Business Problem

Since the beginning of the spread of Covid-19 governments have tried different ways to attempt to curb the effects of the virus. States and counties have implemented mandates to varying degrees. What effect have these mandates had on the fallout of the virus? This study attempts to measure the effect certain regulations have on the number of deaths from Covid-19. Does the regulation reduce the number of deaths? Can accurate models to predict deaths be created to model outcome with and without the regulations?

Background/History

The Center for Disease Control (CDC) has been collecting Covid-19 data and the data available for each county. The data is available both as latest numbers and a day-by-day report. Using the daily data, a time series data frame can be created to follow trends through the spread of the virus.

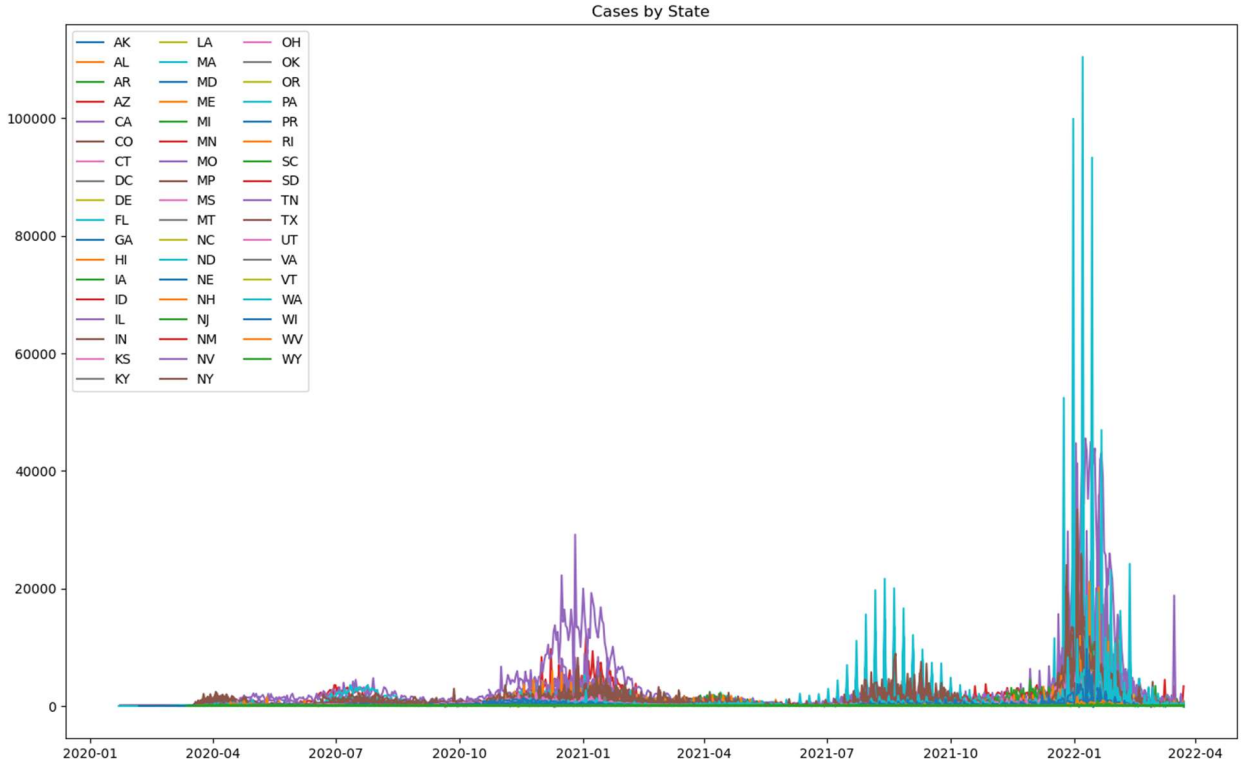
Information regarding Covid-19 regulations has been made available through the Covid Act Now API. The data is available as a collection of data frames with information regarding regulations for each county for the given dates. Using these data sets together it is possible to obtain the death rate for each county along with information regarding the regulations.

By analyzing the data and comparing the results for different counties it should be possible to measure the effectiveness of different regulations. Using modeling techniques will make it possible to see how implementing or choosing not to enact mandates would have affected the spread of Covid-19.

Data Explanation

As mentioned above the CDC data contains information for each county. The data is available as a JSON file, with each county containing information for a county. When unpacking the information, the data was divided into two separate data frames. One contains all the latest data available as static data, the second was expanded to a time series panel data frame. The second data frame was then merged with

the data from Covid Act Now to create one data frame.



