Spectral Analysis of COVID-19 to Determine Seasonal Trends

Simon Lee
Baskin School of Engineering
siaulee@ucsc.edu

https://github.com/Simonlee711/Research/blob/master/Spectrum_Analysis/covidcase/analysis.py

Abstract

The Covid-19 pandemic has been a global issue for well over 2 years and counting, but recent reports suggest that it may be coming to an end. It is not to say that the virus will be swept under the rug, but health officials have claimed to treat it as a seasonal endemic. Therefore to test the accuracy of these claims, our research runs spectral analysis which is a method of analysis with respect to spectrums that erect in frequencies on 8 countries Covid-19 time-series data (https://github.com/datasets/covid-19/blob/main/data/time-series-19-covid-combined.csv). As part of our methodology, we wrote a python script that takes our time series data and converts it to our frequency representation through the use of an algorithm called the Fast Fourier Transform (FFT). We wish to take this data and make it useful to interpret. Possibly to be controversial, our data disagrees with health officials and we believe it is premature to claim an end to this pandem

1 Introduction

As of today, The Covid-19 pandemic has stretched out for over two-plus years. And the hot topic in the news as of late has been the bold claim that the COVID-19 "pandemic" is coming to an end. "We have seen now that this is likely to become a seasonal endemic disease here in the United States and really around the world" - Rochelle Walensky (Director of the CDC) [1]. This news headline is among many globally that are claiming a premature end to the pandemic and a start to an endemic. However how accurate are these claims that this is truly the direction in which this global pandemic is trending. Luckily with the accumulation of 2 years of daily cases and morality data, we can run mathematical analysis to distinguish our existing status and even make predictions based on this time series. So in this paper, we wish to explore the following question: Is there a seasonal trend in the COVID-19 pandemic? If so can we call this a seasonal endemic like the CDC suggests?

To answer our question we wish to explore the COVID-19 cases time series data of 8 countries (4 from the northern hemisphere & 4 from the southern hemisphere) and see whether there tends to be "Covid spikes" in specific seasons. We wish to look for these "spikes" through a time-domain approach as well as a frequency-domain approach (spectral analysis) that may show impulses (signals) within a specified period. To back up our points, we will generate two sets of graphs for each country, a weekly cases time-series graph as well as a Power Spectrum Density graph (Frequency Domain). So with a better idea of what's to be discussed we form the following hypothesis based on health officials. We believe that if mask mandates are being lifted, COVID-cases do indeed tend to be worse in the winter seasons and better in the spring and summer months.

2 Model & Methods

Before we get to our analysis, there is a level of understanding we need mathematically to understand what is going on. In the following subsections, we will be cruising through the concepts of the time vs.

frequency domain as well as the mathematics behind the Fast Fourier Transform, which is the algorithm we will be using to generate our data.

2.1 Spectral Analysis

What is *spectral analysis?* In the simplest terms, it's a method of observing spectrums that appear within frequencies. [2]. Especially when modeling worldwide pandemics, time-series data tends to show a periodic behavior. However, this may not be so obvious due to our data containing *noise.* Noise is often associated with entropy or uncertainty and in short, it means there is some corruption to the data [3]. Unfortunately, Covid-19 data is among some of the noisiest data and we need a better way to see patterns within our time series data. Therefore we will use spectral analysis, which is a method that helps us determine periodicities within our data. In order to do so, we first must understand the difference between the time and frequency domain.

The *time domain* is typically some dynamical system that generates an output within an evenly spaced amount of time. In terms of our data, we are specifically counting the weekly cases that were recorded. Next, we have the *frequency domain* which is an analytical space in which "signals" or "impulses" are conveyed in terms of frequencies. However, instead of measuring just some signal in frequencies, we will be taking this and be converting it to a periodic representation to see when these impulses are shot for proper analysis. The main advantage to observing our data in terms of the frequency domain was because Covid-19 data was rather noisy. So while we can observe Covid spikes in the time domain, we may not truly understand the trends of Covid-19 in this general approach. Especially since Covid cases are very subject to how much testing is done, whether the test is accurate (false positives, false negatives), as well as how cases are not reported on weekends, we see our data fluctuate substantially from week to week. Therefore using spectral analysis, we get the advantage of observing strong impulses that aren't so easily seen from just looking at our time series and we can use this method to see whether there is a seasonal effect with Covid-19.

