Understanding the Impact of COVID-19 on the United States: A Time Series Analysis

Presented by Andrew Reese

1 Introduction

1.1 Background on the COVID-19 Pandemic

The global COVID-19 pandemic stems from an outbreak of the infectious disease caused by severe acute respiratory syndrome coronavirus 2. It has since been determined to be one of the world's most detrimental global health crises of the 21st century.

In late 2019, patients in Wuhan, China began experiencing similar symptoms to pneumonia that did not respond well to treatments typically used to treat such illnesses. On January 20th, 2020, the CDC reported the first case of COVID-19 in the United States from a sample taken on January 18, 2020 in Washington state. The disease rapidly spread across the world through travel and person-to-person contact and by March 11, 2020, the World Health Organization declared COVID-19 a pandemic.

This study aims to provide a comprehensive time series analysis of COVID-19 cases and deaths across the United States from January 2020 to March 2023. Models, trends, and patterns describe the advancement of the disease over the past 5 years and the factors that influence the spread. This study also highlights the potential relationship between the number of cases and deaths and the implementation of health policies, the release of the COVID-19 vaccine, seasonal patterns, and newly developed medical care.

Understanding the patterns and trends of the COVID-19 pandemic throughout time not only helps experts assess past responses to the pandemic but also to aid with future response and preparedness strategies for future pandemics. The conclusions from this time series analysis report provide insights into how different policies advances in technology and medicine influenced the trajectory of the disease.

1.2 Data Description

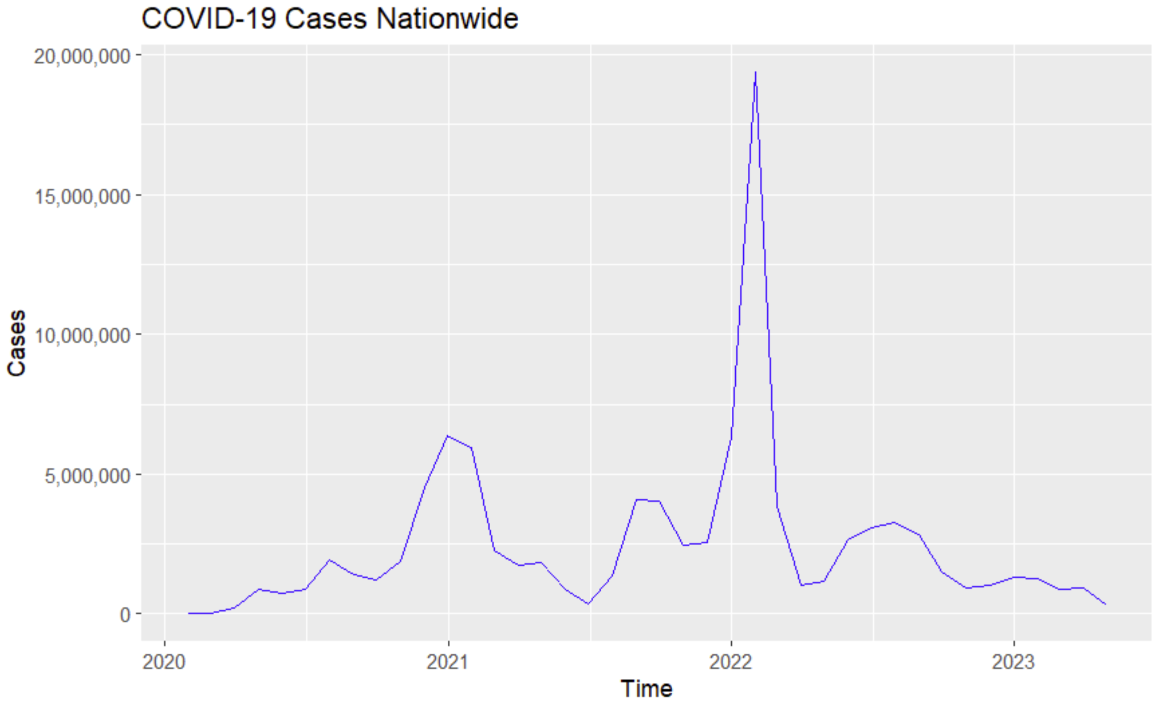
The analysis begins with two datasets chosen from USAFacts**[1]**. The first dataset breaks down daily total cases of COVID-19 per individual county in the United States, and the second set breaks down daily total deaths from COVID-19 per individual county. Each set begins with data starting in January, 2020 and concludes with data from April, 2023. To simplify the plotting of these data, each dataset was modified to be an aggregate, nationwide total of both cases and deaths respectively over that span of months. After doing so, each updated set has two columns - the first being months and the second being national totals over that month.

