABSTRACT

With the onset of COVID-19 affecting hundreds of thousands globally, we wanted to explore how cultural aspects play a role in the rate of infection and the spread of the virus. In addition, we are interested in using available COVID-19 data in order to perform time series forecasting and predict peak projections. The goal of this project is to study the effect of social distancing and government intervention factors on the spread of the curve of COVID-19. We want to learn from the forecasted models whether our assumptions were valid by comparing those two methods of modeling a pandemic. As a result, we can gain insight from countries who have already experienced a peak in confirmed cases and learn whether social distancing and government intervention factors work, in order to help other countries minimize the rate of infection.

1 INTRODUCTION

COVID-19 is a disease caused by the virus SARS Coronavirus-2, which is a a type of coronavirus. It is currently a global pandemic and has caused mass casualties. This virus has caused over 400,000 deaths globally and about 100,000 in the United States itself, by date. We have analysed COVID-19 effects in mainly two ways: Peak prediction to predict the infection curve for different countries and SEIR Model Simulations to simulate the effects of different social and cultural aspects of different countries on COVID-19 numbers.

The peak predictions are based on real data about infection rate, confirmed cases, deaths and recovered cases. This analysis helped us predict the future curve in the countries and states under consideration. On the other hand we made simulations for infection curves based of the reproduction number in different states/ countries. These simulations help us understand the different effects of COVID-19 in socially and culturally different places and if these socio-cultural aspects have affected the spread of COVID-19.

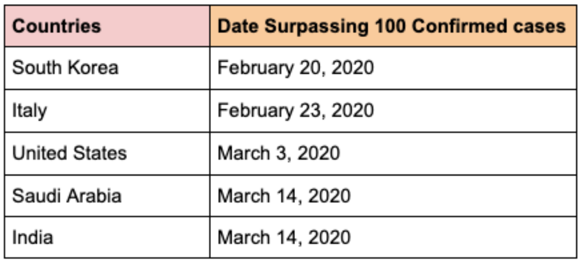
2 EXPLORATORY ANALYSIS

Data sets that were used to perform our data analysis and spread forecasts include the COVID-19 data repository provided by Johns Hopkins University Center for Science and Engineering (JHU CSSE) [11]. This data repository is open to the public and contains worldwide time series data beginning January 22, 2020 for a comprehensive list of province/state and country/regions. The time range we performed our analysis on was from January 22, 2020 to May 29, 2020. Johns Hopkins data repository includes data for the total number of confirmed cases, deaths, and recovered. We decided to choose a subset of countries so we can focus on select features for our simulation models and compare the results to our spread forecasts. The countries chosen were: India, Saudi Arabia, and the United States. Within the United States, we focused on three states: New York, California, and Iowa. We selected these states and countries because of the differences in population, location, and cultural intervention factors.

2.1 Confirmed Cases of COVID-19 Visualization

These graphs help to provide a visual analysis for what the trends look like for confirmed cases in Saudi Arabia, India, and the United States. Furthermore, we used these visualizations to decide which model would best fit our data for spread forecasting of COVID-19. Among these three graphs, we see in Figure 2, that the United States has the highest amount of cases as of May 29 with over a million confirmed cases. The amount of cases with respect to date provides a line that looks almost linear. This is quite different compared to Saudi Arabia and India, Figures 3 and 4 respectively, which both have a line that looks more like a logistic curve.

We assume that at some point in the future, the curve will flatten as the rate of infection slows down and number of cases eventually decrease which would represent the logistic curve [6]. We make this assumption due to data exploration from countries that have been affected earlier and have already seen the curve of confirmed cases flatten such as Wuhan China, South Korea, and Italy. Figure 1 shows the dates for when the total number of confirmed cases have passed 100 cases.



Countries	Date Surpassing 100 Confirmed cases
South Korea	February 20, 2020
Italy	February 23, 2020
United States	March 3, 2020
Saudi Arabia	March 14, 2020
India	March 14, 2020

