Statistical Analysis of Temporal and Spatial Trends in US Covid-19 Cases and Deaths

Jason Gong and Micah Swann

Introduction

Covid-19 is a novel, highly contagious, acute respiratory virus that was first identified in December 2019 in Wuhan, China. Over the course of the following 14 months, this virus spread rapidly to every corner of the globe, becoming one of the deadliest pandemics in recorded history. In the United States, the first confirmed Covid-19 case was identified in January 2020 and by mid-March there were confirmed cases in every single state and North American territory. In the midst of this rapid pandemic spread, epidemiologists and modelers struggled to accurately forecast the spatial and temporal trends in cases and deaths. However, with regularly updated, publicly-available covid tracking data, a sufficient amount of data now exists to retroactively examine how cases and deaths evolved over the course of this 14 month period. This study utilizes the New York Times Covid Tracking Data to statistically analyze trends in the timeseries of Covid-19 cases and deaths as well as the spatial development of cases at the state level across the united states. Through the us of this timeseries data, an autoregressive integrated moving average (ARIMA) forecast model is employed to predict future cases and deaths at the national level. Furthermore cluster analysis is employed to find clustering trends in cases and deaths at the state level. Demographic data is further incoropated to examine the signficance of different factors in the model development process.

Data Description

Data Sources

Due to the fragmented nature of the US public health system, there is no centralized governmental data repository that is updated daily with Covid-19 case and death data. Instead, this study obtained data from the New York Times (NYTimes) Covid-19 Tracking Project (https://github.com/nytimes/covid-19-data). The NYTimes relies on dozens of reporters across multiple time zones to regularly update this tracking database with new information from press conferences, report releases, and local databases. Datasets utilized in this analysis reported the daily cumulative case and death counts in the US aggregated at the national, state and county level (US.csv, US-states.csv, US-counties.csv), respectively. Demographic data on state populations and economies were also obtained from the US Census Bureau. ## Data Formatting

All data analysis and visualization for this study was conducted in RStudio. The dataset was filtered to only examine cases and deaths reported from the beginning of March 2020 through the middle of February 2021. Raw data was reported as cumulative cases and deaths through time. To examine daily statistics, a filtering function was applied to calculate the finite difference between each consecutive reporting day.

Exploratory Data Analysis

Timeseries Visualization

An initial exploratory data analysis (EDA) was conducted to both elucidate trends and characteristics of the dataset and to guide the model development process. To better understand the temporal evolution of daily new cases and deaths in the US, a timeseries for both of these parameters was first generated (Figure 1). The timeseries of daily US Covid-19 cases depicts four distinct regimes in the change in daily covid 19 cases throughout the course of the pandemic. From March through the end of May 2020, the number of cases grew logistically; growing exponentially in March before asymptoting at a maximum daily new case load of 25,000-30,000 individuals through April and May. Similar but larger magnitude growth trends are evident in June through August, asymptoting around 65,000 daily new cases, and October through December 2020, asymptoting around 250,000 daily new cases. Note that the sharp drop in cases around the end of December is likely a reflection of a decline in reporting around the winter holidays and not a reflection of the actual drop in the real case load. From January 2021 onwards, the case trend differs from the early regimes, with a noticeable linear decline in the reported case numbers through December. Another interesting aspect of this dataset is the seven-day oscillation in the case numbers. New case numbers always tend to be lower on Saturdays and Sundays than during weekdays, reflecting the fact that many labs do not report case numbers on the weekend. The timeseries of daily Covid-19 deaths shows a similar logistic growth rate to the case rate in the early spring 2020. However, the number of deaths, drops significantly from mid-May 2020 and oscillates around 1000 cases a days until November 2020 when the number of daily new deaths rises again, fluctuating around 3000 deaths per days. The seven oscillations, observed in the timeseries of new cases, is even more prominent in the death data, with significant drops in reported deaths during the weekend. This initial visualization and review makes evident that that there are clear similarities and differences in the functional trends between both datasets.

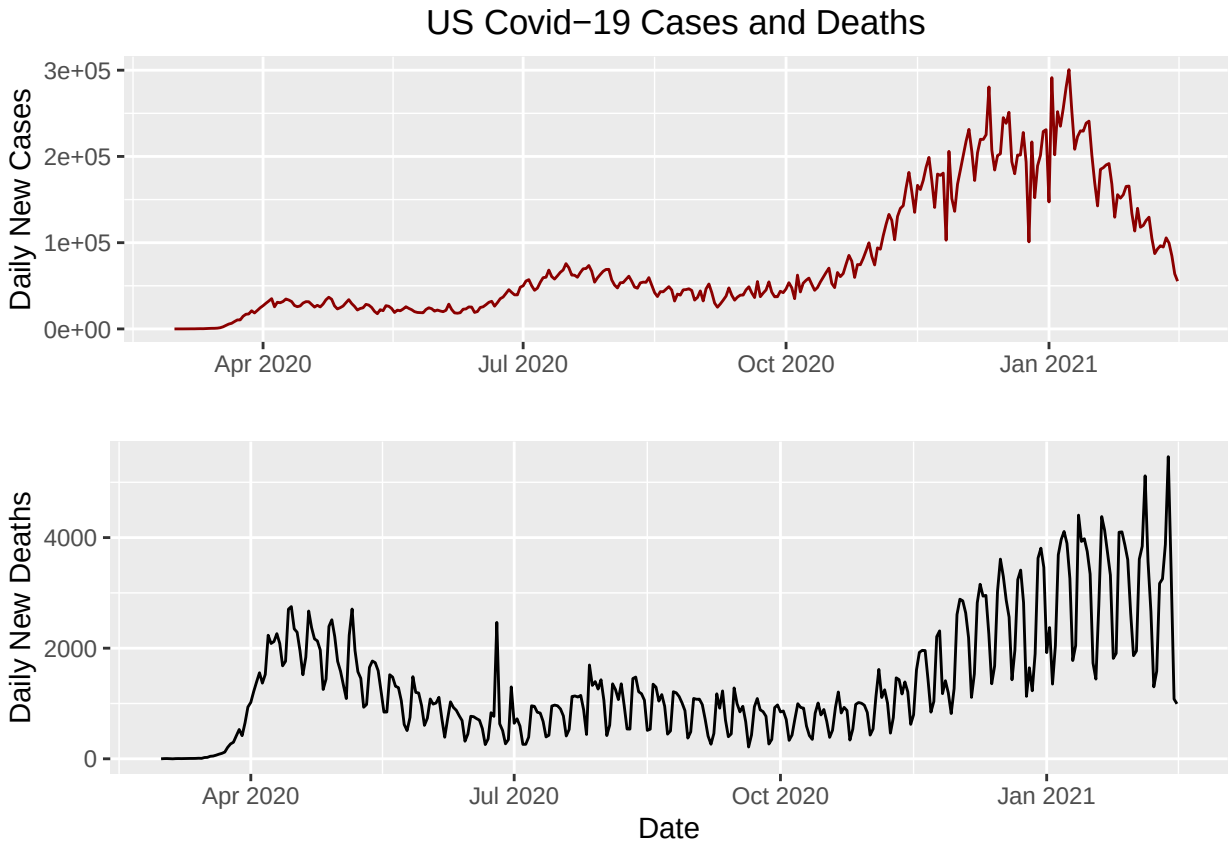

Figure 1 - Timeseries of daily new Covid-19 cases (top) and deaths (bottom) in the United States, March 2020 - February 2021.

Case-Death Scatter Plot

To further examine how the relationship between daily new US covid cases and deaths changed through time, a scatter plot of these two variables was generated (Figure 2). Having previously identified four distinct regimes in the case growth rate, these points were colored by the season for each observation. The scatter plot further highlights the different trends in the daily case to death ratios during each of these time periods. In the spring 2020, there is an significantly evident positive correlation in the case to death ratio. The points in this season are all closely clustered with the deaths growing exponentially with cases. In the summer 2020 data, there is no clear positive or negative trend in the correlation between the two parameters with the points scatter roughly in a circle. In the fall season, again a positive correlation is evident, however the case to death ratio is four to five times that observed in the spring. Finally in the winter (December 2020 – February 2021), the ratio of cases to deaths decreases but is still positive.