1.3 Research Focuses

The main scope of the analysis attributes fluctuations in monthly cases and deaths to certain governmental policies such as quarantine mandates, mask protocols, and also the introduction of vaccinations. We aim to determine if these policies had a statistically significant influence on the reported values.

2 Preliminary Data Analysis

2.1 Visualizing the Data

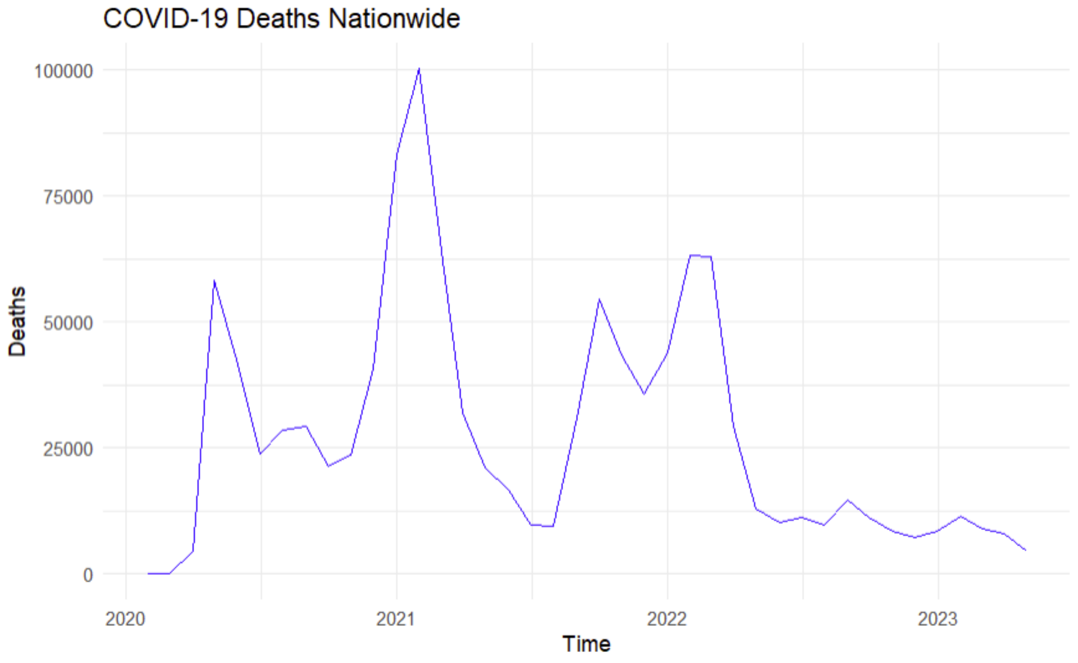


Above is the time-series plot of the monthly, nationwide cases of COVID-19 in the United States. As the plot makes evident, there are several major spikes in monthly cases. The first notable spike falls between October 2020 and January 2021. Per the CDC website, a new variant of the coronavirus called the “Alpha” variant first made its appearance in the US on December 29, 2020. This new variant spread exponentially and is the main cause of the spike. Another important factor that may have influenced the increase in cases is the season in which the spike took place. The winter months are generally much colder which causes people to spend more time inside. Viruses are known to spread more efficiently indoors because of the constant exposure to other people. On top of that, major holidays like Thanksgiving, Christmas, and New Years call for an increase in traveling and hosting family gatherings; both of which increase exposure to the virus and thus contribute to the spike.

The next significant spike falls between June and August of 2021. During this time, the Delta variant was introduced to the United States and spread rapidly. At this point, it is likely that CDC guidelines such as mandated quarantine, public mask policies, and others are beginning to be ignored due to the extended period of time under which the policies have been enforced.

People began to grow tired with staying home and practicing social distancing which likely also contributed to this particular spike.