Figure 1: Confirmed cases surpassing 100 cases from Johns Hopkins Data

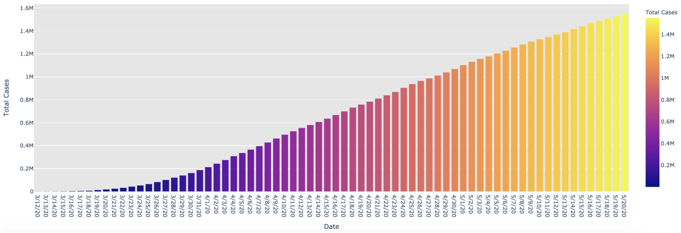


Figure 2: United States Confirmed Cases May 29: 1,746,019 cases

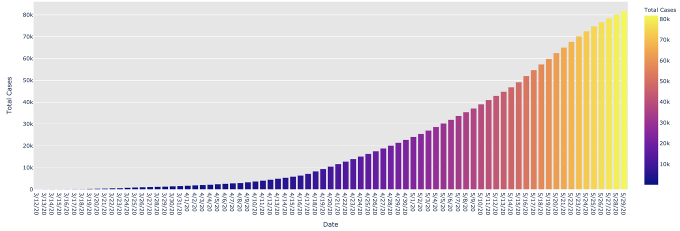


Figure 3: Saudi Arabia Confirmed Cases May 29: 81,766 cases

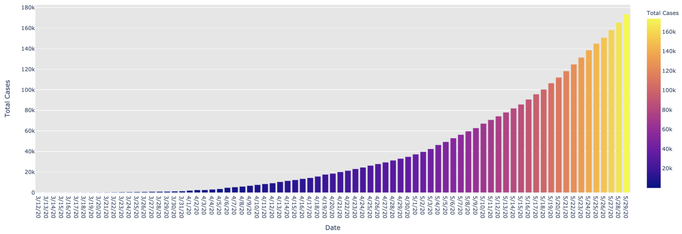


Figure 4: India Confirmed Cases May 29: 173,491 cases

3 SPREAD FORECASTS

Our goal was to use the available data as well as the information we had gained about the data through exploratory analysis to forecast the spread of COVID-19. The approach we took was to optimize a parametric curve fit of the currently available data with two well known growth functions: the logistic curve and the Gaussian curve. Then, we extended this curve to dates in the future to create a forecast for the spread. For regions that had not hit their peak cases yet, this also projects a value and date for the peak. These forecasts were done on both total cases and active cases for India, Saudi Arabia, South Korea, United States, and a few states within the US: California, New York, and Iowa.

We believe that using a curve fitting approach would be a good method for forecasts because there exists popular off-the-shelf forecasting models such as Prophet (made by Facebook's Core Data Science team) which is also a type of curve fitting model. The Prophet model takes into account trend, seasonality, and holidays by using either a piecewise linear or logistic growth curve for modeling time-series data [3]. Our model does not take into account effects of holidays or seasonality but it is based on real data provided by Johns Hopkins. The following sections will explain each growth function used and our results.

3.1 Logistic Curve Fitting

For modeling the total number of confirmed cases, we decided to use a logistic function since the number of confirmed cases can not go beyond the total population of the selected country and we'd eventually see the curve flatten [6]. The logistic function is a growth function that can model scenarios in which there is an increasing growth rate in the beginning and a decreasing growth rate towards the end. This makes it a good choice for scenarios such as popular growth or the spread of a disease in terms of total cases, which is the way we are using it here.

The logistic function we used is shown below:

$$f(x) = \frac{c}{(1 + e^{-k(x-m)})}$$

c = curve's capacity
m = sigmoid's midpoint
k = logistic growth rate
x : x ε ℝ

In order to minimize the error of the logistic fit on the time series data we have available, we used SciPy, a free open-source Python library [12] that contains a wide range of computational methods such as their *optimize curve fit* module. This module was used to find the most optimal coefficients and parameters to optimize the curve fit. Once we had the best fitting model, we could use it to create a forecast. The forecast was done by extending the dates into the future and continuing the curve to produce daily predictions. Figures 5, 6, 7, 8, and 9 show the logistic fit and forecasting for the countries of India and Saudi Arabia, as well as the US states of California, New York, and Iowa.

