Visualization and the Trend of Confirmed COVID-19 Cases and Deaths in the U.S

Ziqing Wang[1], Yuetong Ling[2], Jinming Nian[3], Chen Wang[4], Chenxi Qin[5]

[1]York University, Toronto, Ontario, M3J 1P3, Canada

[2]University of Toronto, Toronto, Ontario, M5S 1A1, Canada

[3]University of California, Davis, California, 95616, United States

[4]Central Catholic High School, Lawrence, Massachusetts, 01841, United States

[5]Bashu Secondary School-John Carroll Program, Chongqing, 400013, China

Abstract

This research project aims to visualize and examine the effects and progress of the ongoing COVID-19. Using the programming language R and its corresponding integrated development environment R studio, the report visualizes COVID-19 data by demonstrating time-series plots of the daily increase in confirmed cases and deaths, examines effects of policy implementations, and makes short-term predictions on the potential trend of the pandemic. In this report, the visualized plots indicated that there are different degrees of fluctuations in the number of cases and deaths for each state, but they all possessed an increasing pattern towards the end of each plot, implying a worsening pandemic since early November. The analysis on policy implementations suggested different degrees of policy influence, and prediction results showed a short-term continuously increasing trend in daily new cases and deaths. As the pandemic is a world-wide concern, the significance of this research is that the normally intangible effects of the pandemic can now be visualized and examined, it provides visible effects of policies on the pandemic which aims to augments public awareness, and it gives a short-term prediction which the public can make reference.

Keywords

COVID-19;Visualization; Confirmed cases and deaths; Short-term prediction; Effects of policies.

1. Introduction:

There is no doubt that the most unprecedented event in 2020 that has shocked all mankind is the COVID-19 which persisted for nearly a year. For the first time, people felt a serious threat to the entire human society from a seemingly small influenza virus to which they have not paid much attention before. Due to its fast spreading speed, the global epidemic situation is getting more and more dangerous, especially in the United States. To different degrees, the pandemic has affected each person, each industry, and each society in various ways. Some facts and statistics that have been reveal and investigated about the pandemic include but not limited to the following: In terms of economics, There has been a slump in the US GDP. According to npr.org, the US GDP shrank at an annual rate of 32.9% in the second quarter of 2020, which has become by far the worst contraction [1]. Another consequence was a drastic increase in unemployment rate. According to usnews.com, US unemployment rate rose to its peak since the Great Depression a decade ago [2]. Forbes.com reveals that the hardest hit US state by unemployment was Nevada, with the unemployment rate increasing from 4.0% in May 2019 to 25.3% in May 2020 [3]. By contrast, the least affected state by unemployment was Nebraska, with unemployment rate increasing from 3.1% to 5.2% [3]. But Nebraska's seemingly small change in unemployment rate still reflects the large impact of the coronavirus when the ratio of these numbers are considered. Similarly, unemployment rate has hit many industries. The most severely affected industry was leisure and hospitality due to mobility control measures. The unemployment rate of the leisure and hospitality industry increased from 5.0% in May 2019 to an unprecedented 35.9% in May 2020 [3]. The least affected industry was the Financial Services industry, with the unemployment rate increasing from 1.7% to 5.7% [3]. Despite this small numerical change, the rate more than tripled for the financial industry, which is an indicator that no industry remained safe from unemployment. In terms of the Financial Market, the US stock market crashed in March 2020, which led investors to unexpected losses. According to Forbes.com, the daily returns of the S&P 500 index reached a bottom on March 16, 2020, at a staggering -12.0%, which became historically the third worst performance [4]. Despite a rising trend of stock market performance after the stock market crash as government establish policies to stimulate the economy, shutdowns and other measures to contain this pandemic are still influencing the stock market. Therefore the volatility remained high as the pandemic persisted. In terms of the society in general, many small businesses went bankrupt, restaurants were shut down, and even many educational institutions reached the end of their lives due to low numbers of enrollment. The above are all events that exemplify the impact of the COVID-19, but the impact is by far not limited to those events, and has an impact on each individual, which is why people all over the world care about this pandemic.