2.2 Requisite Mathematical Background

Now that we are familiar with this concept of the time and frequency domain, we now will discuss the mathematics required to begin our analysis. In this subsection, we will discuss the sine equation, Fourier transforms, Fast Fourier Transforms (FFT), and the Power Spectrum Density (PSD). With varying levels of mathematical understanding, it is important to note that all these mathematical concepts being discussed build off of one another.

2.2.1 Sine Equation

Let's first begin by describing the basis of the Fourier transform which is the sine equation. The sine equation can be seen as the following:

$$y(t) = \lambda \sin(\omega t + \phi) \tag{1}$$

In the above Equation 1, we have our sine function. Within this sine function we have three components: the amplitude λ, phase ϕ, & frequency ω. Assuming that not everyone has seen this, we will briefly describe what each coefficient governs within this function. The amplitude λ is the maximum and minimum height that the sine wave hits from its zero bases (x-axis). The phase ϕ is what we call the time shift, and it is the left and right movements we can make within the function. And lastly, we have the frequency ω which is the number of complete cycles it gets through within a given period. In the context of our equation, ω is actually the angular frequency which is the frequency multiplied by 2π.

2.2.2 Fourier Transforms

With that, we can take a massive leap in difficulty and discuss Fourier Transforms which will be the basis of our algorithm the Fast Fourier Transform (FFT). Applying our knowledge acquired from time and frequency domains, a *Fourier Transform* is a mathematical transform that decomposes functions depending on time into functions depending on temporal frequency [4]. It can be expressed in the following way as a series of infinite sums of sinusoidal functions.

$$f(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty}(a_k \cos 2\pi kt + b_k \sin 2\pi kt) \tag{2}$$

This series representation of an infinite sum of sinusoidal functions can represent any mathematical function. In fact, the goal of a Fourier transform is to re-imagine the function as a series of sinusoidal functions. However, an issue that arises is how we solve for the Fourier coefficients a_k & b_k. Solving these coefficients can allow us to determine the shape of our signals. Therefore Fourier himself arrived at the following integral in order to solve these two coefficients.

$$\hat{f}(\xi) = \int_{\infty}^{\infty} f(x)e^{-2\pi ix}dx \tag{3}$$

Mathematically what's going on is that we are taking the integral of a function and producing a complex number. However, we once again arrive at an issue because our data is not a mathematical function but a data set of some fixed output with dependence on time. Therefore we must further take on more math into a specific type of Fourier Transform known as the *Discrete Fourier Transform (DFT).*

2.2.3 Discrete Fourier Transforms & Fast Fourier Transforms

The fundamental concept behind the Discrete Fourier Transforms is that it takes a data set instead of a function and transforms it into another data set that contains the Fourier coefficients. This new data set can be visualized as the following:

$$X_k = X_0, X_1, X_1, X_2, ...X_N \text{ such that} |X| = N \tag{4}$$

$$X_k = \sum_{n=0}^{N-1} x_n e^{\frac{-2\pi i(kn)}{N}} = \sum_{n=0}^{N-1} x_n[\cos(\frac{2\pi kn}{N}) - i\sin(\frac{2\pi kn}{N})] \tag{5}$$

where N is the total number of samples and n is the current sample. We also see the value k which is the current frequency within the boundaries $k \in [0, N-1]$ and x_n which is the value of the current sample. With all those parts we compute the X value which produces a complex number $(a + ib)$ computed by Equation 5.

Now that the mathematical foundation has been laid out, we need to discuss how we get these results numerically. To that, we introduce the *Fast Fourier Transform (FFT)* which is one the most prominent algorithms that computes the DFT's of any given time series data. This algorithm is a divide and conquer algorithm, where it divides up the signal into smaller signals, computes the DFT of the smaller signals, and joins them together. This completely changes the time complexity of such computations where a DFT is $O(n^2)$ while a FFT is $O(n \text{ log } n)$. This is why it is one of the most widely used and famous algorithms of the 20th century.