3.2 Logistic Curve Fitting Results

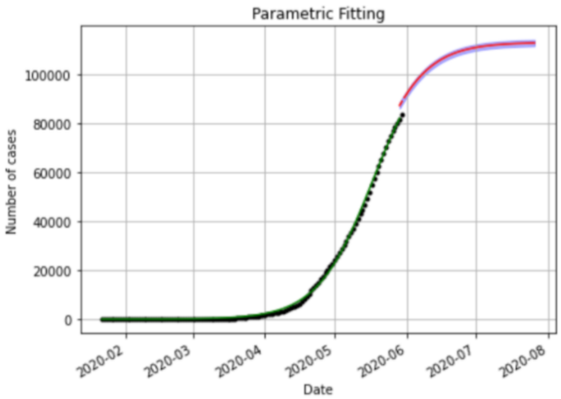


Figure 5: Saudi Arabia Logistic Fit for Confirmed Cases

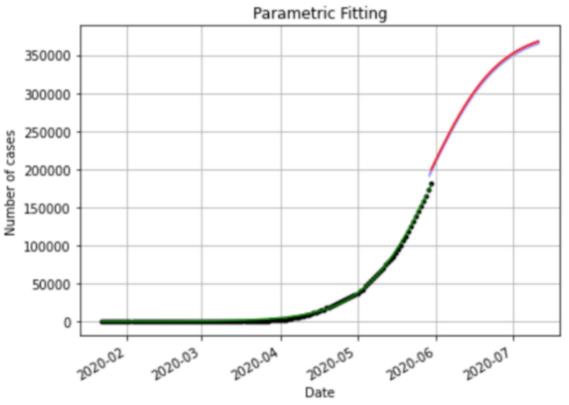


Figure 6: India
Logistic Fit for Total Confirmed Cases

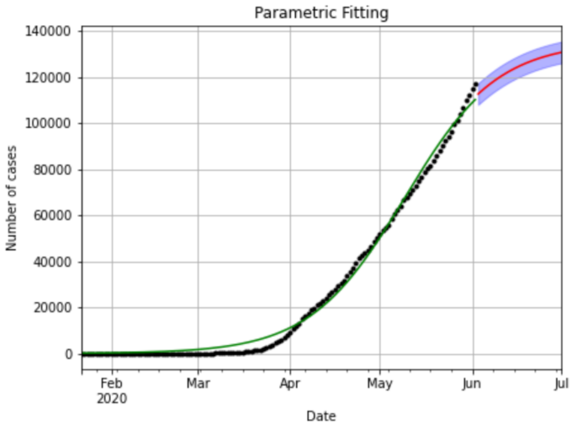


Figure 7: California
Logistic Fit for Total Confirmed Cases

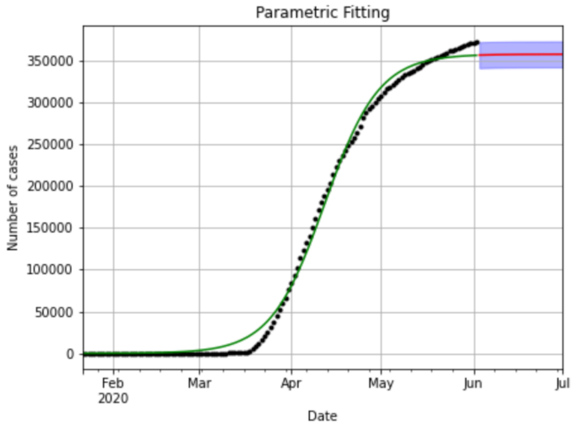


Figure 8: New York
Logistic Fit for Total Confirmed Cases

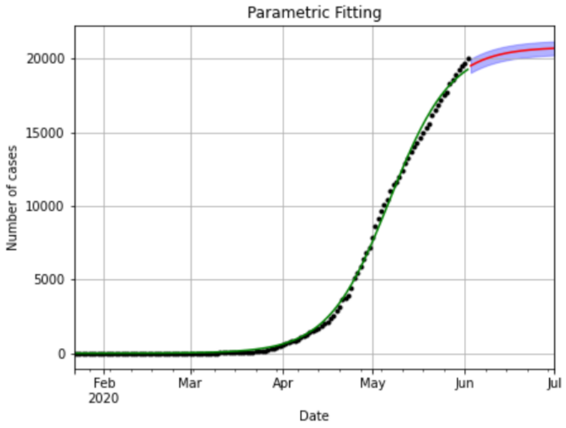


Figure 9: Iowa
Logistic Fit for Total Confirmed Cases

3.3 Gaussian Curve Fitting

To model the active cases on a given day, we used the Gaussian curve. Although this is not a perfect choice for this data, the normal distribution is a good choice for modeling many natural phenomena, including active cases of an infection. Similarly to the logistic curve, this shows a rapid growth in active cases followed by a rapid decline. For India and Saudi Arabia, we were able to calculate daily active cases by subtracting total recoveries and total deaths from the total confirmed cases. For the US states, however, this data was not available, so we instead modeled on new cases data. This is not an exact representation of true active case data, but we felt that it was a close approximation of the general trend using the data that was available to us.

Just as in the logistic curve fitting, we used the *optimize curve fit* module from SciPy to fit the above Gaussian function. Using the generated best fit parameters, we then plot the function on top of known data. We extend this function to generate forecasts for dates in the future. Figures 10, 11, 12, 13, and 14 show the Gaussian fit and forecasting for the countries of India and Saudi Arabia, as well as the US states of California, New York, and Iowa.

The Gaussian function we used to curve fit is shown below:

$$f(x) \;= ae^{-\frac{(x-b)^2}{2c^2}}$$

a = height of curve's peak
b = center of the peak
c = standard deviation
(controls width of the bell)

3.4 Gaussian Curve Fitting Results

These graph results contain black dots which show the data points from Johns Hopkins data repository. The green line is the model we've fitted and the red line is the prediction. As you can see below, our Gaussian curve fits better for the Active Cases in Saudi Arabia and India, compared to the new confirmed cases for the states New York, California, and Iowa.