This report purports to show the progression, evaluate policy effectiveness, and make reasonable short-term predictions to inform readers about this on-going pandemic, using datasets on COVID-19 cases and deaths found on the New York Times website, as well as policy data found on healthdata.gov. The following sections of this report begins with the data selection and cleaning processes. Then moves on to Explorations and Analysis, in which findings are discussed, inferences are made and analysis are performed. This is then followed by a conclusion section, in which results are summarized, future outlooks are provided, model limitations and potential questions are addressed. The Last section of this report will devote to acknowledgements and references.

2. Data and Analysis

2.1. Data Description

The data we obtained for this project are US state-level data. Data on daily cumulative cases and deaths of each state was found on the New York Times website. The variables used from the New York Times dataset were the date, cases, and deaths, with cases and deaths being integer values. The daily cases and deaths in this dataset were cumulative, therefore the daily number of new cases and deaths can be obtained by subtracting the cumulative cases of each day by the cumulative cases of the previous day, and the plots of cases and deaths against the days spanned since the first day of available data for each state. Hence the independent variables were the number of new cases and the number of new deaths, and the dependent variable was the number of days spanned since the first day of available data. Out of all US states, five states which were more representative and had interesting patterns were selected. Those states are: California, Maryland, Florida and the District of Columbia. The policy dataset was found on healthdata.gov. For the policy dataset, a data cleaning process was performed on the dates of the policies. The pre-cleaning dates were in the format of MM/DD/YY. In order to match the date format of the New York Times dataset, changing the data format to YYYY-MM-DD using the gsub function in R was performed. Then three policies were selected for visualization and analysis: Face Covering Requirement, Shelter in Place Order, and State of Emergency. The policies were selected based on the rationale that these were the most widely implemented policies, and are likely to have an impact on controlling the pandemic.

2.2. Analysis

After extracting and sorting the data according to their dates, the difference was taken between any two given dates to create a plot with "number of new cases" as the y-axis, and a timeline as the x-axis. This would be the upper plot with the name "cases" after the state name. The lower plot with the name "(State name) deaths" is a plot created with the same methods as the upper one, showing the reported COVID-19 death cases. A trend line (colored in light pink) was fitted through the data points of these two plots using the built-in scatter plot smoothing function "lowess ()" for predictions that will be discussed later in this paper. The plot portraying "new cases" was mainly focused on instead of the plot for deaths. This was because putting COVID-19 out of the equation, death cases are influenced by many other factors, for example: previously existing long-term disease, poor immune system, high age, etc. The death plots are important, however, because the main purpose is to show exactly what happened instead of drawing conclusions from the data. Five states were explicitly picked to present in this paper. They represent two different types of COVID-19 confirmed cases growth trend.

Furthermore, a prediction model was built up based on the last 14 days of the lowess curve. Although the lowess curve starts at abscissa equals 0 (the spring of 2020), only the last 14 days were used to build the model since early data tends to have a negligible impact on the future. On the other hand, redundant data will have an adverse effect on the accuracy of the forecast. Moreover, generally, the incubation period of the COVID-19 is 1~14 days. Given the lowess curve is approximately monotonous at the last 14 days, a quadratic term is introduced into the model whereas a higher-order term will not be considered. The prediction line (colored in deep sky blue) was also be drawn from the last 14 days to the upcoming 14 days. Besides, from the prediction function of the model, the future trends in 14 days could be predicted.

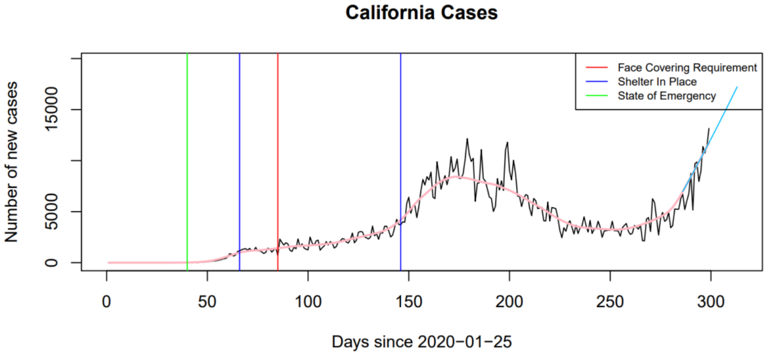

Figure 1. Time series plots of daily new cases in California.