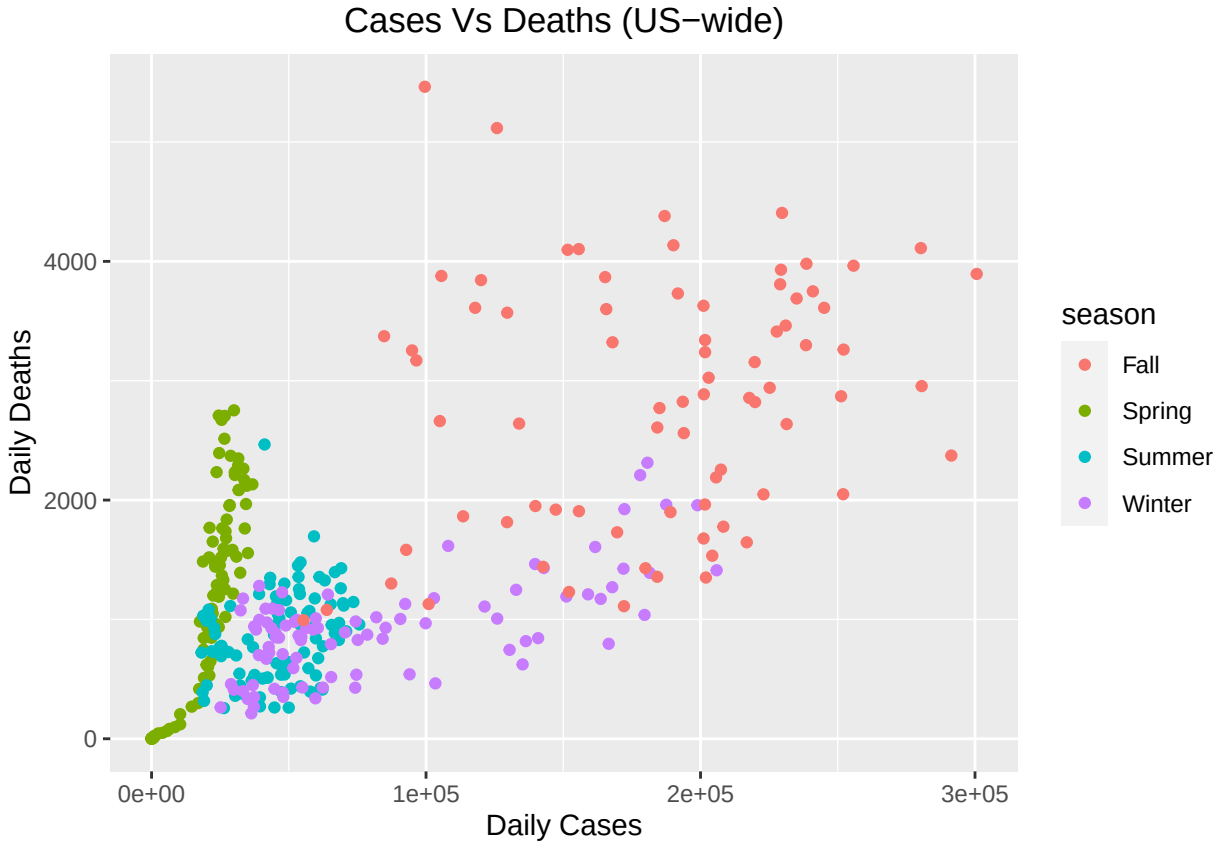

Figure 2 - Scatterplot of daily new Covid-19 cases vs deaths in United States. Point colors indicate season

State-level case and death boxplots

To invetigate the temporal trends in cases and deaths at the state level boxplots were generated that visualized the distribution of average daily cases and deaths in each state binned by month (Figures 3 and 4 respectively). Cases and deaths are shown as per capita values to normalize for state population. While both boxplots follow similiar trends to those observed in the raw timeseries data, the large spread in per capita rates show that a few outlier states were mainly responsible for the peaks observed in time. For example in November, 2020 the per capita case rate in California was 0.0016, more than twice the median state case rate. This single state was responible for a large percentage of the increase in Covid cases during the fall and winter of 2020. It's also interesting to note that the while the highest per capita death rates for a single state were observed in April 2020 (New York and New Jersey), the highest median death rates were observed in December 2020 and January 2021, reflecting the widespread death toll of the virus across the entire country.

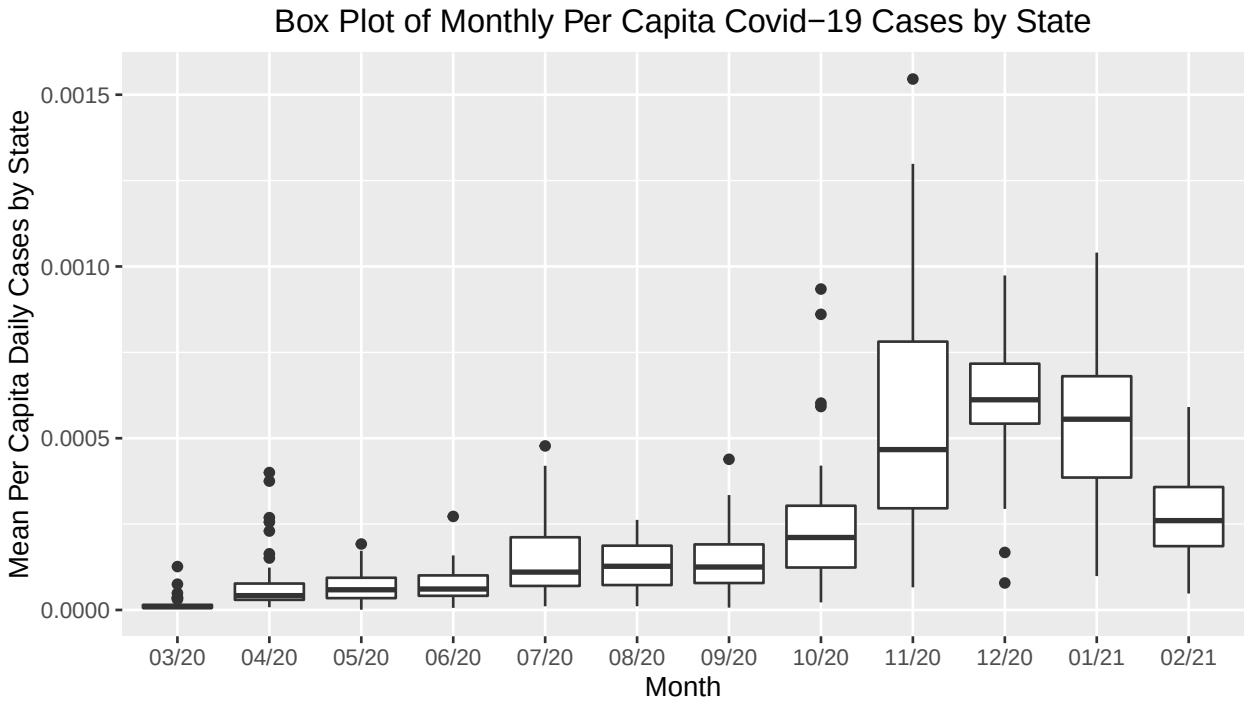

Figure 3 - Boxplot of daily average new Covid-19 cases by state across US. Data is binned by month