Shortly following the delta variant came the most powerful variant yet, Omicron. Omicron hit the US on December 1st, 2021 and very quickly became the most significant, COVID-19 related issue yet. Its presence was the leading cause of the massive surge in cases that is seen at the end of 2021 and the beginning of 2022. **[2]**



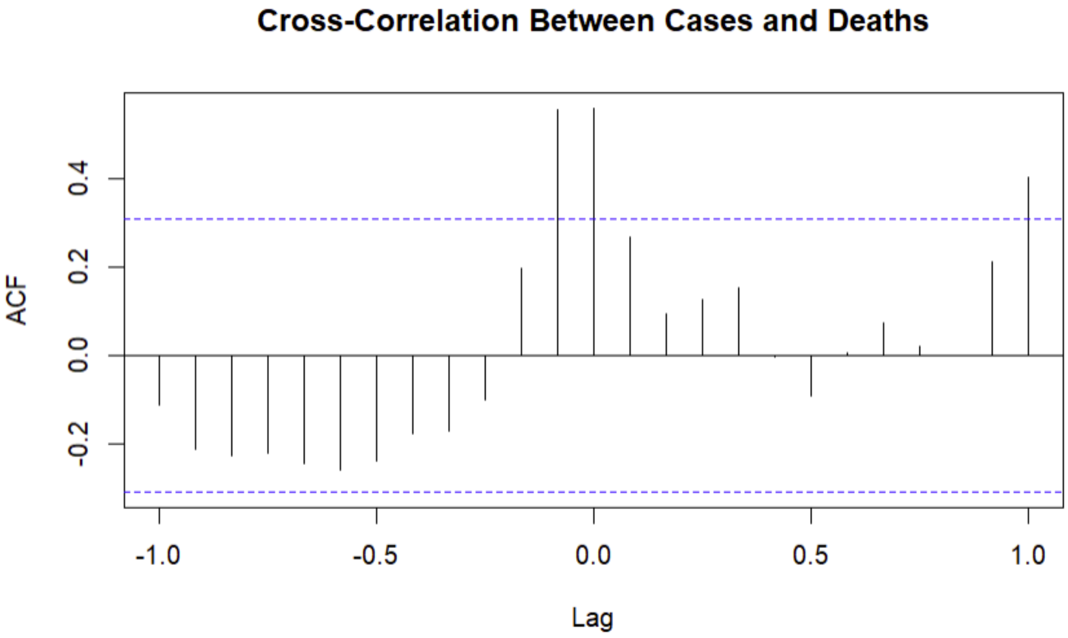
This is the plot of monthly deaths due to COVID-19 in the United States. While the values here are significantly smaller than those regarding COVID cases, we can see some correlation between the two plots. There are four main peaks in this plot with the first being between March and April of 2020. This peak is logical given the fact that the coronavirus made its way into the US in early 2020 and began to spread rapidly through the middle of the year. Since the virus was brand new at the time, the country was still figuring out the most efficient way to stay safe and gain control, so there were much less CDC guidelines to stop the spread. This quickly explains why there was such a surge during that period in 2020, but what about the other peaks on this plot?

Well, the remaining three spikes seen on this graph overlap almost perfectly with the three peaks seen in the COVID Cases vs. Time plot. This makes sense because the addition of new variants such as Alpha, Delta, and Omicron rendered the population defenseless. Any previous biological immunity to the virus was not helpful when facing a new variant, which led to the increase in deaths. **[2]**

2.2 Correlation Analysis

To analyze the correlation between the two data sets, we will use the cross correlation function in R. Cross correlation works similarly to the regular autocorrelation function, but there are few distinct differences when interpreting results.

1) The x-axis has both positive and negative values. The negative values, in this case, would indicate that the spikes on the COVID death plot precede spikes in the cases plot. The positive lags indicate that the spikes on the cases plot precede spikes on the death plot.
2) The y-axis also has positive and negative values, where the positive values indicate that the graphs move in the same direction at the same time, and the negative values mean that they move in opposite directions at the same time. **[5]**