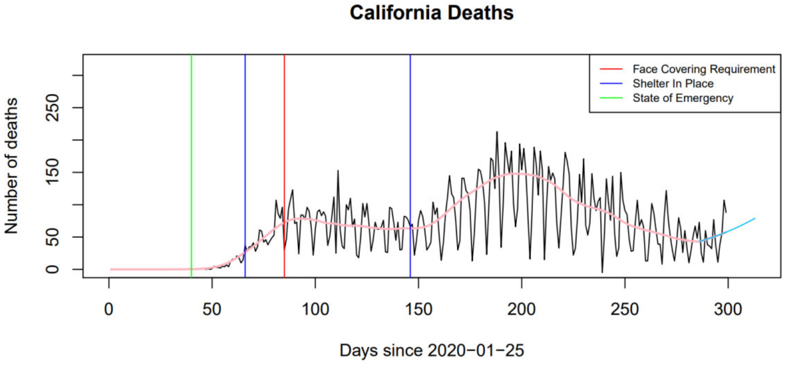

Figure 2. Time series plots of daily deaths in California.

Prediction function of cases:

$$cases = -22770 - 147.9\,day + 0.8810\,day^2$$

Prediction function of deaths:

$$deaths = 1092.04653 - 8.22906\,day + 0.01595\,day^2$$

The shape of "California cases" can be categorized as the first type, where the state is heavily populated, and there are numerous large cities in them. As seen on the plot, cases started to sky-rock around the middle of May. This was about two months after the quarantine in California. Around this time, people started to go out more, while the importance of wearing masks was still widely discussed. A peak of confirmed cases was reached around the end of July, where new confirmed cases were as high as 12000 per day. After that, new cases went down and stabled at around 4000 per day. Then it began to diverge and showed a scary steep curve starting from the middle of October. Looking at the lower plot, the "deaths" plot, a similar shape to the cases plot showed up, but shifted to the future for about 7-14 days.

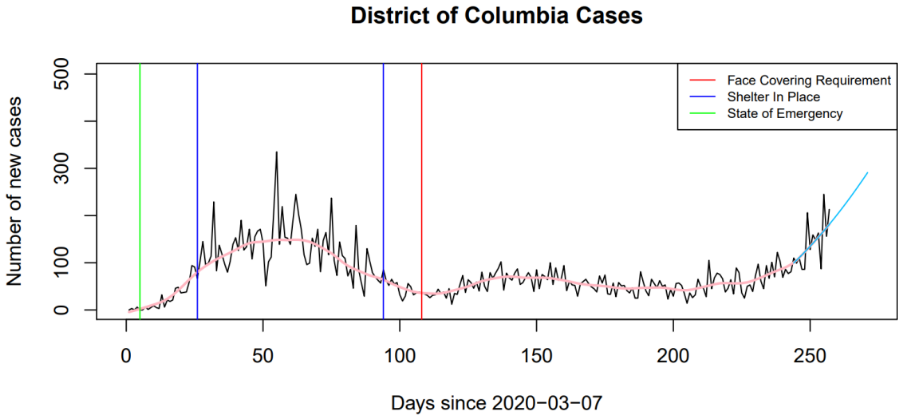

Figure 3. Time series plots of daily new cases in District of Columbia

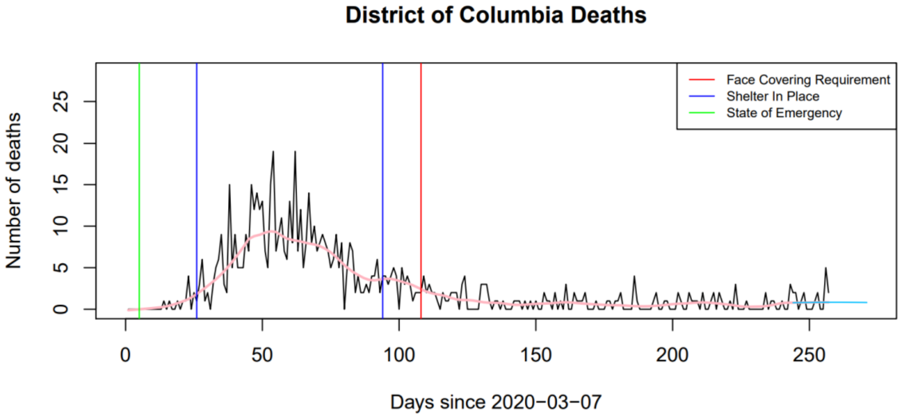

Figure 4. Time series plots of daily deaths in District of Columbia

Prediction function of cases:

$$cases = 2199 - 22.74\,day + 0.05793\,day^2$$

Prediction function of deaths:

$$deaths = -12.80 + 0.1049\,day - 0.0002019\,day^2$$

The shape of "District of Columbia cases" can be categorized as the second type. These two plots for D.C. almost look identical to the plots for Maryland. The overall confirmed cases in D.C. is about one-tenth of Maryland. This is especially interesting because if the plot for Maryland is used and only reduce the y-axis by a factor of 10, an almost identical "cases" plot to D.C.'s plot will be obtained. This could be caused by the geographical closeness of Maryland and D.C., but most importantly, it shows consistency between the shape of these plots and the state's geographical location. In other words, the trend in confirmed cases has a dependence on the region.

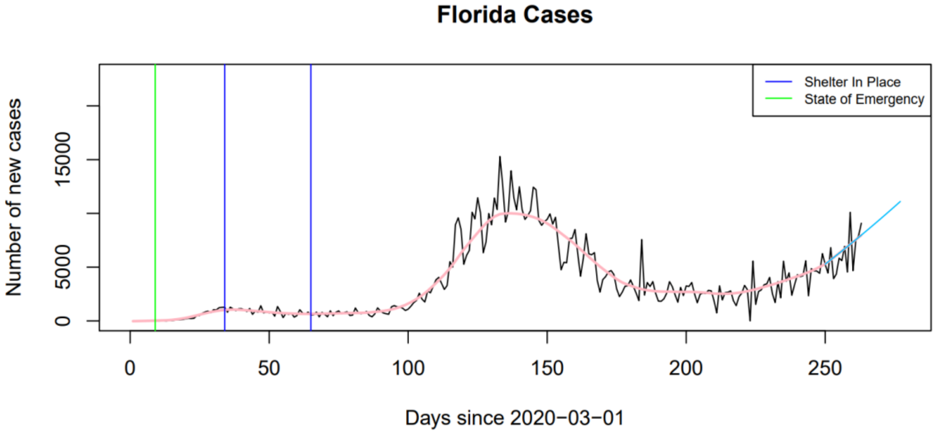

Figure 5. Time series plots of daily new cases in Florida

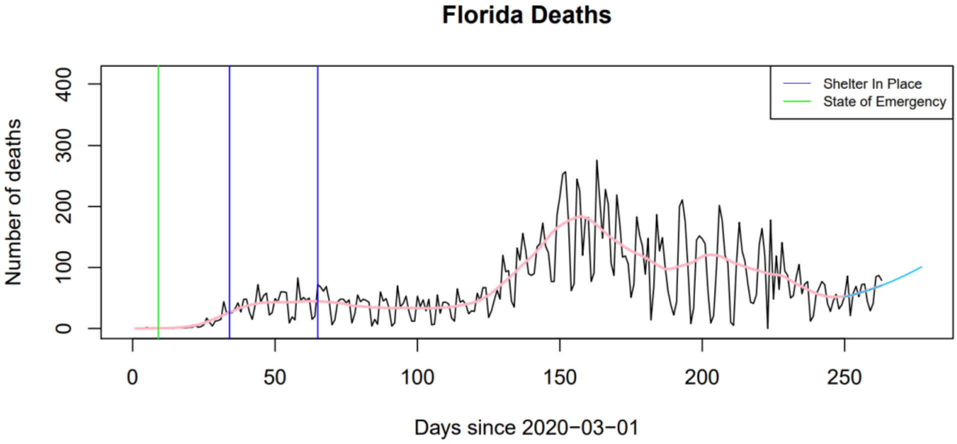

Figure 6. Time series plots of daily deaths in Florida

Prediction function of cases:

$$cases = -9160.2552 - 84.8044\, day + 0.5702\, day^2$$

Prediction function of deaths:

$$deaths = 1487.0398 - 12.5651\, day + 0.0273\, day^2$$

Similar to the plots for California, this can be categorized as the first type. New confirmed cases in Florida peaked around mid-July, a few weeks after California. This peak came to as high as 15,000 new cases daily. Then the curve gradually went down to a steady 3000 new cases daily since the beginning of September. The sad news is then, similar to many other big states like California, cases started to increase more fiercely from the middle of October. There was an apparent 7-14 days shift to the future in the deaths plot compared to the cases plot.