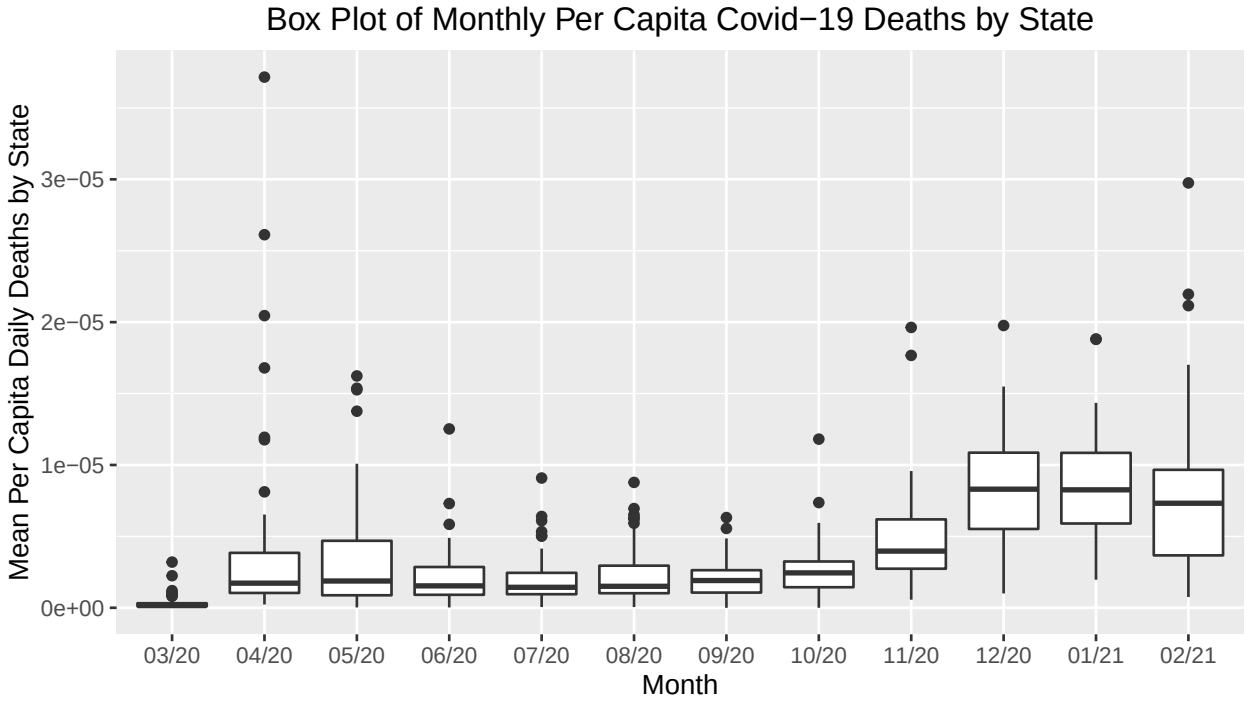

Figure 4 - Boxplot of daily average new Covid-19 deaths by state across US. Data is binned by month

Regional and Demographic Trends

State level spatial and demographic trends are also investigated by generating boxplots of monthly-binned, mean daily per capita case and death rates grouped by region (Figures 6 and 7). The four regions (Northeast, South, North Central, West) are provided in the state.region default dataset in r (Figure 5). Alaska and Hawaii, not pictured, are included in the Western region. Boxplots were also generated using density as a grouping factor (Figures 8 and 9). Regional patterns provide into how the virus spread around the US and what parts of the country drove spikes in cases and deaths at different points in time throughout the study period. For example, the Northeast had the highest magnitude and median of cases and deaths in March, April and May, largest driving the early spring rise in cases. Subsequently, the South the highest totals in July and August, promoting the summer rise. However both of these peaks were dwarfed by the magnitudes observed in the Western States, primarily in Arizona and California, in October and November 2020, constituting the majority of the winter rise.

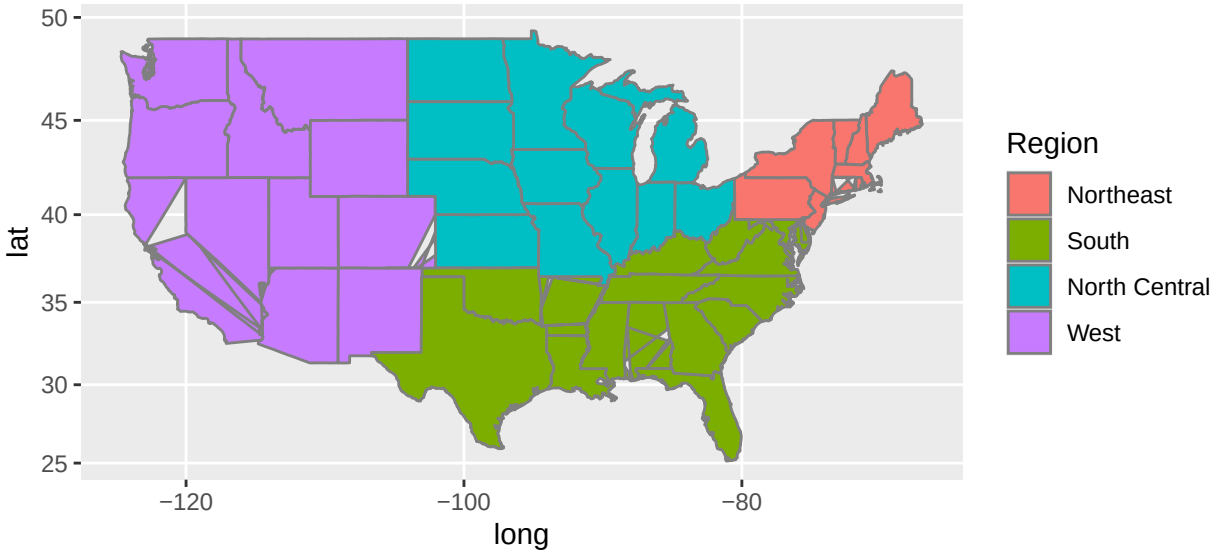

Figure 5 - Map of US States grouped by region

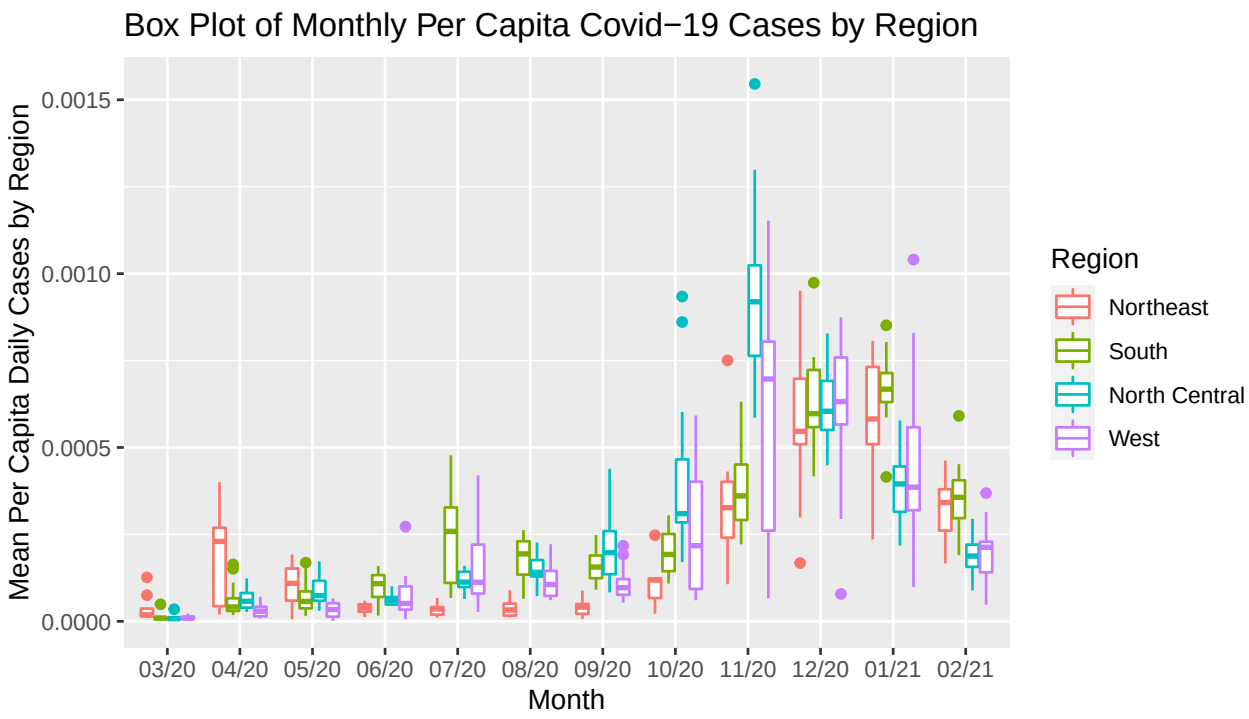

Figure 6 - Boxplot of daily average new Covid-19 cases by state across US. Data is binned by month and categorized by region.