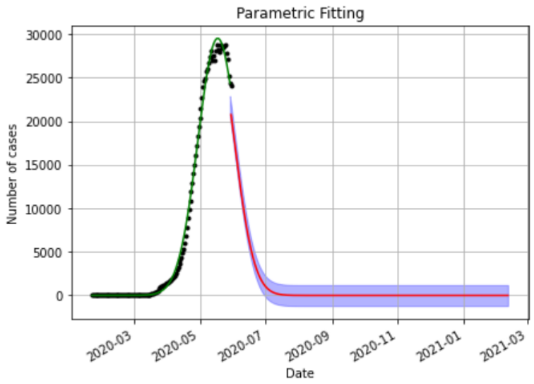


Figure 10: Saudi Arabia
Gaussian Fit for Active Cases

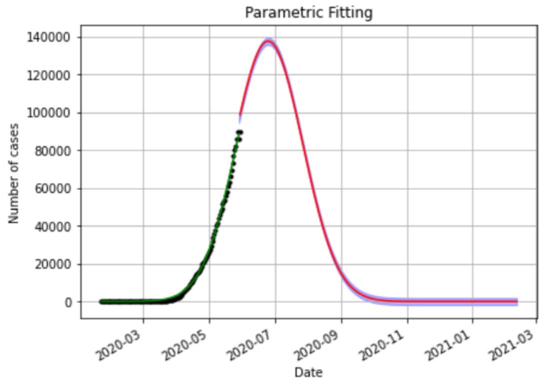


Figure 11: India
Gaussian Fit for Active Confirmed Cases

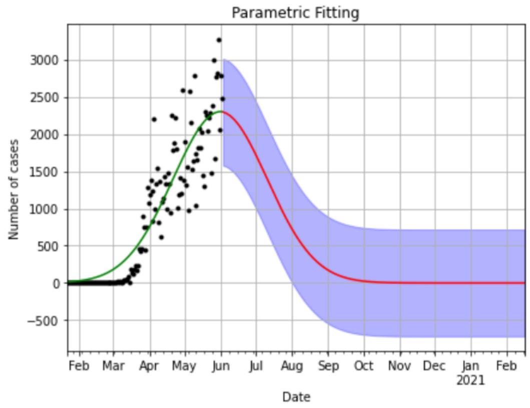


Figure 12: California
Gaussian Fit for New Confirmed Cases

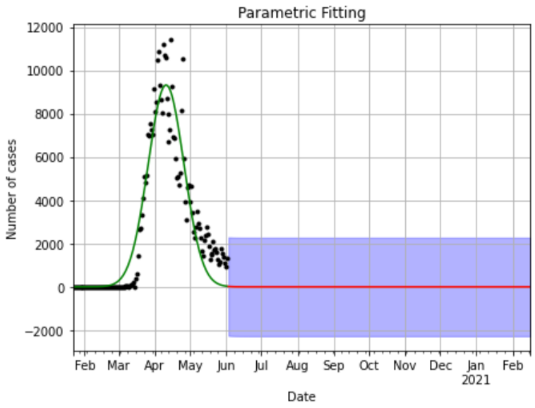


Figure 13: New York
Gaussian Fit for New Confirmed Cases

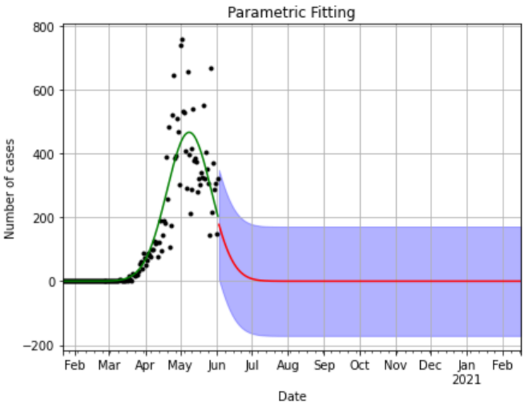


Figure 14: Iowa
Gaussian Fit for New Confirmed Cases

4 CULTURAL AND INTERVENTION EXPLORATION

In order to learn more about the factors affecting the spread of COVID-19, we approach this by simulating spreads using the parameters we believe have the most effect on the spread. This will enable us to look into the future and see what would a spread look like based on different variations of these parameters. Looking at the simulated spreads, we can compare them to the forecasted spreads and deduce the parameters that actually affected those trends based on the similarities and differences between different spreads.

In this section, we will start with an overview of the model used for simulating the pandemic. Next, we will discuss the different factors and parameters considered and collected for the simulations. Finally, we share the simulations for the countries chosen before based on the different variants of the factors in consideration.

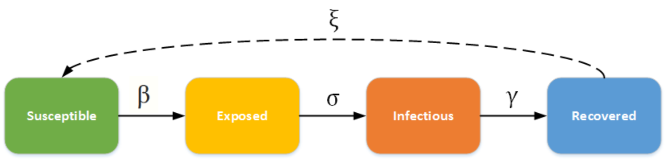


Figure 15: SEIR model parameters

4.1 Model for Simulating a Pandemic

Here we will talk about the SEIR model, it's parameters and how it ties with our purposes. The SEIR Model is a compartmental model in epidemiology which is used to study the effects of a pandemic on a given population. In this model the population is divided into four groups: Susceptible, Exposed, Infectious and Recovered, as shown in Figure 15. The sum of these groups of people is assumed to be constant. The infection rate β is the rate at which susceptible people become infected. σ is the incubation rate which represents the time duration when an individual is infected but doesn't show any symptoms of the disease. The rate of recovery is represented by the letter γ. The most important number is the Reproduction Number which is the ratio of the Infection rate to the Recovery rate. This number represents the number of Susceptible people that can get infected at an average by an infected person. We used this parameter in our simulations to represent some of the social and cultural parameters.