Many data points are missing as not all dates are available from all data sources. Missing data will be dealt with at time of use of data as different models appear to deal with missing data differently. If necessary missing data will be dropped. Otherwise, certain assumptions will be made to fill the data.

The target variable for this study will be daily new cases as a percentage of the county's population. This has been chosen despite that there may be indications of change in reporting. It is known that testing at the start of the spread of Covid-19 was not as widely available as later. This may skew the number of cases reported and make it appear that later dates have higher instances of Covid-19 as compared to earlier. The percentage of cases to population has been chosen over the raw new case numbers to counterbalance the wide variance of population size between counties.

Methods

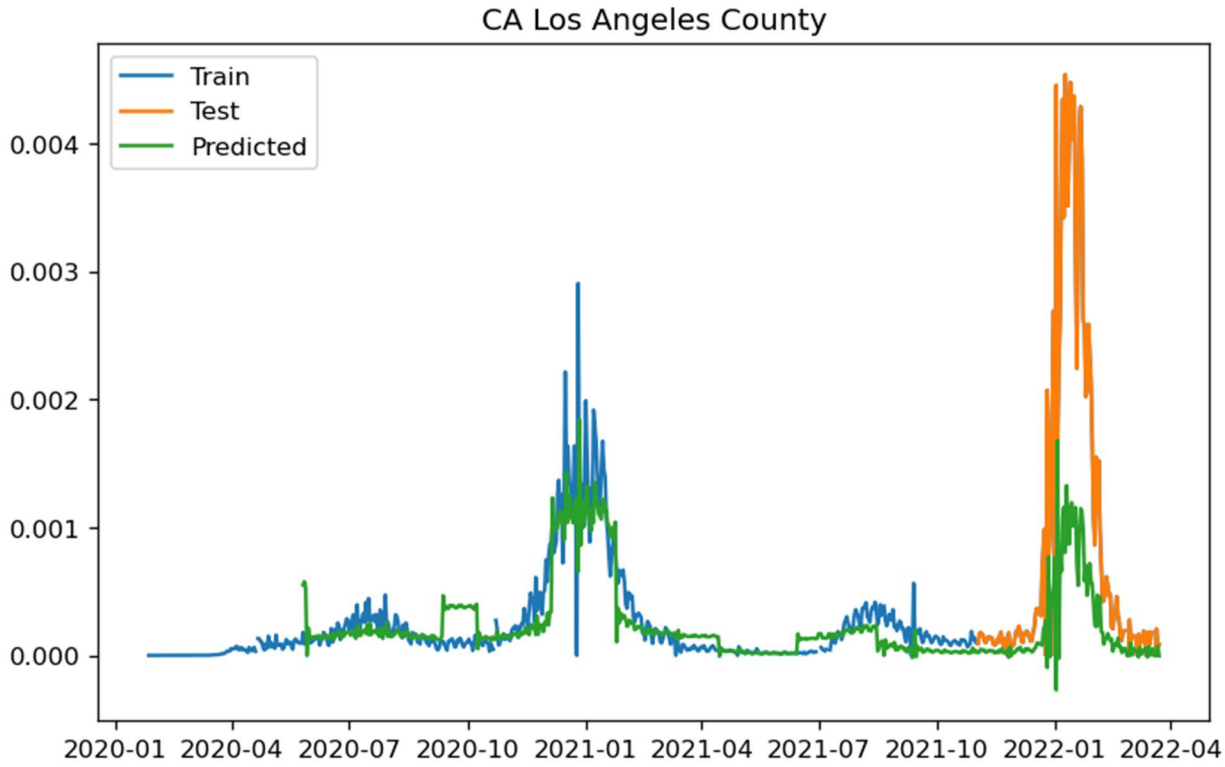
Exploratory data analysis was preformed on all parts of the data to check for outliers and other anomalies. When possible missing data was filled from information gathered from other variables. This was possible for order codes and their meanings as these were one-to-one matches and was possible to fill one from the other.

For time series analysis the categorical orders were used, and one hot encoding implemented. Statsmodels' SARIMAX model used to predict trends. This was applied to each county individually and an r2 score returned. While the model does not appear to work well for all counties and has an over all negative r2 score. The model was run using regulations as a predictor as well as without. The mean r2 score for the two models shows little difference. Adding regulations to the model has a mean r2 score of -0.115. Without the regulations as a predictor the average r2 score is -0.113. To be able to illustrate the model's performance Los Angeles County was chosen.