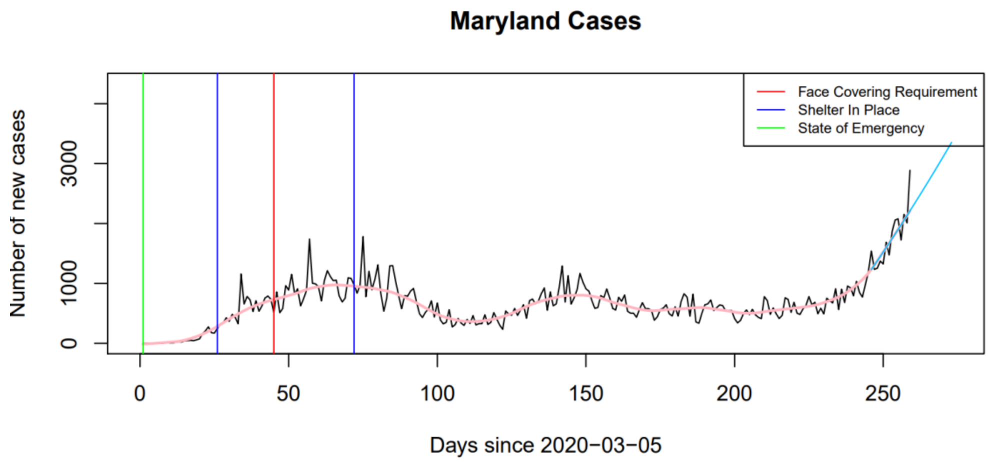

Figure 7. Time series plots of daily new cases in Maryland

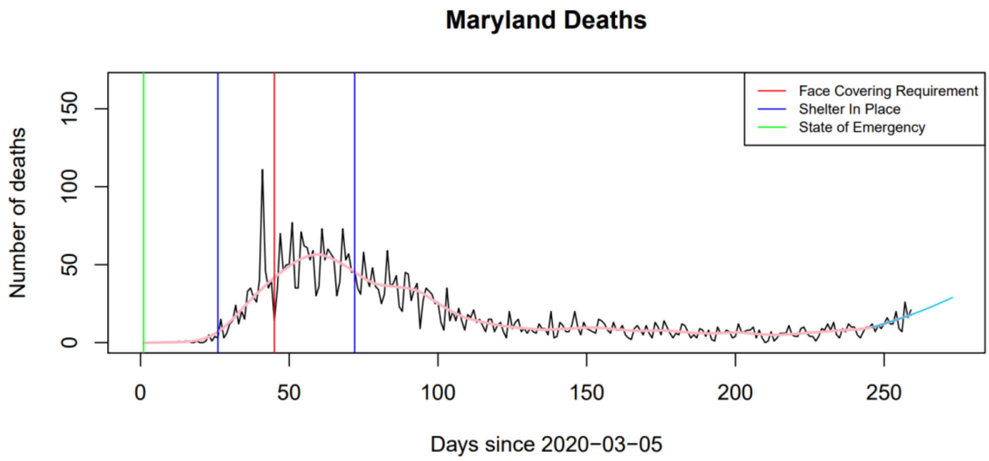

Figure 8. Time series plots of daily deaths in Maryland

Prediction function of cases:

$$cases = -2468 - 42.21\, day + 0.2327\, day^2$$

Prediction function of deaths:

$$deaths = 264.543297 - 2.607357\, day + 0.006390\, day^2$$

The shape of “Maryland cases” can be categorized as the second type, where the population is less than 10 million, and the peak does not exceed 5000 cases per day. In the upper plot, cases peaked at around the middle of May. Then, there are minor fluctuations from the end of May to the middle of October at around 500 new cases each day. However, after October, cases started to grow almost exponentially. The good news is that the death cases became stable around mid-June, and it only started to rise a little around the tail of the plot which is the beginning of November.

The colored vertical lines in the plots shown above represent the dates on which a specific policy was issued. Some policies ended before the last day of our data set, therefore, there is often a second blue line representing the end date of “shelter in place” policy. The correlation between policies issued by the state governments and the trend in the “cases” plot was investigated. The policies issued are for slowing down the spread of COVID-19. Therefore, any flattening of the “cases” curve was that could be possibly due to a certain policy was looked for.