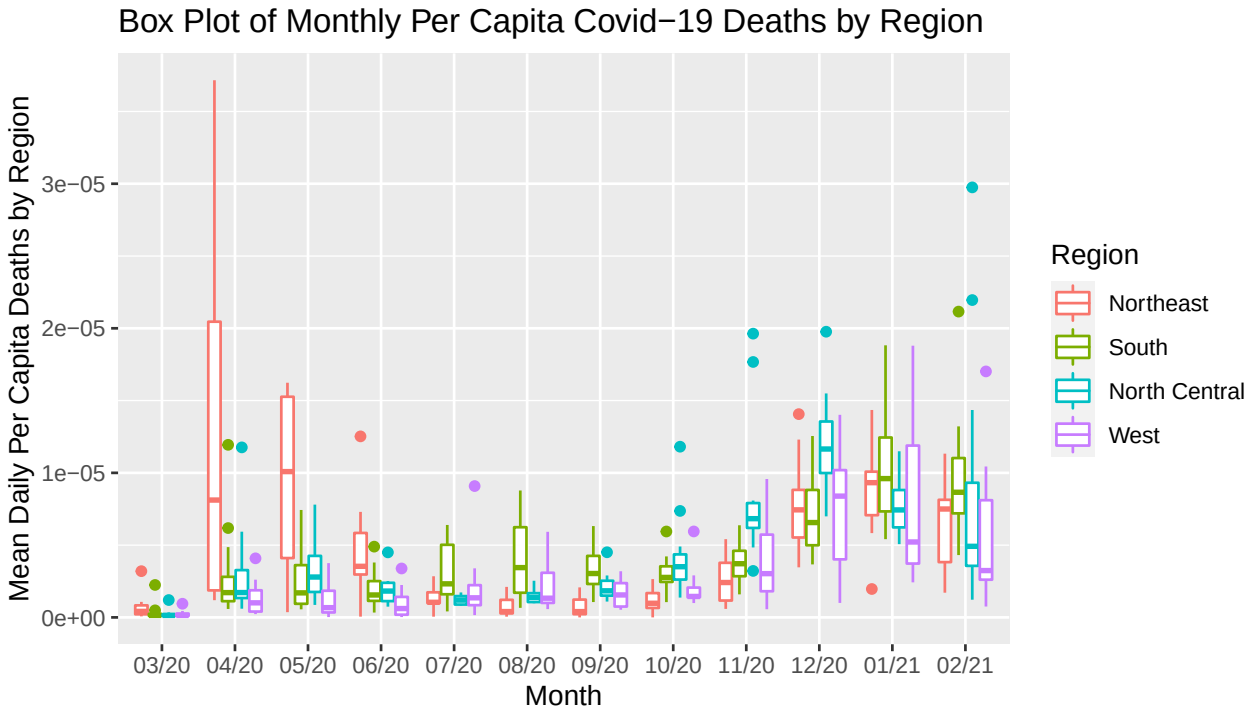

Figure 7 - Boxplot of daily average new Covid-19 deaths by state across US. Data is binned by month and categorized by region.

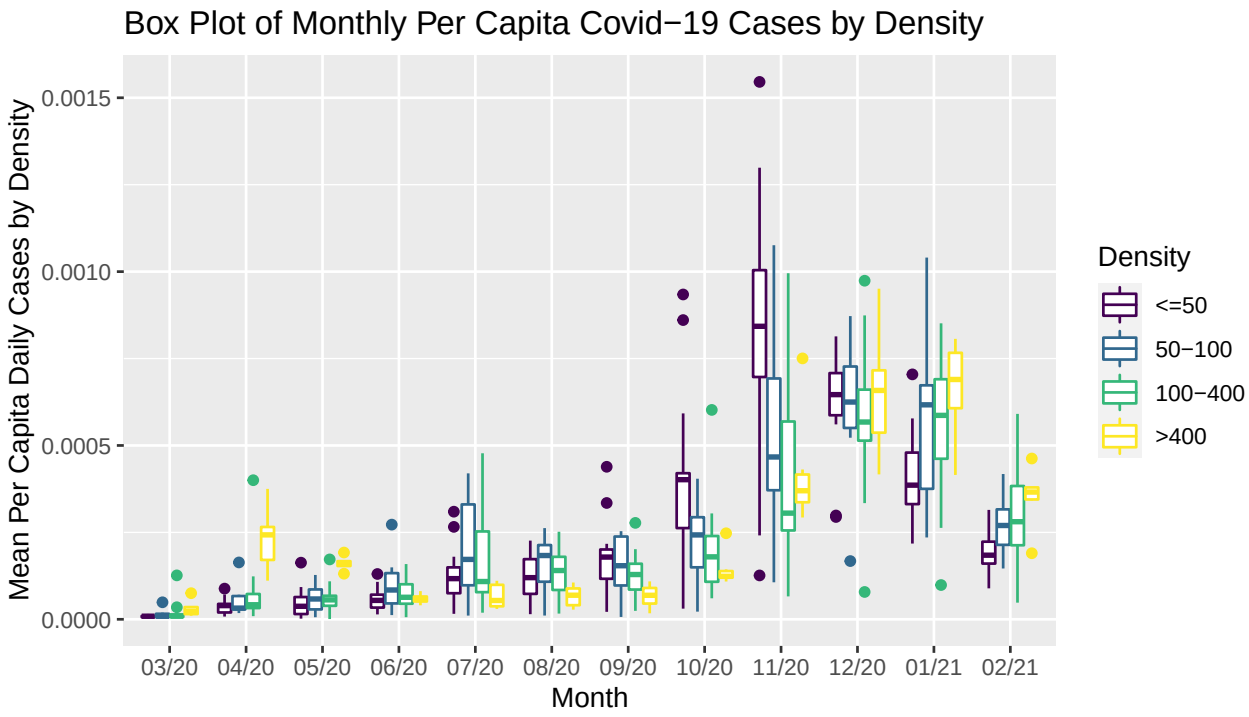

Figure 8 - Boxplot of daily average new Covid-19 cases by state across US. Data is binned by month and categorized by density [person per square mile]

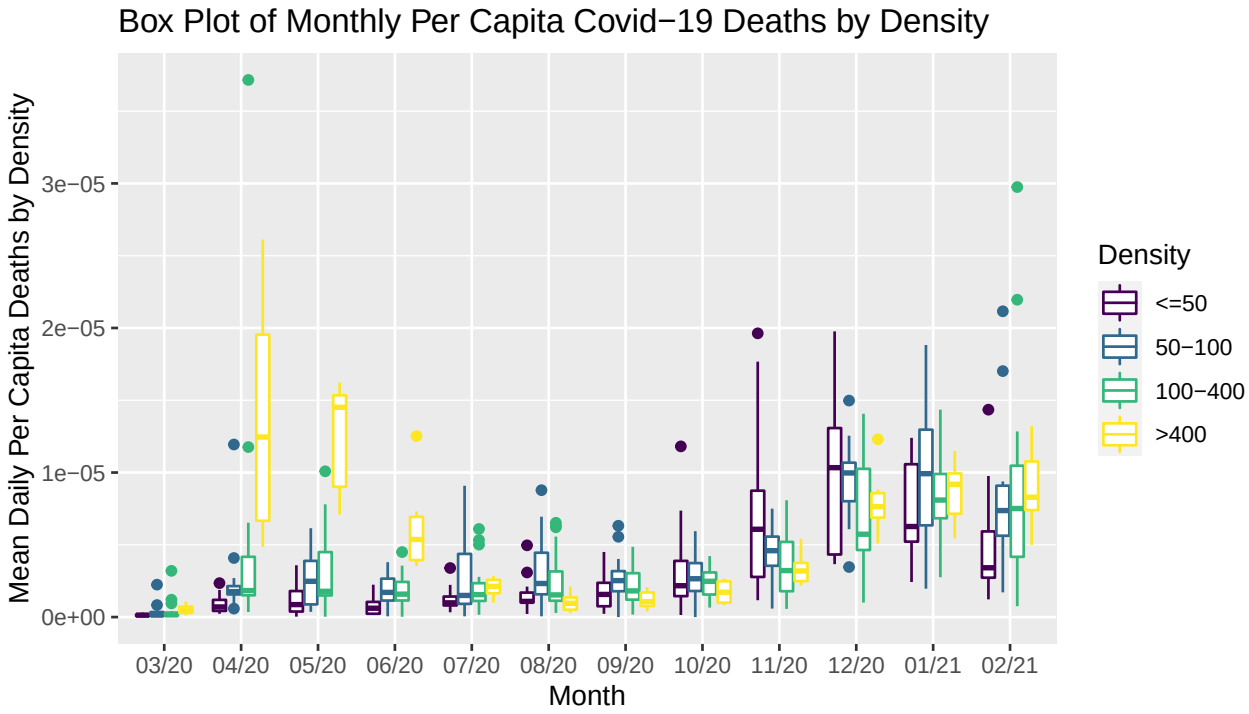

Figure 9 - Boxplot of daily average new Covid-19 deaths by state across US. Data is binned by month and categorized by density [person per square mile]

Algorithms

ARIMA Timeseries Forecast Model

Utilizing the timeseries data of US Covid-19 cases and deaths an ARIMA (Autoregressive Integrated Moving Average) forecast model was developed to predict the change in these two parameters over the subsequent 30 days following the final day that data was reported in this dataset, February 15th, 2021. For this analysis, univariate time series forecasting was conducted; only previous values of a single parameter were used to forecast future parameter values. ARIMA is a family of linear regression models that seek to statistically 'explain' trends in time series data and forecast the future. This model combines the mathematical formulations of two less sophisticated models: a simple autoregressive model (AR) and a moving average (MA) model. In a purely autoregressive model, a modeled value at time t (Y_t) is soley a function of previous lags and modeled coefficients.