2.2.4 Power Spectrum Density

Last of all, we want to talk about the Power Spectrum Density (PSD). A power spectrum $S_{xx}(f)$ of a time series $x(t)$ describes the distribution of power into a frequency that shows up in the form of a signal/impulse. [5] It is a popular method in Fourier analysis, and we can use it to see physical signals that are shot up over a periodic range. The reason we care about the PSD at all is that it is visually easier to interpret compared to a DFT. Therefore in the context of our paper, we are looking to see what periodic signals (spikes) we can be seen within a range of time (seasons). The following can be computed with the below equation:

$$\text{PSD} = \left(\frac{1}{\omega * \text{len(covid data)}}\right) * abs(\text{FFT Countries Results})^2 \tag{6}$$

However, this still does not totally overcome our noise problem. Within our data, we will see that many impulses are actually shot up in our PSD graphs. So a popular trick in signal processing is to "denoise" our data by setting some threshold that will drown out those tiny signals. This makes a more clear and easier interpretation and it will be very much useful in analyzing the possible seasonal trends we might see in the Covid-19 data.

2.3 Data Preparation & Methods

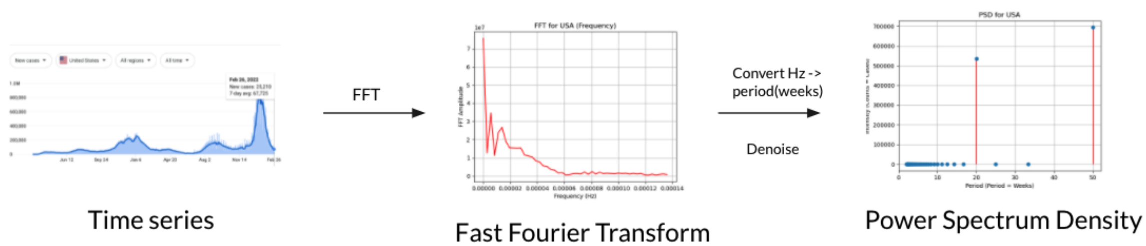

Figure 1: An overview of our Methodology to obtain our results

With that, we can begin addressing our original question on whether there is a seasonal trend in the COVID-19 pandemic. In order to see whether there is some seasonal effect, we must take into account the difference in seasons that occur naturally from the northern and southern hemispheres. Therefore we take a deeper look into 4 countries from each respective hemisphere: Argentina (Southern), Brazil (Southern), India (Northern), Indonesia (Southern), Russia (Northern), South Africa (Southern), United Kingdom (Northern), & The United States of America (Northern). The selection of these countries was motivated by mainly their population size as well as the countries that globally experienced the worst conditions during Covid. With our original hypothesis stating that winters are the worst seasons for infections, we want to see whether these claims are true within both hemispheres.

Luckily for us, we obtained the Covid-19 time-series data from John Hopkins University's Center for Systems Sciences and Engineering (https://github.com/datasets/covid-19/blob/main/data/time-series-19-covid-combined.csv), and they are the ones responsible for updating the daily Covid cases time series that you see on Google. We observed Covid cases data from March 5th, 2020 all the way up until February 7th, 2022. The reason for the delayed start was that a lot of countries did not receive a patient zero until the beginning of March. Now with access to a data set as well as a specified timeline, we begin our data preparation. First, we took our daily cases time series and converted it into a weekly time series. The reason is to mitigate some early noise within the data. So to do this we wrote a basic python script that would scrape the CSV files and store the dates as well as the case counts into separate lists. We repeat this process until we extract all the daily case counts for each country. Next, we look to convert our daily case counts to weekly and we can do this by simply taking the summation of the case numbers and storing this into another list. We reset our summation every time our index hits a modulus of 7.

