Making Sense of Covid-19

Kathy Wu, Yewen Zhou

Abstract: The coronavirus, also referred to as "Covid-19" has spread out to the entire world and caused thousands of deaths. In this project, we are more interested in the domestic situation of the pandemic, specifically, we are trying to answer questions such as "is the total number of cases still growing?" and "how has social distancing affected the overall trend?". By conducting Exploratory Data Analysis (EDA), we arrived at the result that based on the data from 4/27/20 to 5/3/20, with 95% confidence, the growth rate of confirmed cases in States such as California and New York are decreasing but are still increasing in States such as New Jersey. We also found that States that implemented social distancing measures early on generally were able to "flatten the curve" by reducing the growth rate of the number of confirmed cases. Finally, we reflected on our whole process and brought up interesting questions for future research.

1. Introduction

In this project, we are interested in answering the following questions: 1, is the total number of confirmed cases still growing? If it is, is it growing faster or slower? When would we possibly reach a peak? 2, how effective is social distancing? In order to answer these questions, we utilized the techniques of EDAs, fitted proper models, and made conclusions accordingly.

2. Description of Data

The datasets that we used include "abridged_couties.csv", "time_series_covid19_confirmed_US.csv", "time_series_covid19_deaths_US.csv", and "05-03-2020.csv". Except for "abridged_couties.csv", all other three datasets were updated up to 5/3/20.

3. Summary of Results

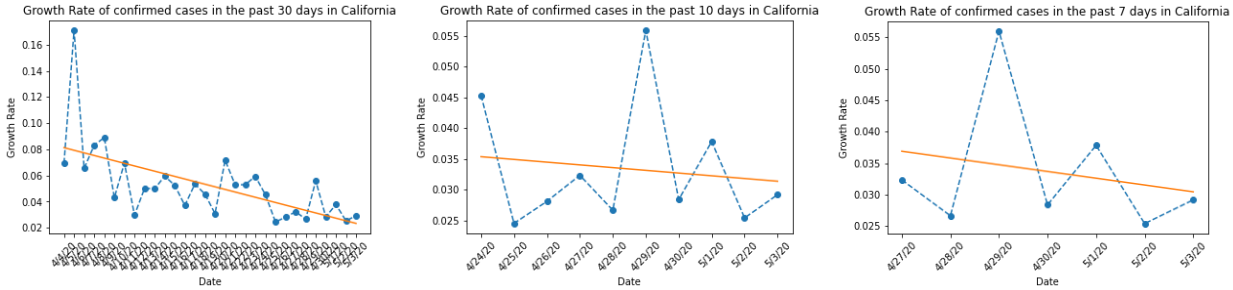
3.1 Are the confirmed cases in California and New York still growing? When would it reach the peak?

Data cleaning and transformations: in the method get_df(state, x), we calculated the sum of all the confirmed cases in a specific State, then we added a column "Days" that represents

the days after 1/22/20 so that it is easier for future analysis, a column of "Growth Rate" that represents the growth rate, a column "log" that is the log transformation of the original data. We did log transformation on the data with the assumption that the original data might have exponential growth and log transformation would transform the data into a linear trend so that it is easier to be dealt with.

Methods: there are four methods that we used here. The first method get_df(state, x) does the data cleaning and data transformation and it returns a DataFrame that includes the confirmed cases in a specific State in the past x days. The second method fit_linear_mode(state, x) fits a linear regression model to the growth rate in the past x days in a specific State, makes plots of the growth rate and the linear regression model, outputs a summary table that shows the statistics of the model, and predicts the date when the State reaches the peak. The third method uses the scikit-learn module and evaluates the model. The fourth method returns a slice of the DataFrame from the start date to the end date inclusively.

Feature selection: we tried to use the "date" as our feature but it was not convenient to work with the datetime object. We also tried to use a multilinear regression model using the growth rate in the past x days as x_train and the growth rate on 5/3/20 as y_train but it did not work because we only had one sample instead of a matrix. We then used "Count Up", which starts with 0 and ends at 5/3/20 as the feature for our linear regression model. We decided to use the data in the past x days with x being a small number such as 7 or 10 because the growth rate is constantly changing so including the most recent data can best reflect the current situation. By comparing the plots and statistics in the past 30, 10, and 7 days, we noticed that with 95% confidence the slope of the regression line in the past 30 days is negative ([-0.003, -0.001]), while it could be positive for both the past 10 days ([-0.003, 0.002]) and the past 7 days ([-0.007, 0.004]), which suggests the possibility that the cases are growing faster again.