Further exploratory analysis was preformed on the static data. Correlation and patterns were analyzed to check for patterns within the data.

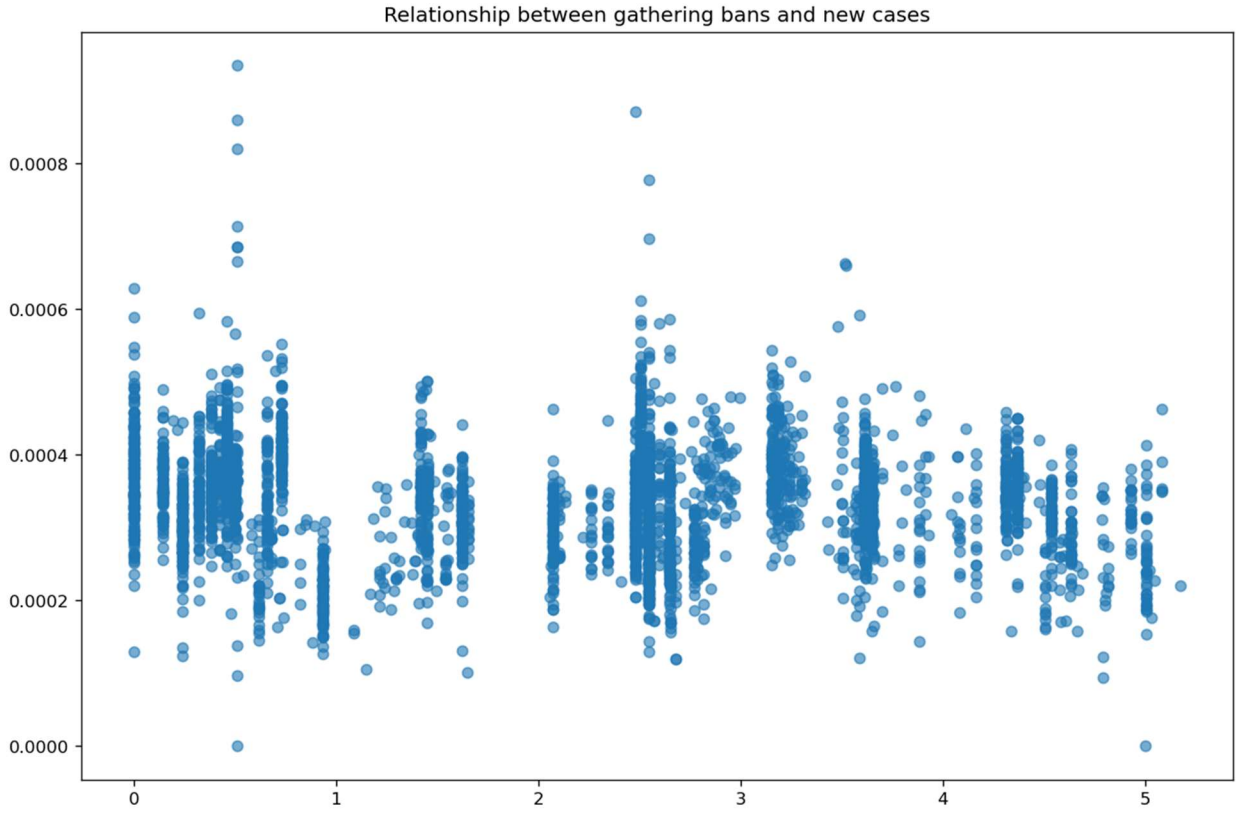
Analysis

Based on the findings of the model and the low predictability it is hard to determine if regulations have any significant effect on the virus. While it is true that r2 for the models are low, the model for Los Angeles which preform better than the average with an r2 score of just 0.103, still sems to be able to spot trends in the data. More analysis would be needed to determine if better parameters would improve accuracy of model. In the Los Angeles County model there is a 0.024 improvement by adding the regulation variables.



Using regulation codes ordered, with most stringent regulations given higher numbers, an average level was determined based on the duration of regulation levels. When using these numbers there is a slight negative correlation between all the regulations and both total cases and death ratios. The largest correlation is -0.15. However, upon examining scatter plots there is concern that the relationship may

not be linear.