The next step we must take is running the Fast Fourier Transform (FFT) algorithm on our weekly time series data to get into a frequency domain (Hz). One of Python's external libraries *SciPy* contains a series of numerical algorithms, and we intend to use Scipy's built-in function for the FFT to compute our Fourier coefficients. Our last step is now to get it into a PSD. In order to do so, we are going to convert our frequency to a period (weeks) as well as set a threshold to drown out noisy signals that may appear in the PSD diagram. Our python script as well as all our data is stored on Github and linked below the title.

3 Results

In this section we begin to analyze the trends of the northern and southern hemispheres.

3.1 Northern Hemisphere

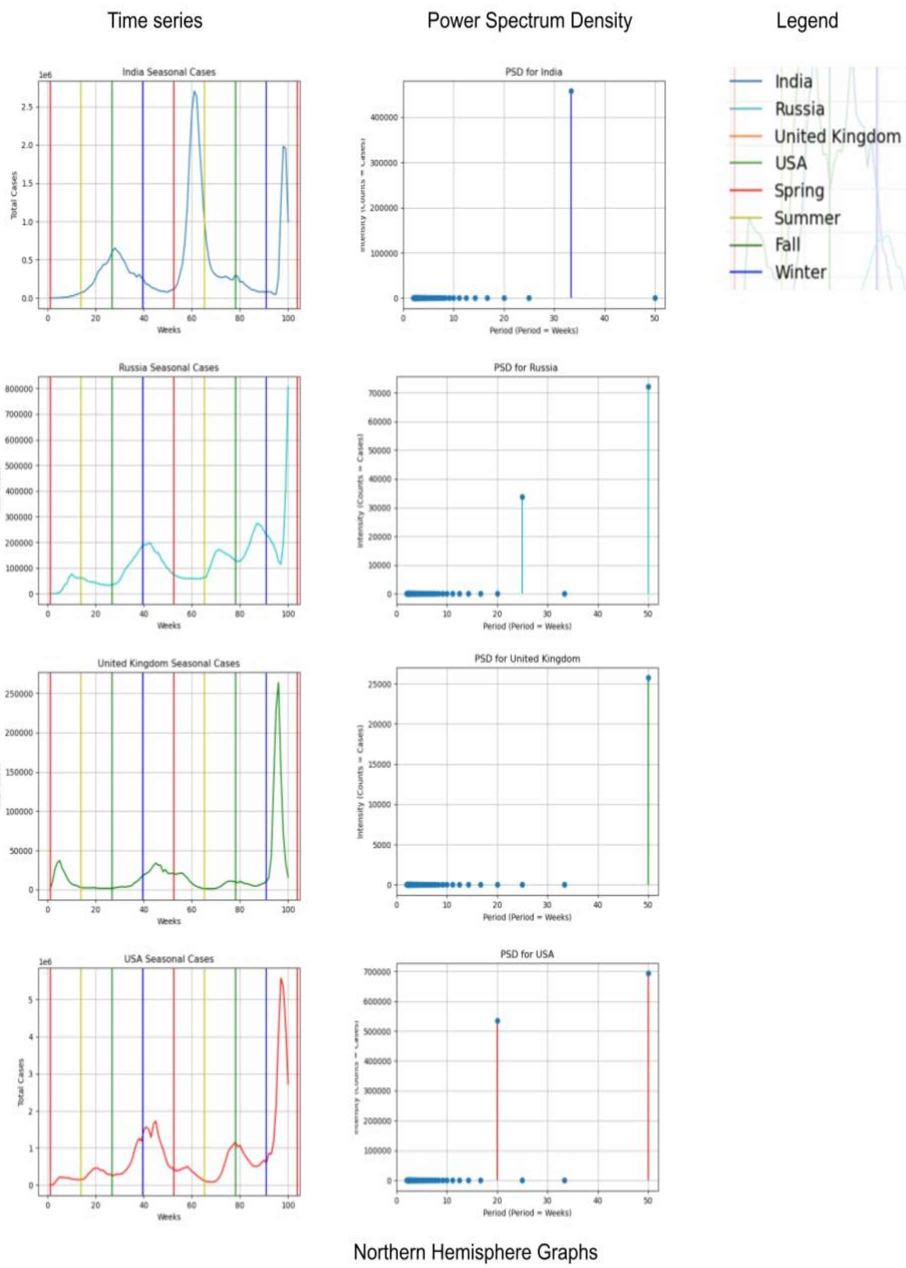

Figure 2: Pictured of the left are the time series graphs. Pictured on the right are the Power Spectrum Density Graphs.