4.2 Factors Considered

There are many factors that could have greatly affected the spread of COVID-19 and determine how flat the curve would be. In order to come up with simulations that reflect the real spread of the curve, we collected data about the two countries (India and Saudi Arabia) and three states (California, New York and Iowa) on the reproduction factors, the healthcare capacity, the dates at which government enforced interventions, and whether curfew or lock-downs were enforced as shown in Table 1. We vary the reproduction rate based on the specific information regarding each country. We considered the US states separately because within the United States, COVID-19 had significantly different effects in each state. The governmental interventions were also quite different in each state, which becomes proportional to the severity of COVID-19 in these states.

Our assumption is as follows, the reproduction rate is a factor controlled not only by the disease itself, but also by the cultural differences between countries. In a country where the population is used to big gatherings and large family sizes, the reproduction rate of COVID-19 is higher than countries that are culturally less social and has a smaller family size. Moreover, populations that adhere to government interventions appear to have lower reproduction rates than countries that don't. In that sense, the reproduction rate carries in it the cultural and intervention effects on the spread based on country.

In addition to the aforementioned factors, we also include the healthcare capacity in some of the simulations to show how the urgency of interventions dramatically affects how overwhelmed the healthcare system is forecasted to be affected.

4.3 Cultural Effect Exploration

We share the simulations for the countries and states with and without the social and cultural differences. As mentioned earlier, we considered a few social and cultural aspects of these countries and states and made simulations for different reproduction numbers. We made simulations for different reproduction numbers, representing the the different parameters shown in Table 1.

The simulation figures for India and Saudi Arabia (shown in Figures 16 and 18) show how the social aspects influence the COVID-19 infection curves. Also reducing the reproduction number by conscious social distancing can reduce the infection rates significantly (shown in Figures 17 and 19) . These simulations are heavily influenced by the parameters like the average family size in the country, population, etc. The effect of this is reflected in the reproduction number in these countries. Figures 20 and 21 show the simulations for curves without any social distancing in California and New York respectively.

In US the reproduction numbers are comparatively lower due to multiple factors like the population density, less number of people per family, and better social distancing. The curves for California and New York are reflective of these parameters.

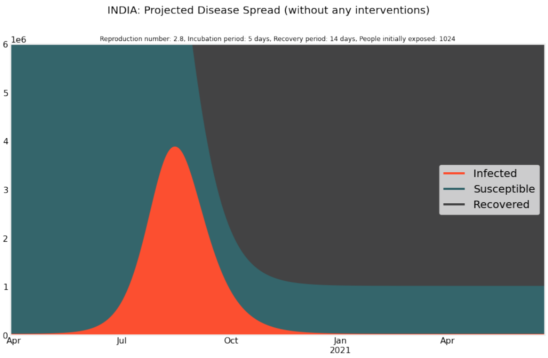


Figure 16: India
Simulation without any social distancing.

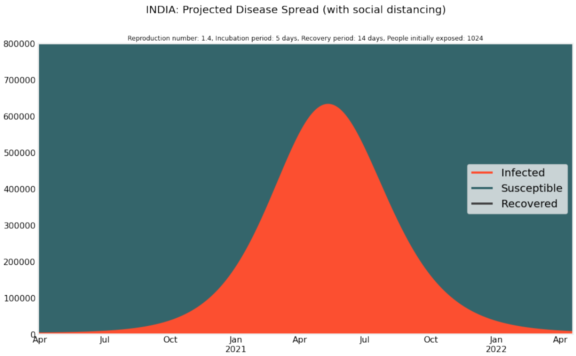


Figure 17: India
Simulation with social distancing.

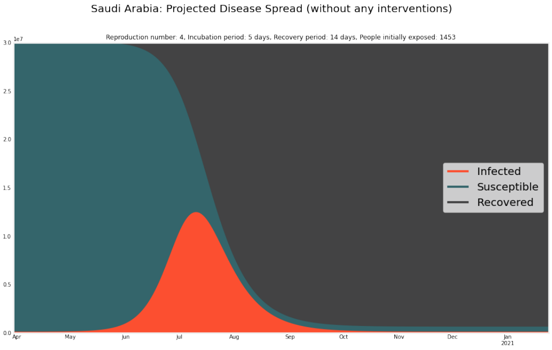


Figure 18: Saudi Arabia
Simulation without any social distancing.

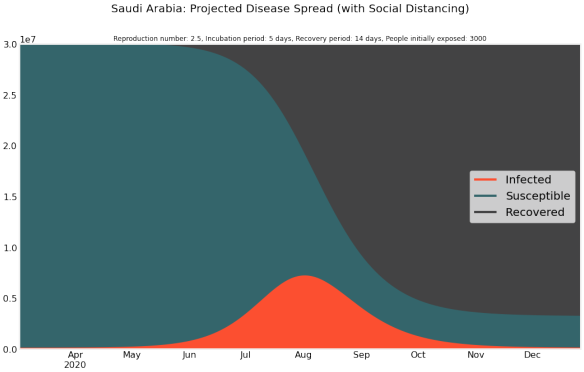


Figure 19: Saudi Arabia
Simulation with social distancing.

