Super Spreader: How Political Rallies Spread Disease During Covid-19

Bickston Laenger | Yanni Ma

Both team members have contributed a similar amount of effort

Introduction

Covid-19 was an international pandemic that disrupted the lives of millions and was the first true pandemic of the information age. It caused us to reevaluate how we could deal with a pandemic at scale with our ever-globalizing world, and in the process, demonstrated how seriously people were willing to take an infectious and deadly disease. There was also a spread of misinformation that came with it.

A large divide in how people chose to react to the pandemic (at least in the United States) was along party lines, with the Republican party being much more skeptical of the effects of the disease as well as the mitigation efforts recommended by the CDC such as mask wearing. On the other hand, Democrats' messaging was more oriented around following the guidelines, and it was trying to err on the side of being too cautious. The goal of this project is to test our hypothesis that the side that did not take the disease as seriously ended up spreading the disease more.

In our project, we aimed to investigate not only the differences among each of the political events but also how they contributed to the spread of the disease. We planned to compare the 3 main event types that were happening during the pandemic: Republican-leaning, Democratic-leaning, and all other events. We employed different epidemiological models to demonstrate which side's political events affected and spread the disease the most.

Data

Our first data source is the Crowd Counting Consortium data repository, a collection of metrics hosted on GitHub by the Nonviolent Action Lab. This repository collects data on various political protest events in the USA, defining protest broadly. Various CSV files from the CCC include information about protests, demonstrations, rallies, and more. Available on GitHub, the CSV files include entries for the number of participants, event type, and geographic location.

The second data source, accessible from the New York Times, is a comprehensive dataset for COVID-19 case and mortality data for the US. For this dataset, we analyzed changes in case numbers, investigating how local infection rates evolved in response to specific SSEs. The dataset provides cumulative counts of COVID-19 cases and deaths at the national, state, and county levels. The goal of this dataset is to compile comprehensive time series data, offering a detailed record of the pandemic's progression. We can then correlate the case and death count by county to compare event happenings on both sides with the case counts over time.

The datasets overlapped in a data parameter called a "FIPS code," a 5-digit code that is used to standardize across counties in the United States. When the datasets were merged, we could then compare events for any given county with the rise in case counts over time.
The CCC dataset has over 72k events countrywide recorded with information such as date, location, number of participants, and, most importantly, political leanings for each. The NYT dataset contained data for each county for each day, totaling over 880k entries for just 2020 and almost 1.2 million entries for 2021. Altogether it was ~500MB of data.

Problem Definition

We aimed to model the spread of COVID-19 at political rallies and compare it to other types of SSEs. Using SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous variables) and SEIR models, we determined whether political events of a specific valence exhibited unique characteristics that amplified disease transmission.

Proposed Approaches

Algorithms and Models

ARIMA Model - Used commonly for time-series forecasting and analyzing the trends and seasonality in data, the Autoregressive Integrated Moving Average model combines three components that can help us isolate the effects of superspreader events on case spikes: Autoregressive to capture the relationship between any given observation (in this case an event) and a number of lagged observations, Integrated to difference the time series and make it stationary, and Moving Average to model the relationship between an observation and it's residual error. We used this model to predict both the comparisons over time of how a certain number of Covid cases could change and to then predict against how they would actually change which we then compared to the ground truth.
SEIR model - The Susceptible-Exposed-Infectious-Recovered model is a model widely used in epidemiology to simulate the spread of infectious diseases. We used this model to predict how the number of events correlated with the spread of the disease.
Combining the results of these two together helped us compare things such as when events happen to when spikes in cases happen right after, showing that there is a correlation, but not necessarily causation of the event leading to the spike. They also showed semi-realistic outputs of what we would expect the disease to do when given a population size similar to the counties that the events happened in.

Experiment

We implemented the models from above and got several interesting results. We ran the models for each county across the US (of which there were 1736 in our datasets) looking for any anomalies or useful outcomes.