In Figure 2, we see our plots of the time series (left) and our Power spectrum density plots (right). Before we begin to analyze the trends, let's first begin by defining seasons. Starting from March 1st to May 31st we define this as the spring season. From June 1st to August 31st we define this as the summer season. From September 1st to November 30th we define this as the fall season. And lastly, from December 1st to February 28th we define this as the winter season.

Let's begin pointing out patterns we see from a time-domain perspective. Within each country, we see varying levels of Covid Spikes, and some spikes are diminished due to the recent events of the Omicron and Delta Variants as shown in the most recent winter season. Therefore, countries like the United Kingdom and Russia still experienced Covid spikes but it is not amplified due to its most recent surge of infections. From a visual standpoint, we can see that Covid-19 has hit all the countries hard during the winter season. However, to our surprise, there also appears to be a "secondary" season in which Covid spikes. So if we look at our corresponding Power Spectrum Density graphs on the right of each countries time-series graphs we can see what impulses are shot up and when.

Before we move on to further analysis, we will briefly describe how to interpret these PSD graphs shown on the right-hand side. Let's look at the PSD of India which has an impulse that erects at around 30-35 weeks. This data is suggesting that there will be a Covid spike every 30-35 weeks. This trend appears to be relatively true if we compare it to the time-series graph where we see that roughly every 30 or so weeks, we do indeed get a Covid spike. Therefore what the Power spectrum density does is erect a clear signal that best captures our noisy data. We can also see that there are a series of other dots or impulses on the x-axis and those were impulses I decided to drown out because they were very insignificant to our analysis. So a lot of this early data suggests that we are getting very clear signals from each country.

However, we also want to point out the interesting observation that shows some countries having 2 impulses shot up on the PSD. This appears to be the case for Russia and USA for the northern hemisphere. In some cases, a periodicity of just one signal is not enough to capture the best representation of the whole time-series data. Therefore there are times when a second impulse will erect suggesting that there are actually two periodicities that we should be aware of in our analysis. We believe some of these impulses have to do with the recent surge of cases that have occurred in the weeks 90-100 period which is when the omicron variant was most prevalent. However, in order to best capture this data, our FFT suggests that these two impulses are something to keep in mind.

Now that there is a level of understanding on how to interpret these graphs, let's begin putting into context some early thoughts about whether season does play a role in Covid-19 infections. In most countries, the winter seasons do agree with the claims from the CDC and infection does tend to be worse based on high positive Covid case counts. However, as mentioned earlier there appears to be a secondary spike that occurs within other seasons. Though this varies from country to country, we want to put into perspective some of the reasons as to why this occurs. Over the spring of last year, India experienced some of its worst cases headlining News outlets saying that India is continuing to "shatter all types of Coronavirus records" [6]. The United States for a while faced a political war on whether beaches should be opened on Independence day resulting in a major spike in cases during the summer months. So with all these countries facing some sort of resentment from pandemic fatigue, it is safe to assume that most countries have experienced close to bi-annual spikes. Applying our data from our PSD graphs also suggests that this is true. Within our four countries that we are analyzing, a spectrum is shot up in all countries excluding the UK within a 25-35 week range which is roughly every 6-8 months. These early signs tell us that perhaps this isn't an endemic and it may be premature to say so.

3.1.1 Southern Hemisphere

Next, we examine the Southern hemisphere which has the opposite seasons from the Northern hemisphere. Therefore we define the seasons in the following. Starting from March 1st to May 31st we define this as the fall season. From June 1st to August 31st we define this as the winter season. From September 1st to November 30th we define this as the spring season. And lastly, from December 1st to February 28th we define this as the summer season. These changes are reflected by the vertical lines seen on the time-series graphs.