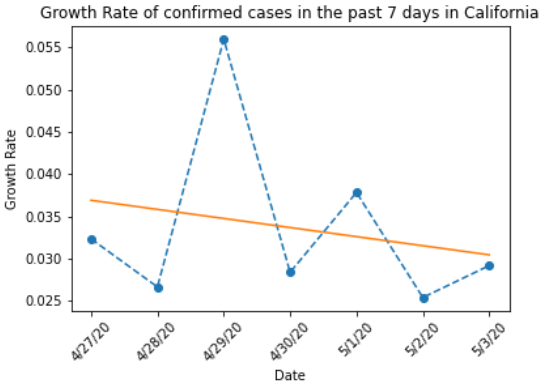


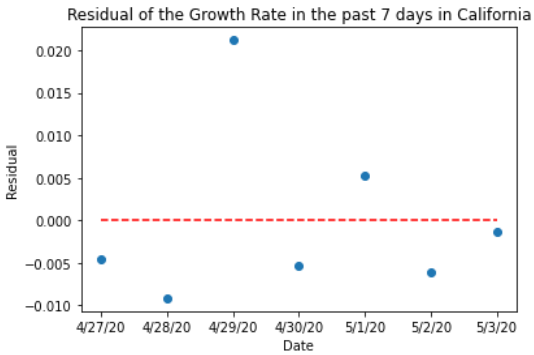
Challenges: one of the challenges was to smoothly handle the datetime object, which was solved by using proper built-in methods. Another challenge was to conveniently calculate the confidence interval of the slope of the regression line, which was solved by introducing the "statsmodel" API which conveniently summarized all relevant statistics.

Model and assumptions: we used a linear regression model to fit the growth rate data in the past x days with the assumption that in the past x days, although the growth rate might be going up and down in a "shaky" way, the overall trend is linear. We also decided to use the linear regression model to make it simple and to only address the question of whether the growth rate is positive or negative and whether the confirmed cases are growing faster or slower or at the same rate as before.

Interpretation of the results:

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0369	0.008	4.751	0.005	0.017	0.057
Count Up	-0.0011	0.002	-0.499	0.639	-0.007	0.004





In the analysis of the State of California, after fitting a linear regression model to the growth rate from 4/27/20 to 5/3/20, it is shown that the slope of the line is negative, which means that based on the data in these 7 days, the growth rate of total confirmed cases is decreasing, the total number of confirmed cases are still increasing but at a lower rate, and the predicted date when the confirmed cases reach a peak is on 5/31/20. From the residual plot, we see that there is not a trend, and the variance of residuals decreases with respect to date, which means that our model is reasonable but not perfect. However, there is 95% confidence that the slope of the

regression line is 0, which means that it is possible that the total cases are still increasing at the same rate.

Evaluations: using scikit-learn, we fit a linear regression model and performed 3-fold cross-validation on our data, the scores were [1, 1, 1] which are the best possible scores.

Limitations: here we adopted a linear regression model, which simply ignores the fluctuation of the data and simplified the data to be a linear trend, it is limited in predicting the possible cycles of the data.

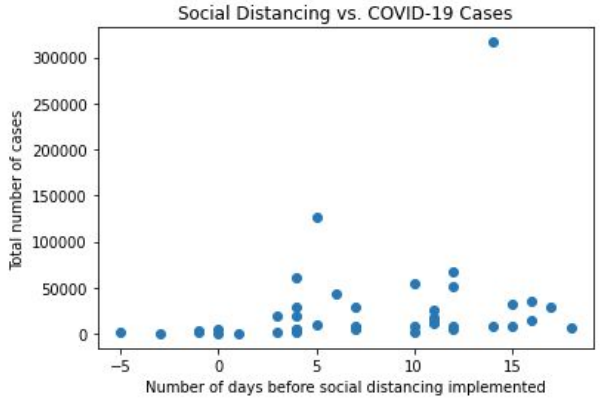
3.2 How have social distancing measures affected the confirmed rate and mortality rate?

Data cleaning and transformations: For this question, we started by using the stay_at_home column of the abridged_counties dataset to sort States by the average date that they began social distancing measures. We grouped counties by State and took the mean of the column, dropping values that were n/a. We then merged the average dates with the confirmed cases dataframe, using States as the index. To normalize the number of cases based on population, we added a population column that was the sum of individual county populations. Finally, we converted the stay at home dates to month/day/year format for easier understanding.

Methods: For the first method, we simply explored the confirmed case rate of States that started social distancing earliest versus the States that started social distancing latest. We then investigated the time it took between the first case and when each State implemented social distancing. We made a scatter plot comparing the time of implementation of the social distance policy with the total confirmed number of cases on 5/3/20. We then looked at individual States such as California and New York to see if the confirmed case growth rate has slowed since shelter-in-place policies were put in place.