Conclusion

Looking at the information collected above, it can tentatively be suggested that there is a small benefit to be had by implementing restrictions. Based on the models used there is no real option for trying to predict outcomes for dropping or implementing restrictions. The model accuracy is too low for any significant assumptions to be drawn for such predictions.

Assumptions

A tentative suggestion was made that there appears to be some benefit in lowering virus cases by implementing restrictions. These suggestions were made based on the correlation between stricter regulations and a lower ratio of cases. It is important to note that this assumption does not take into account when the orders were implemented or when the cases occurred. The rule that “correlation does not imply causation” is something to keep in mind. Furthermore, all regulations are treated as equal and the different level numbers are not necessarily equally distributed to draw any conclusive insight.

Limitations

The results of this study are based on the data available. As discussed earlier there is some concern as to the reliability of the numbers provided. It is known that testing has increased as time went on this may have led to lower initial numbers. It would be interesting to see if those with higher numbers earlier have had lower numbers later as a result of natural immunity. This may lead to an appearance of correlation between regulations that those counties implemented as a result of high numbers.

Challenges

Dealing with time series analysis comes along with its own set of rules. Trying to apply the analysis to multiple counties has not been successful. The added dimension of time requires further study and research in order to find the right model.

Future Uses/Additional Applications

If an accurate model with definitive results can be obtained, similar studies can be applied to an analysis of economic impact of said regulations. While it may be true that there are benefits to the regulations it should be weighed against economic impact.

Recommendations

As of now the only recommendation that can be made is that further study is warranted. There appears there is some benefit to be reached from mandating regulations. When and how to implement them needs more study. A deep learning model that can assess all counties and the effect of the mandates would be recommended as the next avenue of exploration.

Ethical Assessment

This study has been done in a way to avoid any areas of ethical concern. The question has been limited to addressing whether regulations are beneficial or not. No attempt has been made to differentiate between counties at all. Other than what one may decide to assume based on county name no other variables have been included that may lead to any discrimination.

References

CDC, COVID-19 Community Intervention and At-Risk Task Force, Monitoring and Evaluation Team & CDC, Center for State, Tribal, Local, and Territorial Support, Public Health Law Program, “State and Territorial COVID-19 Orders and Proclamations for Individuals to Stay Home,” (August 15, 2021).

CDC, COVID-19 Community Intervention and Critical Populations Task Force, Monitoring and Evaluation Team & CDC, Center for State, Tribal, Local, and Territorial Support, Public Health Law Program, “State and Territorial COVID-19 Orders and Proclamations Banning Gatherings,” (August 15, 2021).

CDC, COVID-19 Community Intervention & Critical Populations Task Force, Monitoring & Evaluation Team, Mitigation Policy Analysis Unit and the CDC, Center for State, Tribal, Local, and Territorial Support, Public Health Law Program, “State and Territorial COVID-19 Orders and Proclamations Closing and Reopening Restaurants” (August 15, 2021).

CDC, COVID-19 Community Intervention & Critical Populations Task Force, Monitoring & Evaluation Team, Mitigation Policy Analysis Unit, the CDC, Center for State, Tribal, Local, and Territorial Support, Public Health Law Program, and Max Gakh, Assistant Professor, School of Public Health, University of Nevada, Las Vegas, “U.S. State and Territorial Orders Requiring Masks in Public,” (August 15, 2021).

Jason Brownlee (August 6, 2018), 11 classical time series forecasting methods in python (cheat sheet). Retrieved from https://machinelearningmastery.com/time-series-forecasting-methods-in-python-cheat-sheet/

Keita Miyaki (August 15, 2019), Time series split with scikit-learn. Retrieved from https://medium.com/keita-starts-data-science/time-series-split-with-scikit-learn-74f5be38489e

Selva Prabhakaran (August 22, 2021). ARIMA Model – complete guide to time series forecasting in python. Retrieved from https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/

Sushmitha Pulagam (Jun 26, 2020). Time Series forecasting using Auto ARIMA in python. Retrieved from https://towardsdatascience.com/time-series-forecasting-using-auto-arima-in-python-bb83e49210cd

Questions

1. Why was the target value of new cases chosen above other variables?
2. How accurate and reliable is the data?
3. Are state testing guidelines controlled for?
4. How does interstate travel affect the results?
5. What role do federal guidelines play?
6. Are penalties for mandates accounted for?
7. What effect does average age of population have?
8. Is it possible to predict the number of cases for a county if no regulations had been implemented?
9. If a modest or very low benefit is determined would the regulation still be recommended?
10. Are mandates enacted as a result of case numbers?