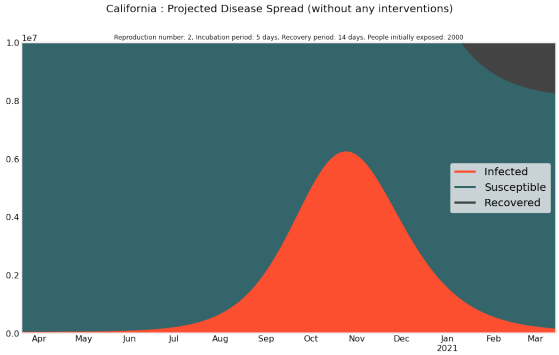


Figure 20: California
Simulation without any social distancing.

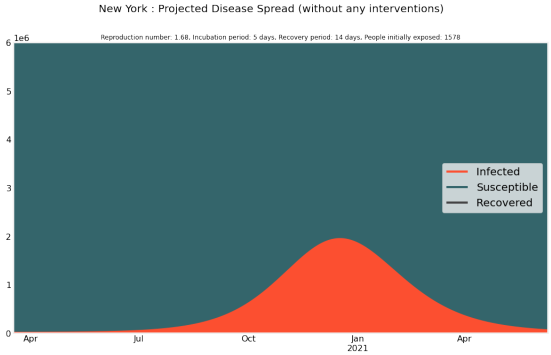


Figure 21: New York
Simulation without any social distancing.

4.4 Governmental Intervention Exploration

In this section we share the simulations with and without the interventions. In India lockdown was implemented after 55 days from the discovery of the first confirmed case. This can be seen in Figure 22. If this intervention was delayed by another 10-30 days then the infection curves would be much higher. This can be seen in the next two Figures 23 and 24. For Saudi Arabia there were two different lockdowns implemented, one partial and then the full lockdown. The first two graphs show curves for these two government Interventions (Figures 25 and 27). We can see here that if the government intervention would have been delayed by another 30 days then the infection spread could have been much worse (Figure 27). The black vertical line in these simulations depicts the time of government intervention and the horizontal blue line represents the health care capacity of the country/state. We can see that if the government intervention was delayed, then the number

Table 1: different factors referenced in the simulations

country/state	family size	Population	healthcare capacity (no. of beds)	government intervention	reproduction rate before interventions	reproduction rate after intervention
Saudi Arabia	6	35M	120K	3/24	3.8	2.5/0.8
India	4.5	1.4B	700K	3/24	2.5	1.29/0.43
California	2.96	39M	75K	3/19	1.8	0.94
New York	2.57	19M	57K	3/22	1.68	0.84
Iowa	2.41	3M	6K	3/15	1.94	0.86

of patients/infected population would have crossed the healthcare capacity for all countries.

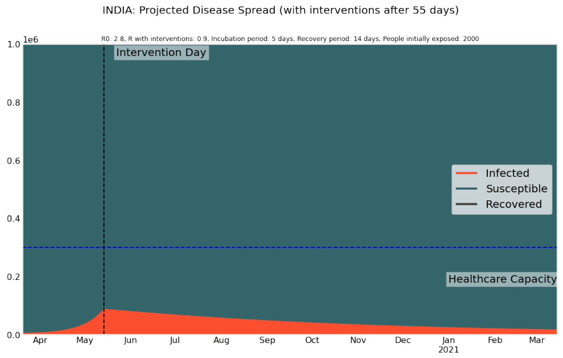


Figure 22: Simulation for India with Government Intervention at 55 days after the first confirmed case.

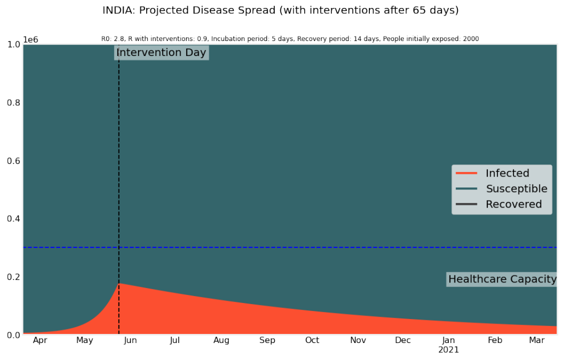


Figure 23: Simulation for India with Government Intervention at 65 days after the first confirmed case.

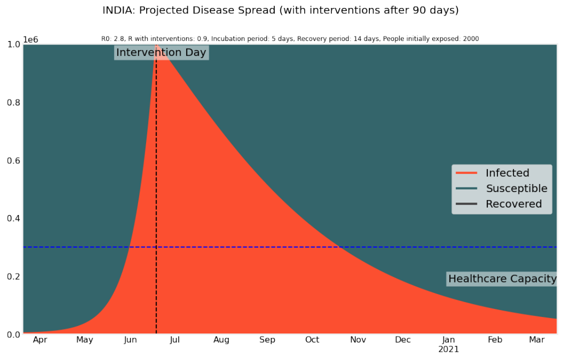


Figure 24: Simulation for India with Government Intervention at 90 days after the first confirmed case.