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} + \epsilon_1$$

Where β_1 is the coefficient of lag 1 that the model estimates, α is in the intercept term and ϵ is the model error

Conversely, in a pure MA model, Y_t is dependent only on the lagged forecast errors (ϵ_t)

$$Y_t = \alpha + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q}$$

Combining these two equations, the ARIMA model formulation can be derived

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} \epsilon_p + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q}$$

ARIMA US Covid Cases Model Development Process

A cursory review of the timeseries in US Covid cases (Figure 1) indicates that the rates of these parameters are not a stationary process. It is evident that there are three distinct periods of increasing trends followed by a decline in 2021. Decomposing this timeseries into long term trends and seasonal variation can provide further indication if this is a non-stationary process. Figure 10 shows the raw US covid case timeseries data along with the decomposed trend, seasonal variation and white noise in the dataset.

This visualization makes evident that this a non-stationary process. This is validated by running an augmented Dickey-Fuller (adf) Test where the null hypothesis is that the dataset is stationary.

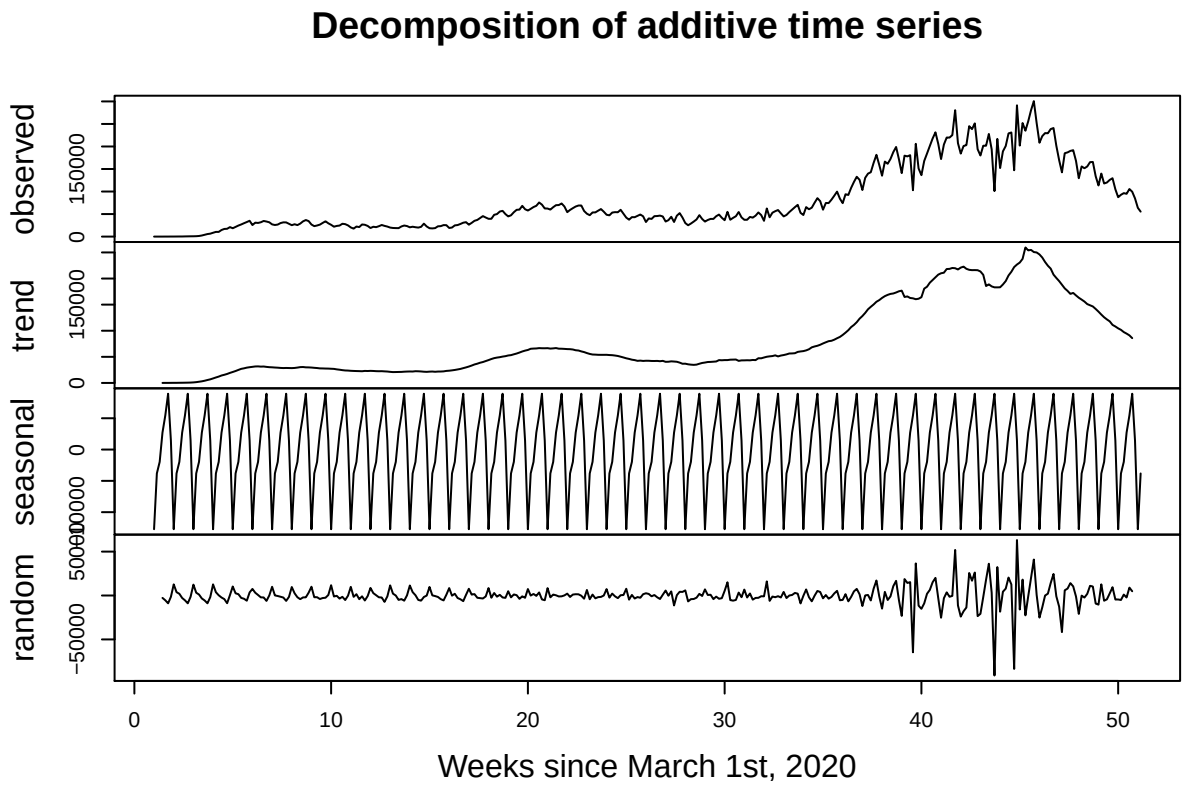

Figure 10 - Decomposed raw US Covid case timeseries Data

```
# Test for Stationarity
tseries::adf.test(new_cases.ts)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  new_cases.ts
## Dickey-Fuller = -0.68928, Lag order = 7, p-value = 0.971
## alternative hypothesis: stationary
```

The large p-value calculated in this test thus shows that the null hypothesis that the timeseries is non-stationary cannot be rejected..For ARIMA model to be valid, the non-stationary aspects of this dataset must thus be removed. This can be achieved by simply taking the difference between each data point. A plot of the decomposed difference-stationarized case timeseries dataset and the corresponding results of the adf test shown below. This dataset now no longer has any long term trends and the p-value for the stationary hypothesi is >0.01, so it is significant.

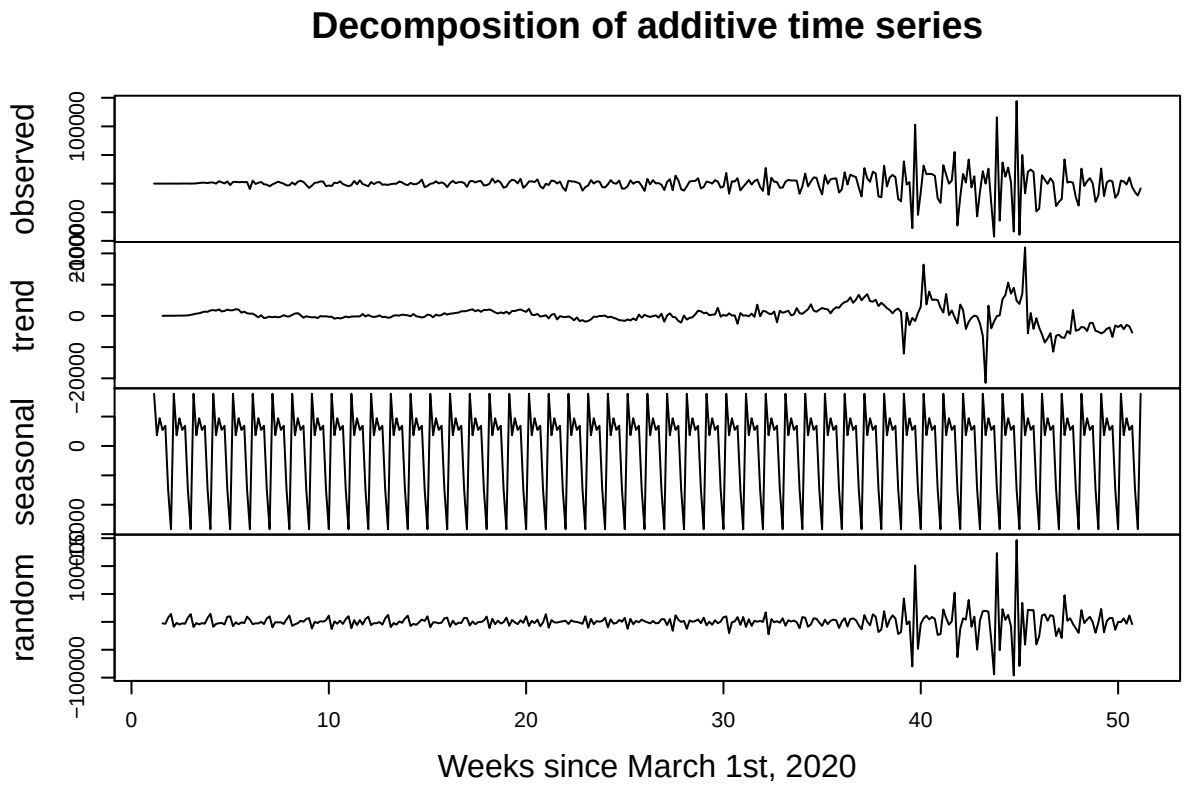

Figure 11 - Decomposition of difference US ovid case data

```
# Test for Stationarity with differencing
tseries::adf.test(diff(new_cases.ts))
```

```
## Warning in tseries::adf.test(diff(new_cases.ts)): p-value smaller than printed
## p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  diff(new_cases.ts)
## Dickey-Fuller = -5.2524, Lag order = 7, p-value = 0.01
## alternative hypothesis: stationary
```