Face Covering Requirement: There are no legal consequences for not wearing a mask, and it would be impossible to monitor or execute it if there were some legal consequences. This policy serves mainly as a suggestion for people to voluntarily follow. As a result, there was not an apparent effect on flattening the curve. Some states, for instance, Florida, did not even issue this policy.

Shelter in place: This policy was present for as long as three months in some states, and as short as a month in some other states. This policy has a great effect on reducing the slope of the curve for many states, as shown in our plots. For heavily populated states like California, the curve during the quarantine period has a much smaller slope compared to other parts of the plot. In states like D.C. and Virginia, the curve even became negative during the quarantine period. Some states like Florida which only quarantined for about a month can be understood because during the quarantine period, there were scarcely any COVID-19 cases.

State of Emergency: Most states announced a State of Emergency around the beginning of March, and none of the states shown in this paper have withdrawn from the State of Emergency until November 19th.

3. Conclusion

There are many reasons to influence and change the trend of the cases and death plot of each state. Due to the time limit of this report, the predictions of daily new cases and deaths are idealized since we assume that in the short term, there will not be any other interfering factors, such as vaccines. Furthermore, the policy selection has not been dedicated to each individual state. In future research, the mobility data from Google which is explicit, accurate, and closely related to the growing momentum of daily new cases can be incorporated. The mobility data from Google, in a societal point of view, can show or even draw conclusions to what are some of the factors that caused the peaks and steep rises that came after the quarantine period and after mid-October in some of the states. In this report, the data of confirmed cases and deaths of coronavirus in four US states were summarized into two categories: population size and regional differences. California, Florida, Colombia's district and Maryland's data analysis state the number of people and state location can influence will be coronavirus confirmed cases and deaths of the two factors, and the four states' peak of the cases numbers are all confirmed in May 2020 by the end of July; however, after a period of slow growth, number of cases started increasing rapidly in October. The models and curves show the rates of cases and deaths in these four states from roughly the beginning of the epidemic to the present. And the data in the models show that, in the short term, the number of cases in these states has not abated and is still increasing, but the number of deaths is growing slower than before. This shows obviously that coronavirus is a very threatening virus infection, but there is a real possibility of being controlled. People should keep calm and vigilant, and continue to shoulder their social responsibilities. These four states are somehow the representatives of other states in the U.S., and the situation of the other places around the country could be estimated based on the conclusions. Washing hands frequently, limiting social gatherings and travel to high-risk areas for the virus, and cleaning and sanitizing frequently would help people protect themselves and others around you. Governments, health authorities, experts in the field and all people should work together to overcome this challenging outbreak with swift and forceful action and to reduce suffering and separation. After overcoming this epidemic, which will certainly happen someday, the world's ability to maintain stability in the face of similar challenges will be fully assessed.

Acknowledgments

On the very outset of this report, we want to express enough thanks of gratitude to our professor John Emerson for his conscientious guidance and encouragement, and the golden opportunity of writing this report, also Teaching Assistant, Max, who has been given a lot of suggestions and recommendations during the process of writing the report.

References

[1] Scott Horsley. (2020) 3 Months Of Hell: U.S. Economy Drops 32.9% In Worst GDP Report Ever. https://www.npr.org/sections/coronavirus-live-updates/2020/07/30/896714437/3-months-of-hell-u-s-economys-worst-quarter-ever

[2] Andrew Soergel. (2020) Unemployment Highest Since Great Depression as Coronavirus Collapses Labor Market. https://www.usnews.com/news/national-news/articles/2020-05-08/unemployment-highest-since-great-depression-as-coronavirus-collapses-labor-market.

[3] Mike Patton. (2020) Pre And Post Coronavirus Unemployment Rates By State, Industry, Age Group, And Race. https://www.forbes.com/sites/mikepatton/2020/06/28/pre-and-post-coronavirus-unemployment-rates-by-state-industry-age-group-and-race/?sh=3837973c555e.

[4] Julie Jason.(2020) The Coronavirus Stock Market: A Market Gone Wild.https://www.forbes.com/sites/juliejason/2020/04/08/the-coronavirus-stock-market-a-market-gone-wild/?sh=6b248923a31f.