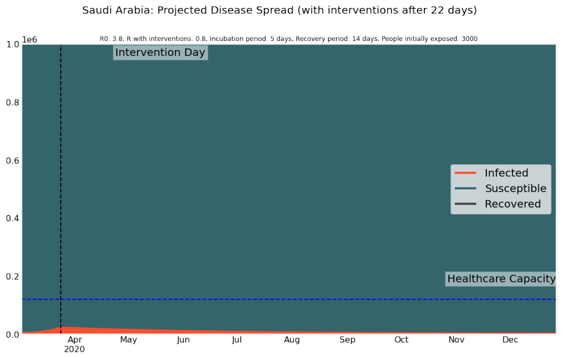


Figure 25: Simulation for Saudi Arabia with Government Intervention at 22 days after the first confirmed case.

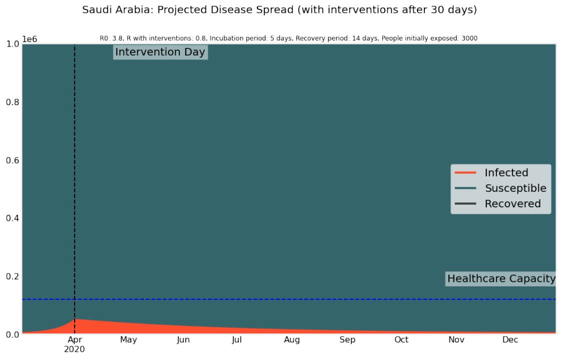


Figure 26: Simulation for Saudi Arabia with Government Intervention at 30 days after the first confirmed case.

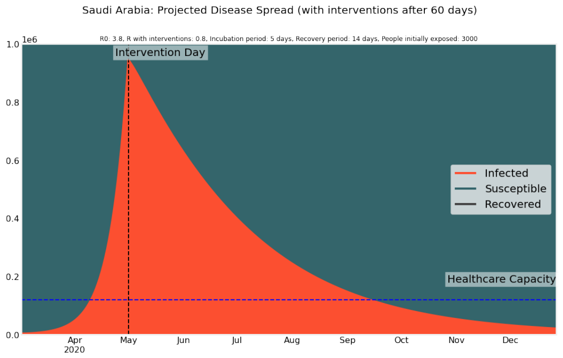


Figure 27: Simulation for Saudi Arabia with Government Intervention at 60 days after the first confirmed case.

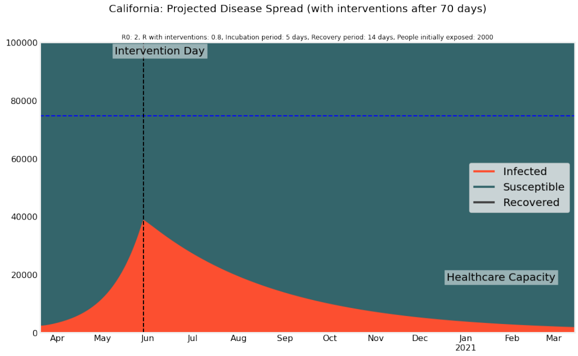


Figure 28: Simulation for California with Government Intervention at 70 days after the first confirmed case.

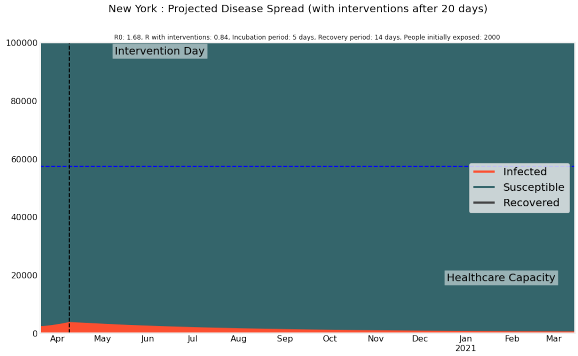


Figure 29: Simulation for NY with Government Intervention at 20 days after the first confirmed case.

5 DISCUSSION

When we examine both the forecast graphs and simulation graphs of the spread for each country/state, we are able to draw conclusions on the effects of government intervention and the importance of social distancing based on the active cases curves.

In Figure 10, the active cases data of Saudi Arabia show a maximum of 30k active cases. The forecasting model predicts the cases to decrease until a full recovery by end of July. The simulation shown in Figure 25 was not exposed to the data used in modeling Figure 10, only to the factors we speculated to have effects on the spread of COVID-19, such as the reproduction number which correlates with social distancing and the government intervention dates (whether partial curfew or a complete lockdown) which in place affects the reproduction number even further. The fact that the simulation peaks at a close value to the real data asserts our assumptions that the main factors shaping a spread curve are the urgency of the interventions and the adherence to social distancing. There are however some discrepancies such as the sudden drop of cases in the simulations graph as opposed to a smoother curve in real data. This is due to the SEIR model not taking into consideration the lag between the intervention date and the correspondence of the population.

In contrast, there is a possibility that the simulations do not match the projected peak of the modeled curve based on the active cases, such as what we see with the simulations for India (Figures 11 and 22). This could be due to numerous reasons based on the observations we make. When the simulations show lower predictions based on the given parameters, it could mean that the population did not adhere to the interventions correlated with these parameters and thus resulted in much higher peaks in real life. This reflects in the projection models forecasting a higher infection rate than the simulated cases.