The next step in the model development process is to find the optimal parameters (p,d,q) for the ARIMA forecast model. Each of these parameters correspond to the AR order, degree of differing and MA order of the model respectively. In this ARIMA model optimal parameters new to specified for both the non-seasonal and seaonal part of the model This can be accomplished by visualizing the autocorrelation function (ACF) and partial autocorrelation function (PCF) for the dataset (Figure 11). The ACF and PCF plots depict what lags have the highest correlation to the a given output. These plots both clearly shown that a seven day lag has the highest correlation. The auto.arima function in r can be used to automate the optimal parameter. Results of this algorithm indicate that the optimal parameters for the seasonal component are 0,1,1 and 0,0,2 for the seasonal component. To verify the validity of these parameters the model residuals were checked (Figure 12). The residuals are normally distributed around 0, indicating that this model is unbiased. The weekly fluctuations are preserved in both models, acurately reflecting the artificial decline in cases during the weekends due to reduced laboratory staff and runs. In actually the case load does not fluctuate with this magnitude as there is no natural process that would explain a weekly cycle in virus spread. This oscillations is even more pronounced in the death reports.

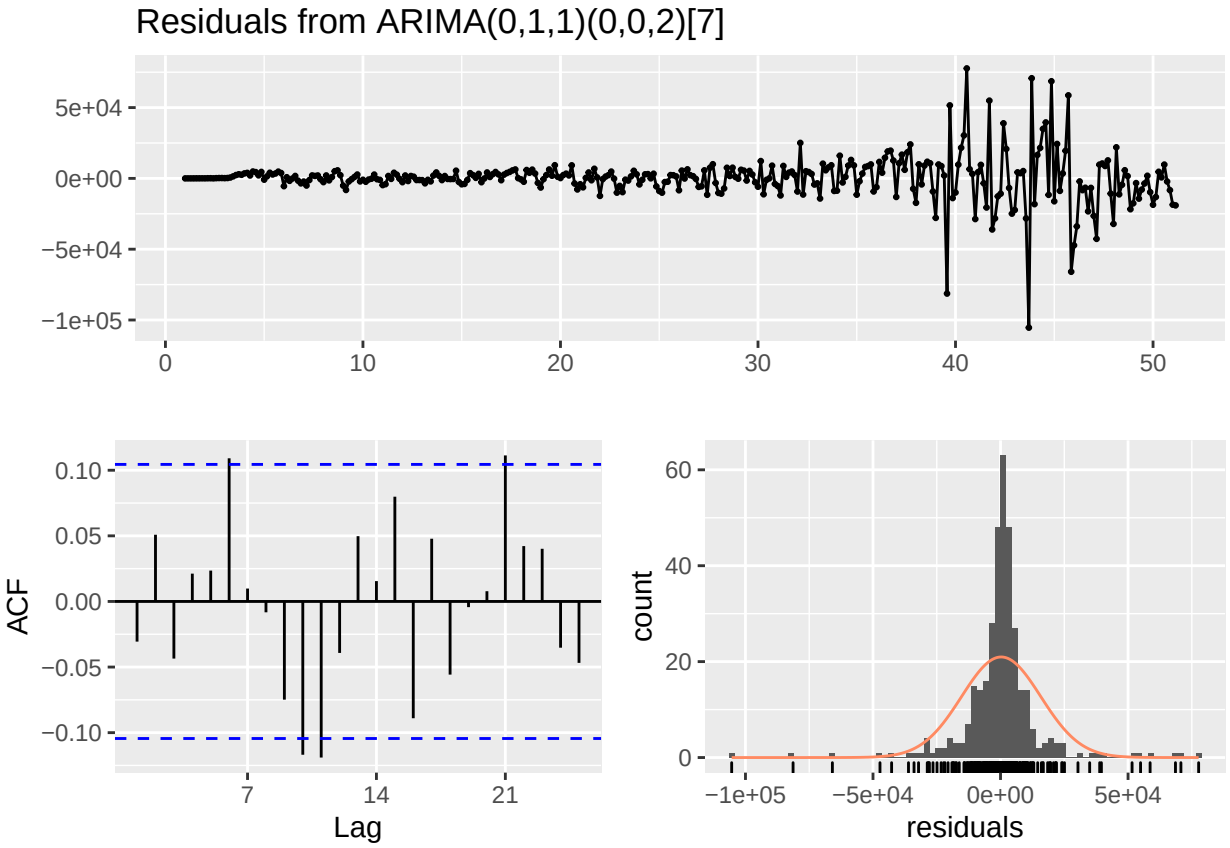

Figure 12 - Residuals Plot for ARIMA Case Forecast Model

ARIMA US Covid Death Model Development Process

The same model development process was repeated for the timeseries of US Covid-19 deaths. The raw data was decomposed and tested with adf to verify non-stationarity. A single finite differencing schemne was then applied to remove temporal trends. Finally the ACF and PCF of the data were plotted up to determine lags and optimal model parameters. In the death data, 7, 14, and 21 days were all signifcant lags. However the fact that these are multiples of seven it was assumed that a lag-7 model would continue to be sufficient. A key difference between the ARIMA models for cases and deaths are the optimal parameters for the non-seasonal and seasonal trends. For the death arima model, the parameters were calculated to be 1,0,2 and 0,1,1 for the non-seasonal and seasonal terms respectively.The residuals in this model too are normally distributed around 0, so the model can be assumed to be unbiased.