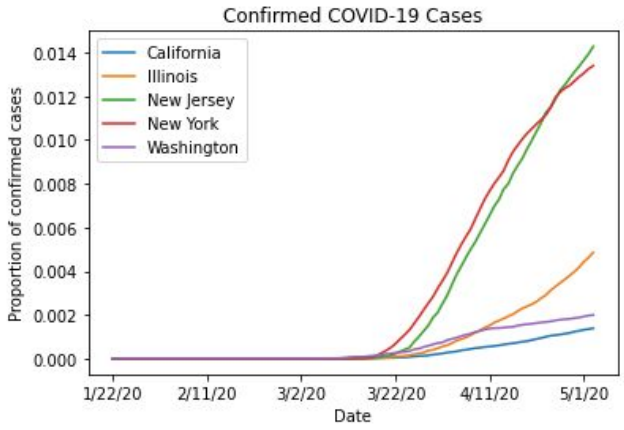
Challenges/Limitations: It was hard to accurately identify how long it took for each State to implement social distancing. Because different counties implemented shelter-in-place at different times, the average date of shelter in place for each State may not have been as accurate as hoped. The average number of cases would also have been better estimated at the county level if we had access to that data.

Results:



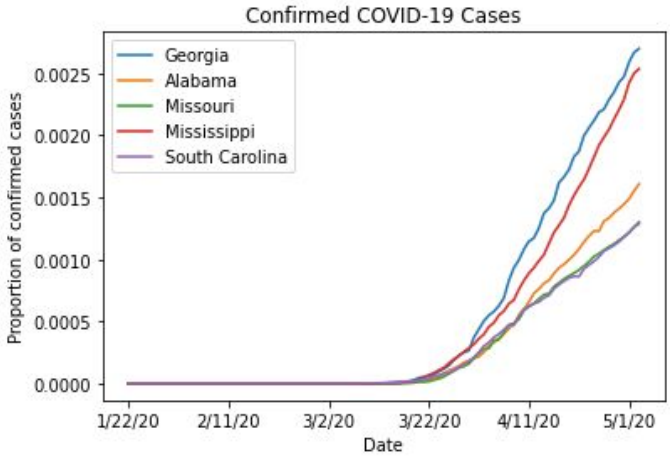
In the above scatter plot, we plotted the number of days before shelter-in-place against the total number of cases for the State at present. We were hoping for a positive correlation between the two variables; i.e. States that took longer to implement shelter-in-place should theoretically have a higher number of cases. However, there turns out to be little correlation—most States have around the same number of cases, with more outliers near the middle. In hindsight, this is expected because many States implemented social distancing around the same time. In addition, we averaged by State (due to data limitations) instead of considering the shelter in place implementations at the county level. Shelter in place can also be limited in success if factors such as hospital capacity, socioeconomic status, etc are not ideal.

In the next two plots, we compared the confirmed Covid-19 rate of States with earlier versus later shelter-in-place dates to see if they were more successful in "flattening the curve".



Above is the graph produced for the states with the earliest social distancing policies. In general, we still saw high growth in the number of cases for New York and New Jersey, but

California, Washington, and Illinois were able to reduce their growth rate by a significant amount by implementing shelter in place early on, around 3/19 to 3/22.



Above are the States with the latest shelter-in-place policies (again factoring in when each State saw its first case).

We see that generally all States had high growth rates, particularly Georgia and Mississippi. Interestingly, all States with the latest social distancing policies are geographically in the South.

4. Ethical Concerns and Possible Solutions

Working with Covid-19 data presents ethical concerns regarding data collection and human bias. For example, we want to collect patient data (underlying conditions, length of illness, mortality) without violating patient confidentiality rights. Data should be anonymized so that it is not possible for the person to be identified, which in this case means that it may not be possible to collect data on a patient's precise location despite the fact that it could produce more accurate predictions. Another ethical concern involves human bias when analyzing the data. We should be careful not to introduce bias when predicting confirmed case outcomes; for instance, we should not go in assuming that the cases will peak at a certain time because of prior knowledge or instinct. Finally, Covid-19 data disproportionately affects certain minority populations and socioeconomic classes. While these may be good factors to use when predicting coronavirus outcomes, we should make sure that we are not targeting populations in a detrimental way.

5. Additional Resources

It would be helpful if we had additional data that includes the daily number of tests so that instead of using the daily confirmed cases alone, we would have a rough idea of how many people are still under the tests. It would also be helpful if we had access to data that has a higher level of granularity, such as the data on the individual level that could give us an idea of the individual's age, gender, etc, so that we could test other hypotheses in greater detail. Knowing the case rate on a county level would also be useful, since much of the other data (social distancing, hospital statistics, etc) were given at the county level as well.

6. Discoveries and Future Work

In this report, we first set off trying to answer the question of "are the cases still growing in California, and how fast it is growing?" which led us to the discovery that the cases are still growing but at a slower rate. However, for States such as New Jersey, we found that based on the data from 4/27/20 to 5/3/20, the growth rate of the confirmed cases are increasing, which suggests that we expect there to be more growth in confirmed cases in the near future. We also found that stay-at-home measures generally reduced the growth rate of confirmed cases, although there were many confounding factors that may have been at play— namely the policy differences between specific counties, as well as healthcare and socioeconomic differences between States. Although we have tried different approaches to model the growth rate, we will leave the question open to future analysis to find a better and more accurate model.