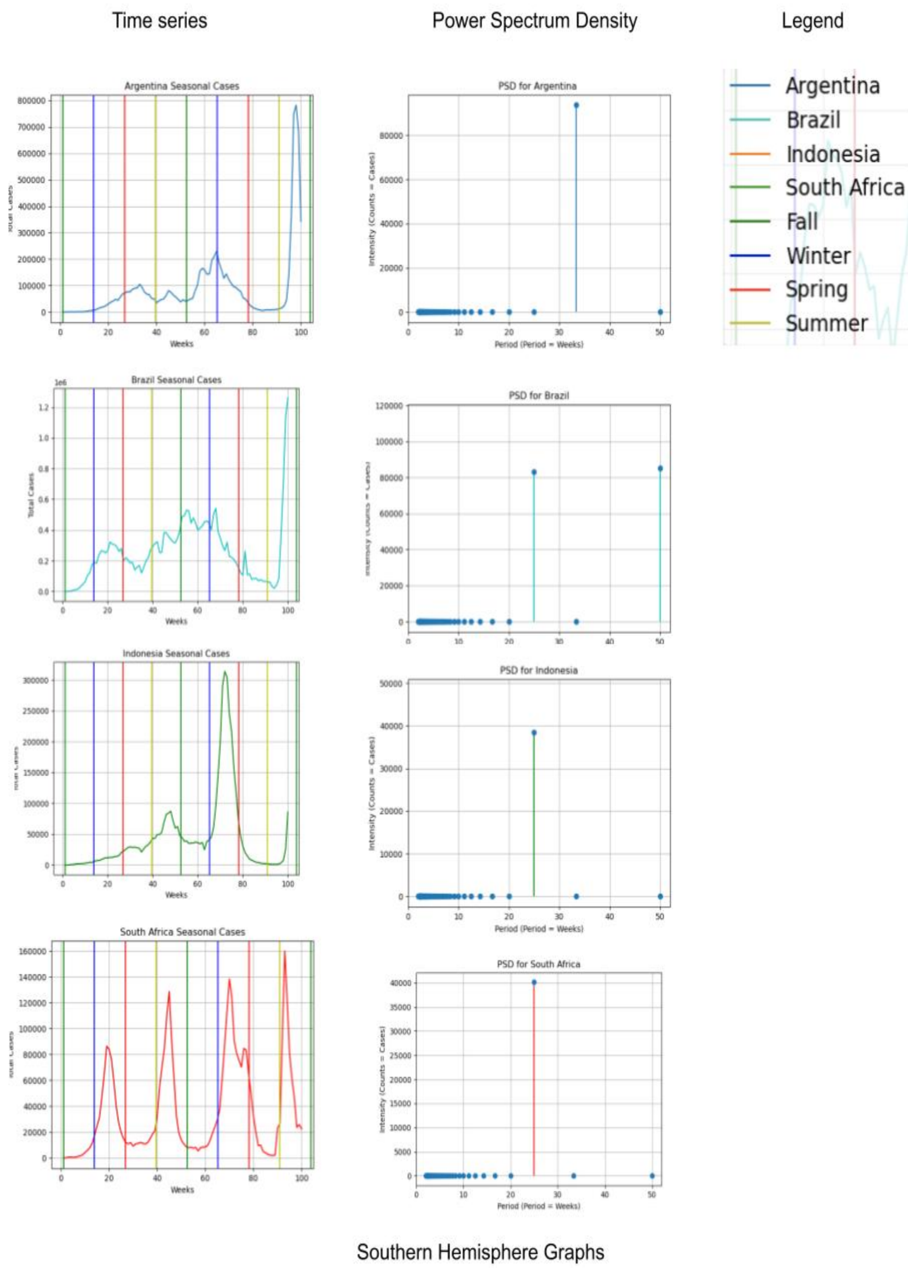

Figure 3: Pictured of the left are the time series graphs. Pictured on the right are the Power Spectrum Density Graphs.