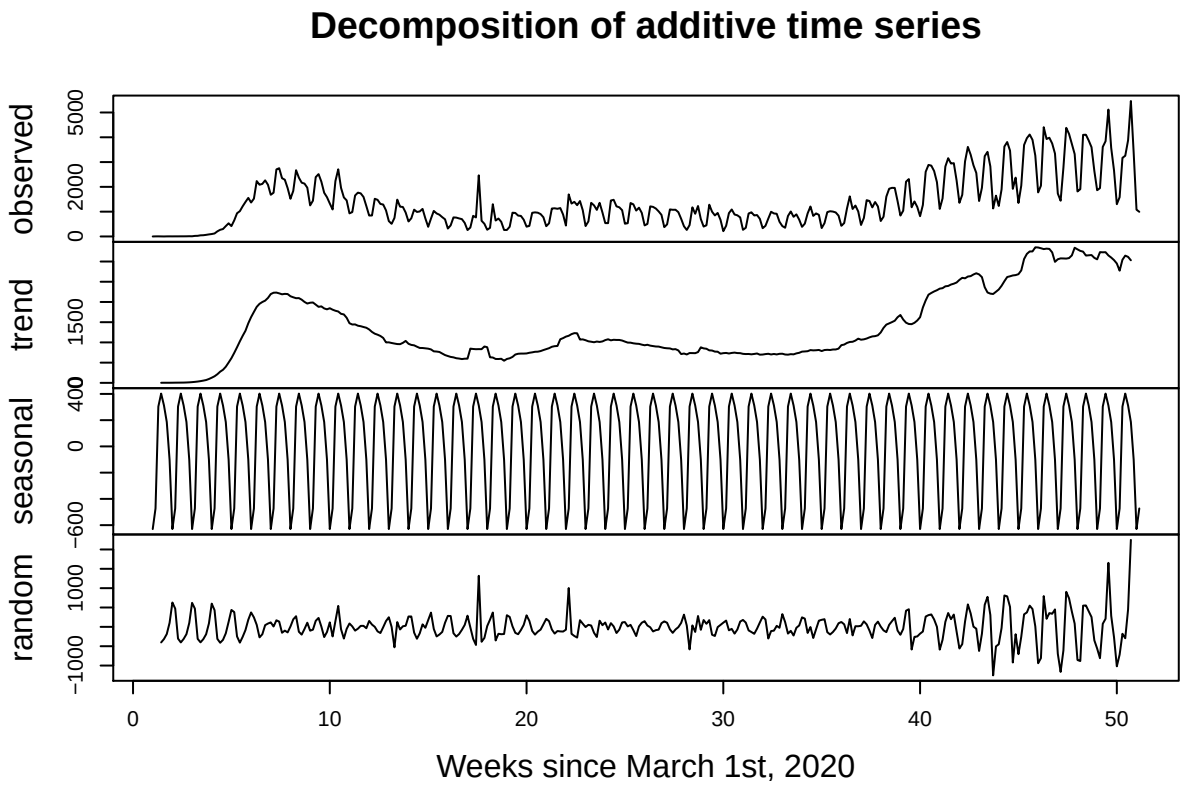

Figure 13 - Decomposition of raw US Covid-19 Death Data

```
# Test for Stationarity
tseries::adf.test(new_deaths.ts)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  new_deaths.ts
## Dickey-Fuller = -1.1801, Lag order = 7, p-value = 0.9087
## alternative hypothesis: stationary
```

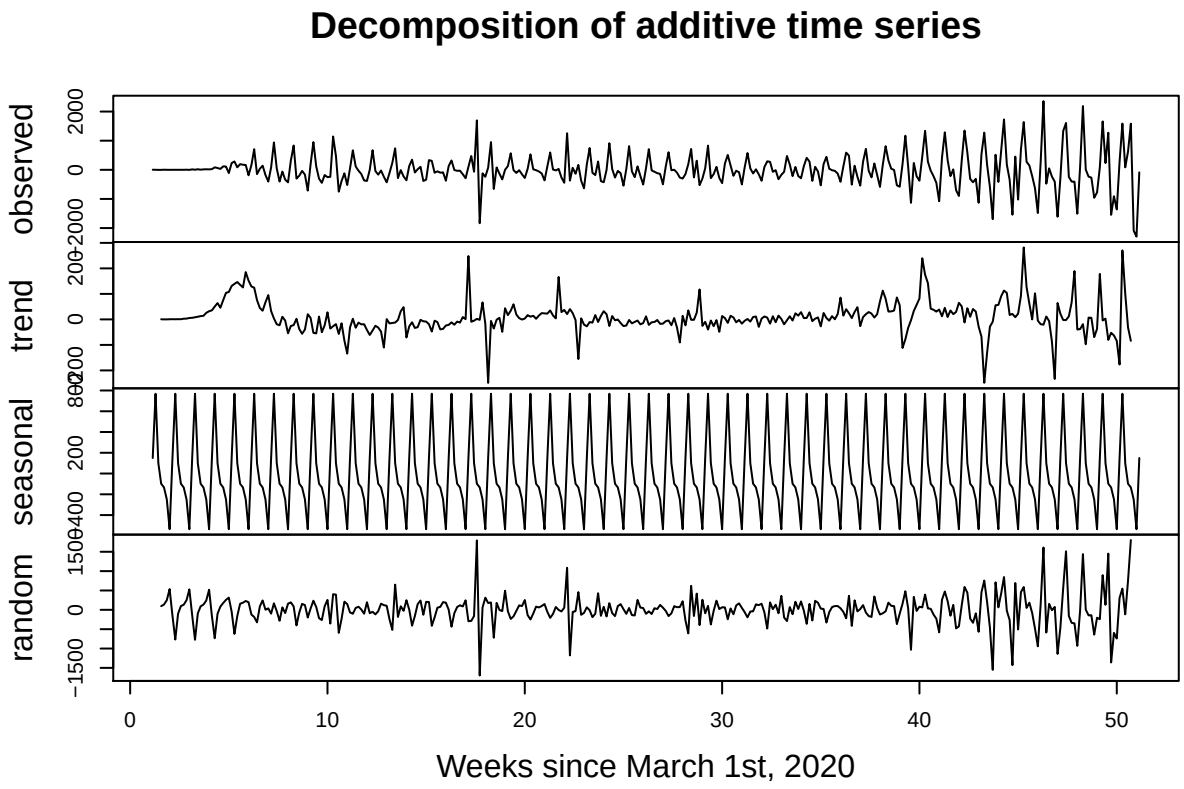

Figure 14 - Decomposition of differenced US Covid-19 Death Data

```
# using differencing method
tseries::adf.test(diff(new_deaths.ts))
```

```
## Warning in tseries::adf.test(diff(new_deaths.ts)): p-value smaller than printed
## p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  diff(new_deaths.ts)
## Dickey-Fuller = -8.1672, Lag order = 7, p-value = 0.01
## alternative hypothesis: stationary
```