On the other hand, if the simulations present higher predictions than what we observe in real life, it could signal a mismatch in data collection and inaccurate or inconsistent cases. This can occur when not all cases are recorded, which is often the case. In other cases, the discrepancy between the projections and simulations could simply mean that the projection model was not accurate. Particularly, it is very hard to predict when the spread curve will peak and bend for countries/states that did not reach the peak number of active cases yet.

To summarize, we deduce these explanations to justify differences between simulations predictions:

- The SEIR model deficiencies: it considers the population to be constant/not considering death cases. It also takes constant values for reproduction number as input, not continuous.
- The lag between declaring intervention dates and the actual population compliance which delays the actual results of lockdown by the incubation period (14 days).
- People not adhering to social distancing measures results in different projections vs simulations.
- The fact that the reproduction rate is different in the simulations vs. the projections. This is because the RN is embedded in the temporal data used to generate the forecasting models, where as in simulations a single constant is fed to the model.
- Healthcare capacity is also dependent on multiple parameters and it also increased significantly in some countries like India during the pandemic due to impromptu beds specifically staffed to combat the pandemic.

6 CONCLUSION

The goal of this project is to closely study the cultural and governmental intervention factors and their effects on the spread of COVID-19. In order to do that, we started with exploratory analysis of multiple countries and states. Next, we implemented multiple curve fitting models such as logistic and Gaussian curves. The model choice depends on the data in question and how it behaves. Subsequently, we collected more data about those countries/states in order to produce simulations of the active cases in each of them. Based on these graphs, we found that the parameters we collected and decided to use for simulation generation had great effects on the spread of COVID-19. However, there were other factors that could overshadow the effects of our parameters and we included them in the discussion section.

This work asserts the positive effects we suspected when studying which parameters have to do with the spread of COVID-19, such as government intervention. An interesting aspect that we noticed is that some of those countries and states that started opening up are noticing a second wave of infections. The curve is no longer Gaussian. We are aware of present mass gatherings and the analysis we have performed was before so. A compelling future work extension is studying how reopening affected the spread.

7 CONTRIBUTION

- Samaa Gazzaz: Mainly worked on the simulations and collecting cultural and intervention parameters for the countries and states we worked on. Worked on final discussion and conclusions.
- Niharika Srivastav: Collected data for the social and cultural parameters for India and California. Did simulation analysis for the same countries.
- Katelyn Suhr: Collected data for India and Saudi Arabia. Performed data exploration, logistic and Gaussian curve fitting/forecasts, and experimented with Prophet model forecasts.
- Priya Padmanaban: Collected data for the US, California, New York, and Iowa. Did initial exploratory analysis on the data and logistic and Gaussian curve fitting/forecasts.

REFERENCES

[1] Abdulrahman Alfozan. 2020. COVID-19-notebook. https://github.com/alfozan/COVID-19-notebook

[2] Lakshmikant Dwivedi Balram Rai, Anandi Shukla. 2020. COVID 19 Analysis using RN and Health Care Capacity. https://www.medrxiv.org/content/10.1101/2020.04.09.20059261v1.full.pdf

[3] Ankit Choudhary. 2018. Generate Quick and Accurate Time Series Forecasts using Facebook's Prophet (with Python R codes). https://www.analyticsvidhya.com/blog/2018/05/generate-accurate-forecasts-facebook-prophet-python-r/

[4] Edureka! 2020. COVID - 19 Outbreak Prediction using Machine Learning | Machine Learning Training | Edureka. https://www.youtube.com/watch?v=_Hi6_JQesSQ

[5] Times of India. 2020. Reproduction Number values for India. https://timesofindia.indiatimes.com/india/coronabytes/msid-75708815,card-75710140.cms

[6] Mauro Di Pietro. 2020. Time Series Forecasting with Parametric Curve Fitting. https://medium.com/analytics-vidhya/how-to-predict-when-the-covid-19-pandemic-will-stop-in-your-country-with-python-d6fbb2425a9f

[7] Rajesh Ranjan. 2020. COVID 19 Analysis using Compartmental Epidemiological Models. https://www.medrxiv.org/content/10.1101/2020.04.02.20051466v1.full.pdf

[8] Mehdi Saeedi, Morteza Saheb Zamani, and Mehdi Sedighi. 2010. A library-based synthesis methodology for reversible logic. *Microelectron. J.* 41, 4 (April 2010), 185–194.

[9] Harry Thornburg. 2001. *Introduction to Bayesian Statistics*. Retrieved March 2, 2005 from http://ccrma.stanford.edu/~jos/bayes/bayes.html

[10] Twitter. 2020. Twitter COVID 19 Statistics for Saudi Arabia. https://twitter.com/SaudiCDC/status/1243521250141577217?s=20

[11] John Hopkins University. 2020. Temporal Data from JHU DataSet for COVID 19. https://github.com/CSSEGISandData/COVID-19

[12] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake Vand erPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1. 0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17 (2020), 261–272. https://doi.org/10.1038/s41592-019-0686-2