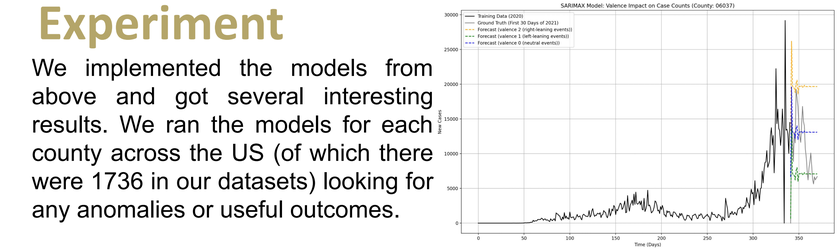

Fig 2: SARIMAX with average event size

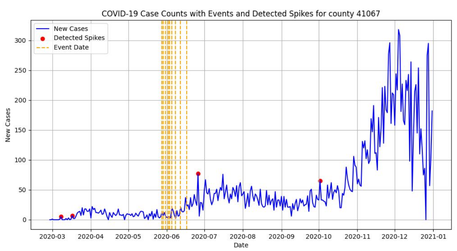

Fig 1: Comparing Covid spikes to when events happened

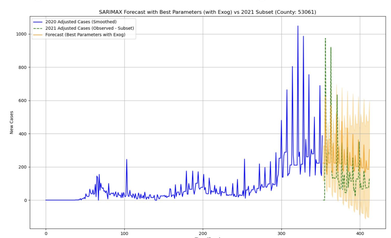

Fig 3: SARIMAX fitting result

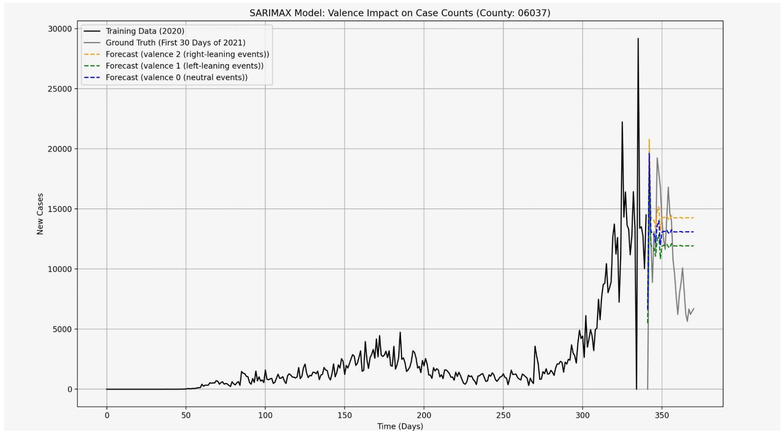

Fig 4: SARIMAX comparing different political events to their projected outcome

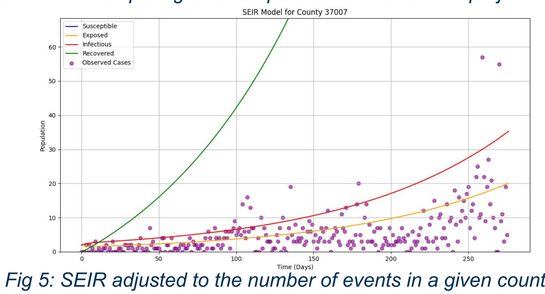

Fig 5: SEIR adjusted to the number of events in a given county

Results

We found two telling results from our outputs: first, that any given event didn't have that much of an effect on causing huge spikes, and second, comparing all events shows that Republican ones were indeed the most likely to spread Covid, followed by control events and then Democratic events. Figure 3 above is from Los Angeles County in California and shows that the predicted cases where the exogenous variable is equal to 2 (Republican-leaning events in our dataset) are higher than both where valence is 0 (no political leanings) and where valence is 1 (Democratic-leaning). This supports our initial hypothesis that Republican-leaning events would be more likely to spread the disease. Still, it was a little surprising how much lower the projected Democratic-leaning spread was even compared to the projected control spread. This follows a general conclusion that ignoring safe disease practices leads to a greater spread of the disease.

Conclusion

The findings of this study show a straightforward but important point: those who did not follow expert recommendations in regards to the pandemic and public health likely contributed the most to increased case counts. Our results suggest that no single event was a major driver of significant spikes in case counts. It's likely that each event contributed incrementally, compounding the case counts over time.

The process of analysis presented several challenges, particularly in working with the datasets. For instance, the valence variable had to be recoded to properly differentiate between the effects of different political events as an exogenous variable in the ARIMA model. The COVID-19 case data was incomplete, with dates showing zero case counts. To address this, we applied data cleaning and imputation strategies to interpolate missing values, aiming to produce more realistic projections. Another complexity was tuning the ARIMA model. The various underlying mechanisms of COVID made it challenging to select parameters that captured the full dynamics of the data, so model optimization took longer than expected.

Looking ahead, we aim to expand our analysis to include a broader set of counties, as time constraints limited us to a subset of the 1,736 counties available. We also plan to explore alternative models to validate whether our conclusions remain consistent across different methodologies. By refining both our approach and the data, we hope to gain deeper insights into the relationships between political events, public health behaviors, and disease spread.