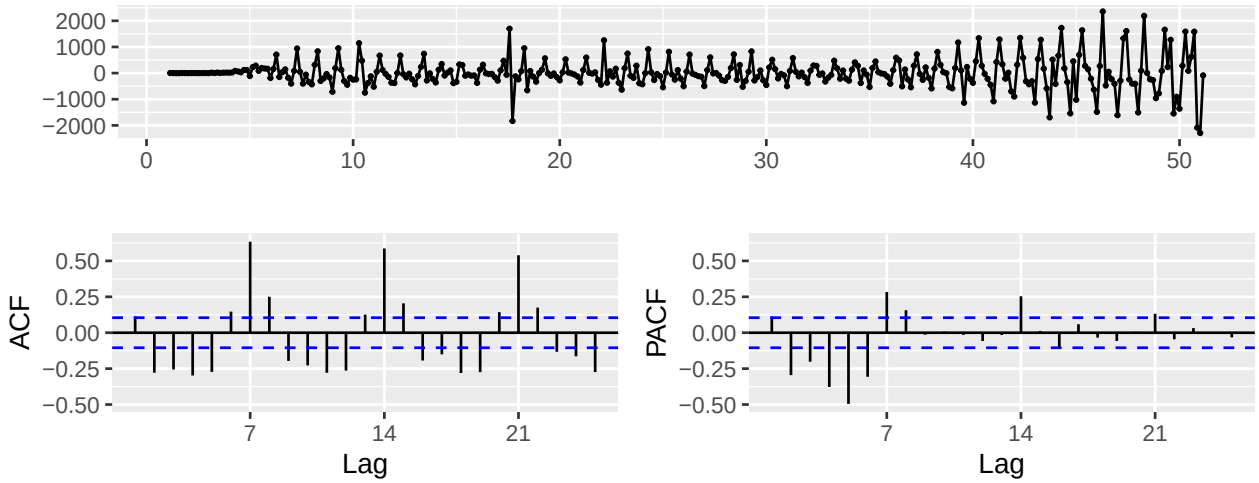

Figure 15 - ACF and PCF for differenced US Covid Death Data

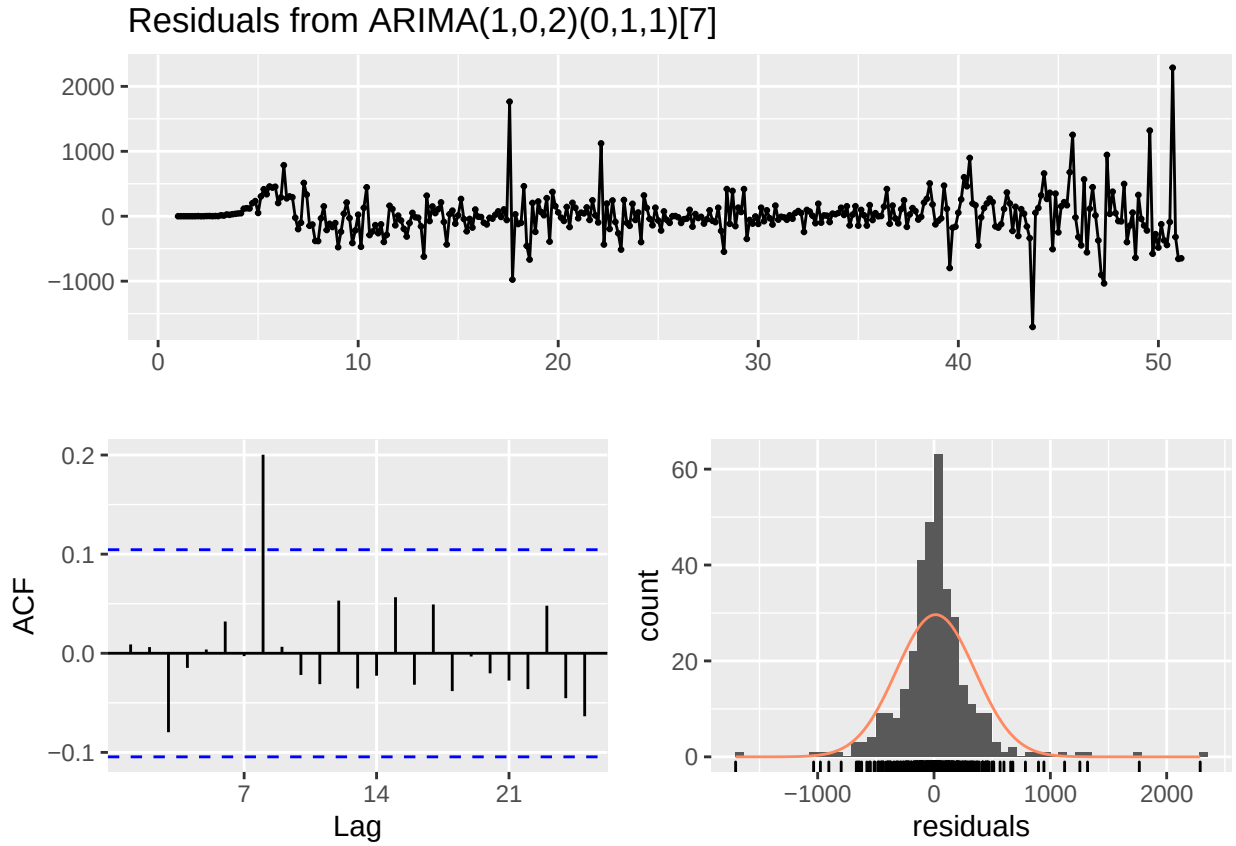

Figure 16 - Residual plot for differenced US Covid Death Data

Cluster Analysis

Results

US-wide Covid-19 ARIMA Forecast Models

The results for a 14 day forecast for US Covid-19 cases and deaths are shown in Figures XX and YY below. The ARIMA case model predicted an average decline in cases from 72,125 to 59.143 per day. On this day, March 1st 2021, the NYtimes reported 56,562 new cases, less than five percent difference from the model value. The ARIMA death model forecasted a decline deaths from 3,161 to 1,311 per day compared with 1,425 deaths per day actually reported, an 8.7 percent increase from the model. Both these models appear to be validated, as they are both within 10% of observed values. The seven day periodic osscilations in cases is preserved in both models. It's important to note that this osscilation is likely an artifact of reporting procedure and not a representation of the true variability in case load. Bergman 2020 conducted a high resolution temporal analysis of New York and Los Angeles testing data and deduced that daily variation in testing signifcantly explained oscilattion in case numbers. Furthermore ossiclation mortality across the US data were attribute as "artificat of reporting".

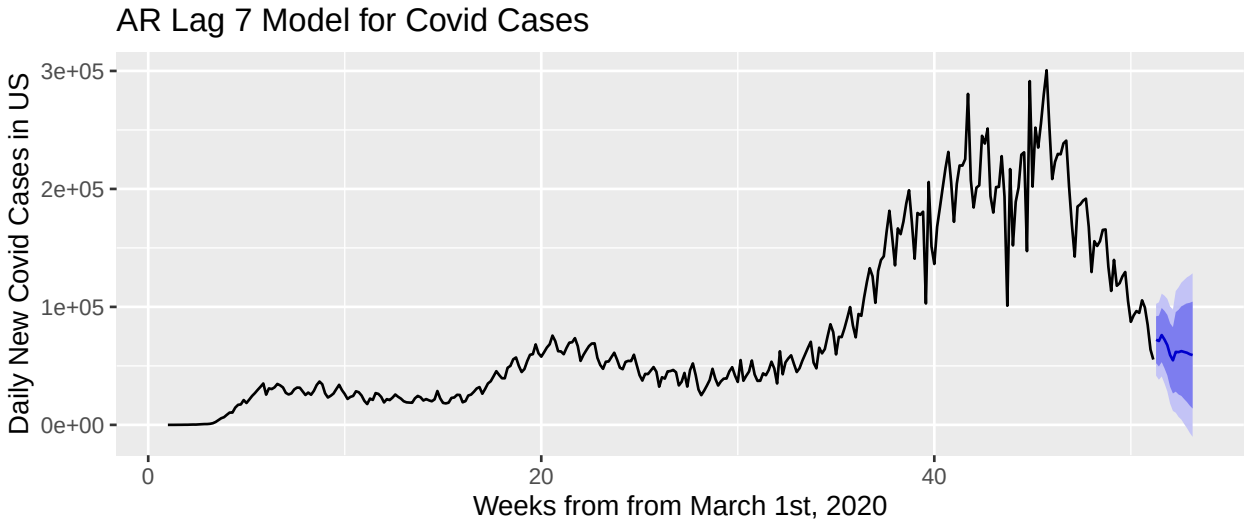

Figure 12 - 30 Day forecast for Covid-19 cases in US using AR Lag-7 Model. Dark blue indicates model results, blue and light blue bands indicate 80 and 95 percent confidence intervals.

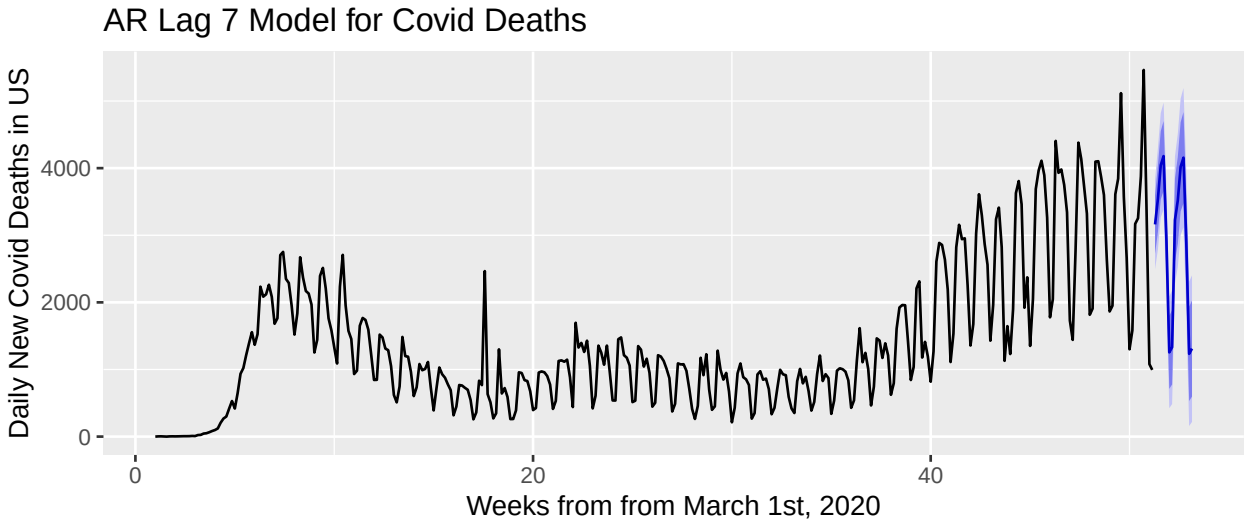

Figure - 30 Day forecast for Covid-19 deaths in US using AR Lag-7 Model. Dark blue indicates model results, blue and light blue bands indicate 80 and 85 percent confidence intervals.

Conclusions

Temporal Trends

Any interesting extension of this analysis would be to develope a multivariate ARIMA forecast model using both case and death parameters as predictors.

Spatial Trends

References

Bergman, A., Sella, Y., Agre, P., & Casadevall, A. (2020). Oscillations in U.S. COVID-19 Incidence and Mortality Data Reflect Diagnostic and Reporting Factors. MSystems, 5(4), e00544-20. https://doi.org/10.1128/mSystems.00544-20