With that, we can begin taking a look at Figure 3 and notice sort of an "inverse" behavior of the Northern hemisphere. Nearly every country observed minus Indonesia has experienced its worst Covid spike in this current summer season (northern hemisphere winter). This lines up with the timeline of the world when the Omicron variant in particular was very prevalent. Therefore it makes sense as to why the summer months were particularly worse for those given spikes. However what we can see is that Brazil, Indonesia, and South

Africa all have an impulse erecting at around 25 weeks. Though not entirely clear in the time-series graph, in these countries we see a "bi-annual" spike occurring in the winter and summer seasons. This data backs up our claim that it is not an endemic and we can clearly see this behavior occur across the world. In fact, South Africa has a beautiful depiction of this behavior as the spikes are very well defined, and then they drop significantly in the other seasons.

Meanwhile, we also have Argentina, which has a periodicity of around 30-35 weeks which is comparable to that of India from the northern hemisphere. This 33-week recurrence provides insight that although there isn't a particular season in which Covid cases spike, there is indeed a periodicity that shows that Covid does make reappearances in its timeline. This is rather important to note because a lot of governments around the world are thinking along the lines of CDC director Rochelle Walensky, and we believe it is very premature to begin claiming an end to the pandemic. In the southern hemisphere, we see a majority of impulses being shot up either within every 25 weeks or in the case of Argentina every 33 or so weeks. This means that roughly every 6 months (8 months for Argentina), there appear to be Covid spikes. Therefore our early data suggests that like the Northern hemisphere, the summer and winter months tend to be the most troublesome and so we disagree that this is an endemic.

4 Discussion

With all our data accumulated, we can truly address the question of whether there is a seasonal effect within the pandemic. And to that, we actually disagree with the claim that this is an endemic at all. Covid-19 is very much a case-by-case situation in every country and we can see that there were multiple Covid spikes that occured throughout any countries time series. While the majority of the countries experienced the worst conditions in the winter, it is also safe to say that this isn't completely necessarily true either. The southern hemisphere and the northern hemisphere experienced their worst peaks within the December 1st, 2022 to February 7, 2022 time frame. As alluded to earlier these were due to the omicron variant which had a highly transmissible quality. But this reassures the idea that there is no season in particular where Covid is worse. On top of that, I think it is safe to say that a lot of countries faced their worst-case counts in this most recent time frame. Therefore it remains skeptical as to why they are claiming an end to the pandemic.

Luckily the power spectrum density is meant for analysis but also predictions based on existing conditions. And our data points in the direction that most countries will likely face another spike within that 25-35 week range globally. Especially with the lifting of mask mandates across many counties, we believe this will yield another possible Covid spike. And with travel restrictions being nonexistent aside from Covid testing, I do not think it is far-fetched to believe that another variant is not possible within our near future. Because the winter and summer months have typically been the worse seasons for Covid infections, it is for these reasons we do not believe we are able to call this an endemic.

While we firmly believe that this is not an endemic but very much a pandemic still, we would like to mention some caveats of our research. It is safe to say that there are definitely covariates within our data that are not considered. The major one is the protection we get from vaccines. The quick development of the vaccine has been one of the biggest achievements of this global pandemic but also the reason this pandemic has been politicized. So with every new variant, there has been some level of concern that the vaccine does not protect against it, and a general rejection from a small population of the world population. However for the most part it is safe to say that the vaccine is very much effective against most strands of the virus. Therefore we would briefly like to mention that our research does not take into account of asymptomatic cases and people who have tested positive but have the vaccine. This is an important consideration because the effects of the virus are far less severe when we do have the vaccination. So putting that caveat and applying it to our data, it is safe to say that health officials are possibly considering the massive Omicron spike to not be an issue due to the lessened effects caused by the vaccine.