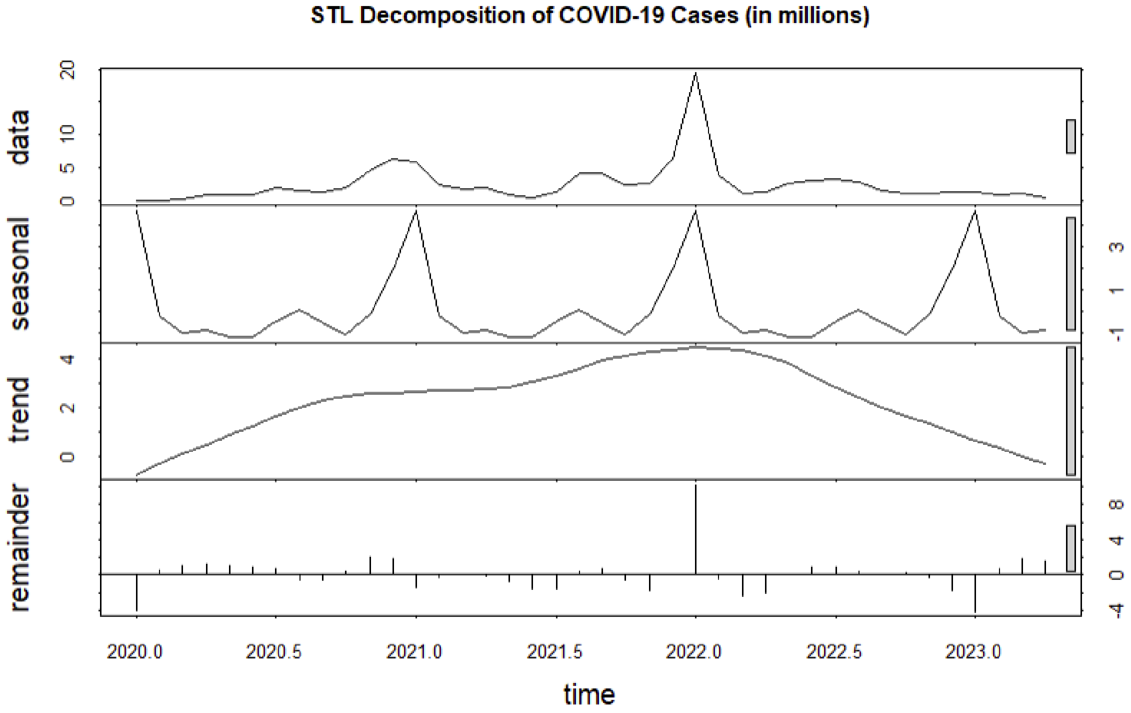
The plot above shows significant positive autocorrelation at around lag 0 and 1. The ACF at lag 0 indicates general overlap between the cases and death time-series plots, and the ACF at lag 1 indicates that a spike in cases precedes a spike in deaths by about 1 lag. Since 1 lag equates to a month in this case, it means that an influx of cases leads to an influx of deaths in around a month's time, which is consistent with the lifespan of the coronavirus. Generally, fatal cases of the coronavirus seem to last around a month's time, so again this makes sense. Another important fact is that there are no significant values of ACF for negative lags, meaning there are very few cases where a spike in deaths would lead to a spike in cases later on. This is the logical conclusion to make, but it is useful to test its truth.

We also computed the correlation coefficient, which came out to be 0.56. This may seem relatively low, but it makes sense when taking the previous results into account. Most of the time, spikes in cases won't usually occur at the same exact time as a spike in deaths. It is instead true that there is usually a time lag between the two events (around 1 month from our Cross Correlation plot). However, the direct correlation that exists can likely be attributed to simultaneously reported cases and deaths following things like new variants or unexpected COVID-19 surges. **[5]**

2.3 Data Decomposition

The next step in the analysis is to decompose the data and focus on seasonality, trends, and residuals. To do so, we make use of the STL function in R, which outputs four plots:

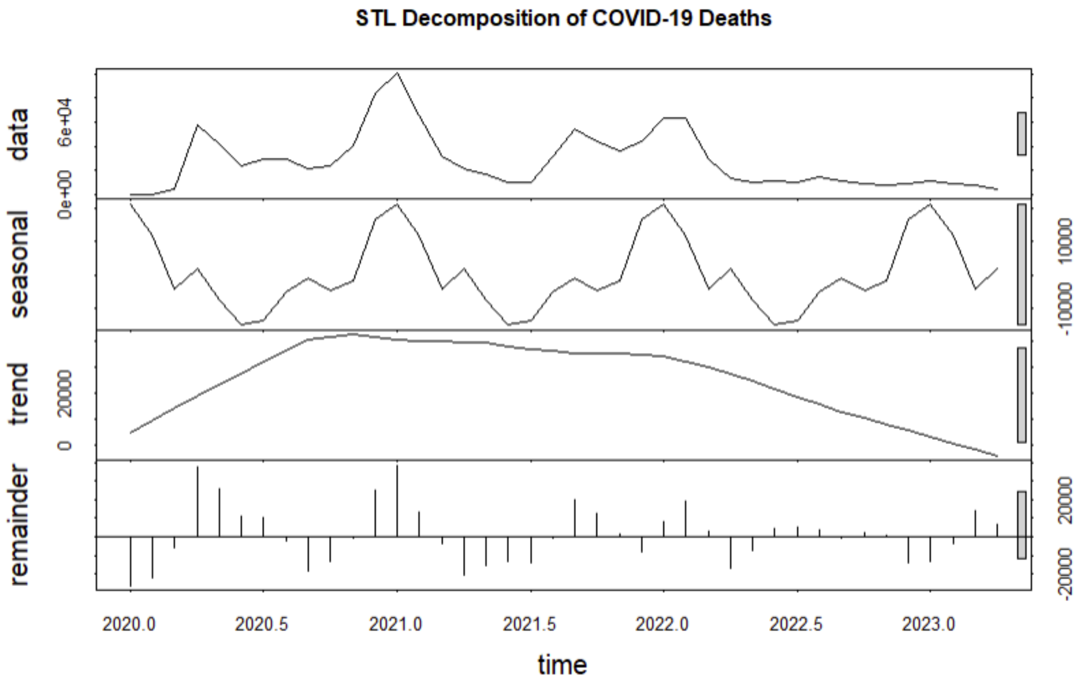
1) Data - The original time series
2) Seasonal - Shows recurring patterns and seasonal fluctuations
3) Trend - Plots the overarching direction of the data with time
4) Remainder - The residual and noise peaks after removing seasonal trends



Above are the STL plots of COVID-19 cases in the United States. To derive more meaning, we can break each plot down one by one. The data plot is the basic time series that we worked with earlier, so that has already been explained. The seasonal trend here shows relatively low values during the spring and summer months, and higher peaks during late fall and winter months. According to **[4]**, the winter months generally see the peak of many kinds of viruses, including COVID-19. This can be attributed to the colder weather which has been shown to decrease human production of extracellular vesicles (produced in the nose to attack incoming viruses). Also, as mentioned before, the holidays during the winter months increase traveling which is certainly a factor here as well.

The trend plot shows a gradual increase in cases in 2020, followed by a relatively flat period until 2022, and ending with a gradual decrease until the middle of 2023. The plateau period overlaps with the introduction of the various vaccinations from companies like Pfizer, Moderna, and Johsnon & Johnson. Outside of clinical trials, the administration of these vaccines began in December of 2020 **[2]**, which can explain the halt in rising case numbers. The decrease from 2022 to 2023 can be attributed to the increase in vaccinations nationwide and also the natural antibodies created against the virus after all that time.

The remainder plot at the bottom of the figure demonstrates any white noise that cannot be explained by seasonal and periodic trends. For the most part, the spikes tend to fluctuate around 0 which indicates the strength the seasonal and trend fit. Beyond that, the one peak at the beginning of 2022 is caused mainly by the Omicron variant which saw its first case in late 2021.



The SLT plots for COVID-19 deaths are very similar to those from COVID cases, but there are a few important distinctions that are worth noting. Looking at the trend plot, there is a relatively steep, linear increase in deaths in 2020, and a steady decline after that. The trend for COVID cases did not see a decrease until 2022. This illustrates the strengths and weaknesses of the vaccinations, which again were introduced in late 2020. It is popularly believed that the vaccines would protect a person from contracting COVID-19 at all, when in reality they are more efficient in lessening the severity of symptoms, likely explaining the earlier decline in deaths than what we see with cases.

2.3.1 Summary of the Decomposition

The previous two images offered quite a few insights into the data sets, and it is important that we recap the main findings:

1) There is a general incline in both cases and deaths around the fall and winter months, which can be attributed to seasonal factors like weather, increased time spent indoors, and more frequent traveling
2) The trend for COVID related deaths begins to decline significantly earlier than the trend for COVID cases, largely because of the introduction of vaccinations in late 2020
3) From the residual plots, it's clear that much of the variation is explained by seasonality and periodic trends, and that the remaining noise is attributed to sudden outbreaks from new variants like Alpha, Delta, and Omicron

3 Insights

3.1 Data Insights

After analyzing the time series datasets regarding total nationwide COVID-19 related deaths and cases from January 2020 to April 2023, we have drawn several insights that help us understand the impacts the disease had on the United States. There were policies implemented by the government, seasonal fluctuations, and new variants that caused both rapid increases and decreases in the number of cases and deaths. One of the major trends that we saw was the consistent increase in cases and deaths during the colder winter months, which are positively correlated with time spent indoors, the number of people traveling, and the number of gatherings for the major winter holidays. In contrast to the winter months, our data analysis demonstrated a decrease in the summer months, also positively correlated with the time spent outdoors.

After further analysis of the data, we noticed that Pfizer, Moderna, and Johnston & Johnston's vaccine trials at the end of 2020 had a significant impact on the time series trends. Cases in 2022 remained stagnant for a period of time and eventually took a steep decrease, but we saw deaths declining far before that, as proof of the successes of the vaccine. While the vaccine did not prevent people from catching the disease, the vaccine prevented people from inhibiting the more severe symptoms that were typically associated with COVID-19. Section 2.2 of our analysis examined the cross correlation between cases and deaths, showing that spikes on the time series plot in deaths occurred about one month about spikes on the time series plot in cases. This seems to follow the pattern of progression that COVID-19 cases resulting in death take.

4 Conclusion

4.1 Limitations of Research

With any research and data analysis, there are limitations to take note of that may hinder the accuracy and benefit of results yielded. In this case, it is important to consider the fact that much of the data, especially later on in the time series, may have been slightly inaccurate since the frequency of COVID testing decreased significantly as time went on. This could mean that the number of cases nationwide are slightly lower than what is actually true, which could skew some of the findings. This could become especially relevant when looking at future forecasts of the virus, so perhaps more in-depth measures would need to be taken to get a sufficient result.

4.2 Discussion of Research

We began this analysis with a general visualization of the data sets. This gave us a preliminary understanding of how the two time series behaved and took note of certain points of interest such as the various peaks and valleys present in both. We were able to attribute these points to different factors such as government policy and CDC guidelines. We then used cross-correlation examinations to determine how the COVID-19 cases set can be used to predict how the COVID-related death plot will react given a certain period of time. This method can be quite helpful, particularly in the medical field, to draw trajectories for future viruses like the coronavirus and it can also give insight on how to efficiently combat the spread. Using a seasonal trends function, we were able to decompose both data sets into four descriptive plots which were used to determine the success that the preventative measures taken to combat COVID-19 had for the nation. Specifically, these plots made it clear that the vaccinations instituted during the late months of 2020 did indeed have a significant impact on the spread and severity of the virus. Overall, after visualizing, correlating, and decomposing our sets, we were able to achieve our research goal in attributing the various plot trends present in the data to worldly influences and assess the efficiency in doing so.

Works Cited

[1] “US COVID-19 Cases and Deaths by State.” *USAFacts*, 23 July 2023, usafacts.org/visualizations/coronavirus-covid-19-spread-map/.

[2] “CDC Museum Covid-19 Timeline.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 15 Mar. 2023, www.cdc.gov/museum/timeline/covid19.html.

[3] “Coronavirus Disease (Covid-19) Pandemic.” *World Health Organization*, World Health Organization, www.who.int/europe/emergencies/situations/covid-19. Accessed 10 Dec. 2024.

[4] “Do People Really Get Sick More Often during the Winter?” *Atlantic Health*, 7 Dec. 2023, ahs.atlantichealth.org/about-us/stay-connected/news/content-central/2023/sick-in-winter.html.

[5] “8.2 Cross Correlation Functions and Lagged Regressions: Stat 510.” *PennState: Statistics Online Courses*, The Pennsylvania State University, online.stat.psu.edu/stat510/lesson/8/8.2. Accessed 11 Dec. 2024.