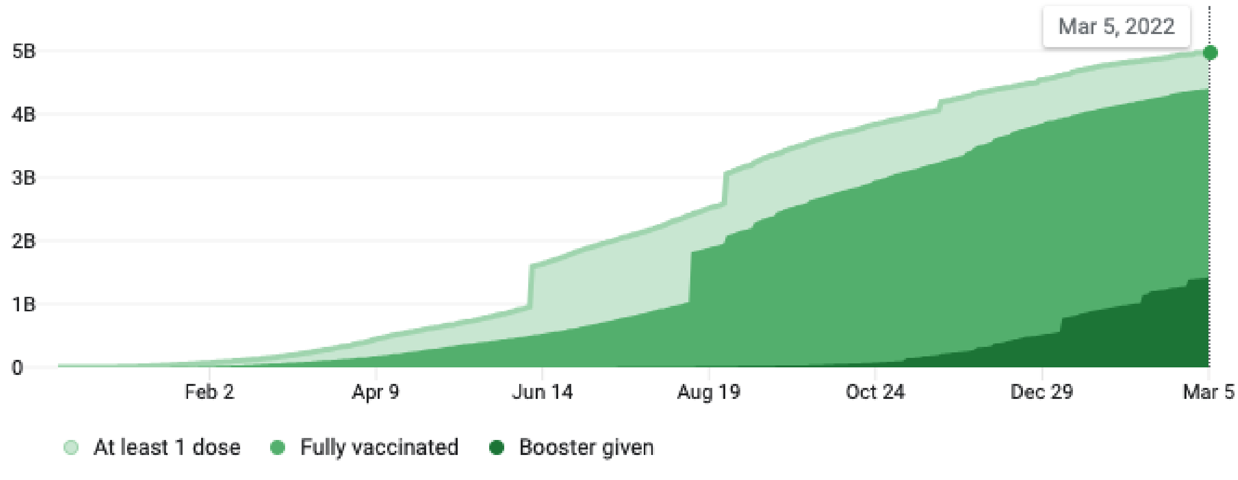

Figure 4: Vaccination numbers time series

While this is one substantial caveat within our data, it is hard to differentiate the vaccination on a country-to-country basis. But according to *Our World Data* [7] who have been in charge of keeping track of global vaccination rates, only a little over half the population is fully vaccinated against the virus at 54.6% as seen in Figure 4. So while vaccinations can substantially downplay the effects of the virus, it is also safe to believe that with a little less than half the population not vaccinated, to take the cautious approach and consider these Covid spikes as a dangerous situation. With all that we stand with the bold and possibly controversial claim that Covid-19 should not be sounded off and should most definitely remain a pandemic.

5 Conclusion

With that being said, we have decided to oppose the claims made by the CDC director, Rochelle Walensky. It is safe to say that this pandemic has an unfortunate political attachment to it and with everyone experiencing a level of pandemic-related fatigue, health officials most of all have faced a high degree of backlash. Though this may not be the motivation for claiming an end, it is safe to say that these words were partially misleading and that it is premature to make such claims. Therefore our research has made the strong remark to keep a level of caution and treat this as a global pandemic.

Our decision to disagree was backed up with spectral analysis and the impulses that were erected to show the general trend of the pandemic from a country by country basis. We also made sure to reassure our results by applying them to our time series graphs to make sure they were agreeing with one another. So through a time and frequency domain approach, we have made this decision to oppose the end of a pandemic. In the meantime, all we can really do is follow local community guidelines and keep up with the latest on possibly new variants that may appear. Though we are all very much tired of the pandemic it is safe to say that it is not over and possibly will continue for quite some time.

References

[1] https://www.washingtonpost.com/wellness/2022/01/20/what-does-endemic-mean/

[2] http://web.stanford.edu/class/earthsys214/notes/series.htmlspectral-analysis

[3] https://www.techtarget.com/searchbusinessanalytics/definition/noisy-data

[4] Bochner, S.; Chandrasekharan, K. (1949), Fourier Transforms, Princeton University Press.

[5] P Stoica R Moses (2005). "Spectral Analysis of Signals" (PDF).

[6] https://www.nytimes.com/live/2021/04/26/world/covid-vaccine-coronavirus-cases

[7] https://ourworldindata.org/covid-vaccinations?country=OWID_WRL