DMD Analysis on COVID-19 Time Series

Henry Charlton

June 9, 2022

1 Introduction

The spread of disease in a population has properties of interest that can be expressed using mathematical systems of equations, but often the complexity or availability of these underlying dynamics render their deduction difficult. In this analysis we seek to provide quantifiable insight from longitudinal COVID-19 case data using dynamic mode decomposition as a data driven analysis technique that is independent of knowledge of the underlying pandemic dynamics. Objective analysis like this can be beneficial to policy makers seeking to make decisions on how resources should be allocated and could also be used to extrapolate dynamics from our current pandemic to future incidences.

2 Methods

Dynamic mode decomposition (DMD) is a data driven technique that will be applied in this analysis to COVID-19 case time series by state and country. It generally can be used to obtain linear reduced order models for higher dimension complex systems, giving us access to the spatial temporal coherence structures of the data. This effectively comprises the dominant patterns of the data, and so is useful for longitudinal data and especially periodic longitudinal data like we have seen with the COVID-19 case time series. The advantage of DMD is that we can obtain quantified temporal information about a system without needing any knowledge of its underlying dynamics. This is particularly useful for our application. While there are many disease models that can approximate individual surges of COVID-19 cases such as the SIR model, few are able to provide useful insight regarding the longitudinal aspect taking into account multiple surges.

The calculation of DMD relies on linear algebra. First, the data is formatted into two matrices, X and X', of which figure 1 provides a visualisation. These matrices are very similar: Each column represents a time snapshot of the data and each row represents a different dimension of the data (in this analysis each row will be a geographic area). The only difference between the two matrices is that X includes all the data except for the last time snapshot (column), and X' includes all the data except for the first time snapshot (column). In this sense, X' is "one step ahead" of X. DMD then functions by finding the best fit linear operator to transform X into X'.

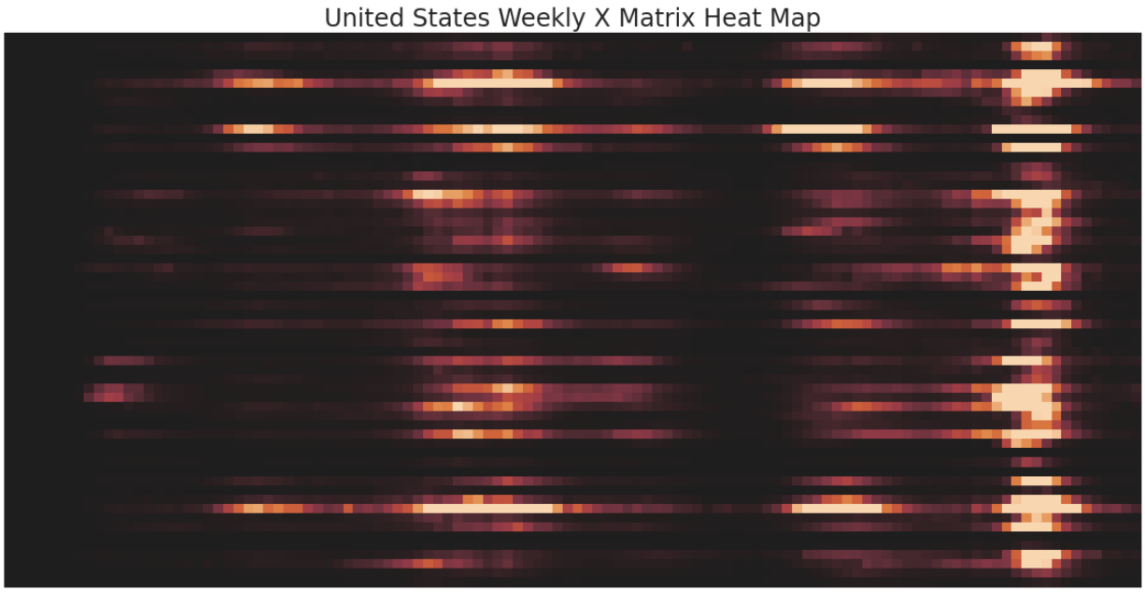

Figure 1: A heat map of the United States weekly data. Each row represents an individual state's time series and each column a snapshot of one week of data. This heat-map is a good way to visualize the X matrix that is used in DMD.

Data for this analysis was retrieved from the Johns Hopkins University (JHU) COVID-19 data repository for global data and from the United States CDC database for US data. The JHU data was delivered in the form of a daily time series for each country, whereas the CDC data required additional cleaning and collation to be formatted into a time series since each state and day together had a separate row. Formatting of this data was done in python using the pandas library. A caveat of the new case time series used is that some countries do not report cases (routinely) on weekends or in other irregular ways, which results in artificially noisy data. To mitigate this we collated a weekly time series instead of daily, which we calculated as a sum of every seven days compiled into a new time series. This eliminated a great deal of noise since most countries consistently report their cases at least once a week. Figures 1 and 2 illustrate that the weekly noise present when the time series is in a daily format is largely eliminated when reduced to a weekly resolution.

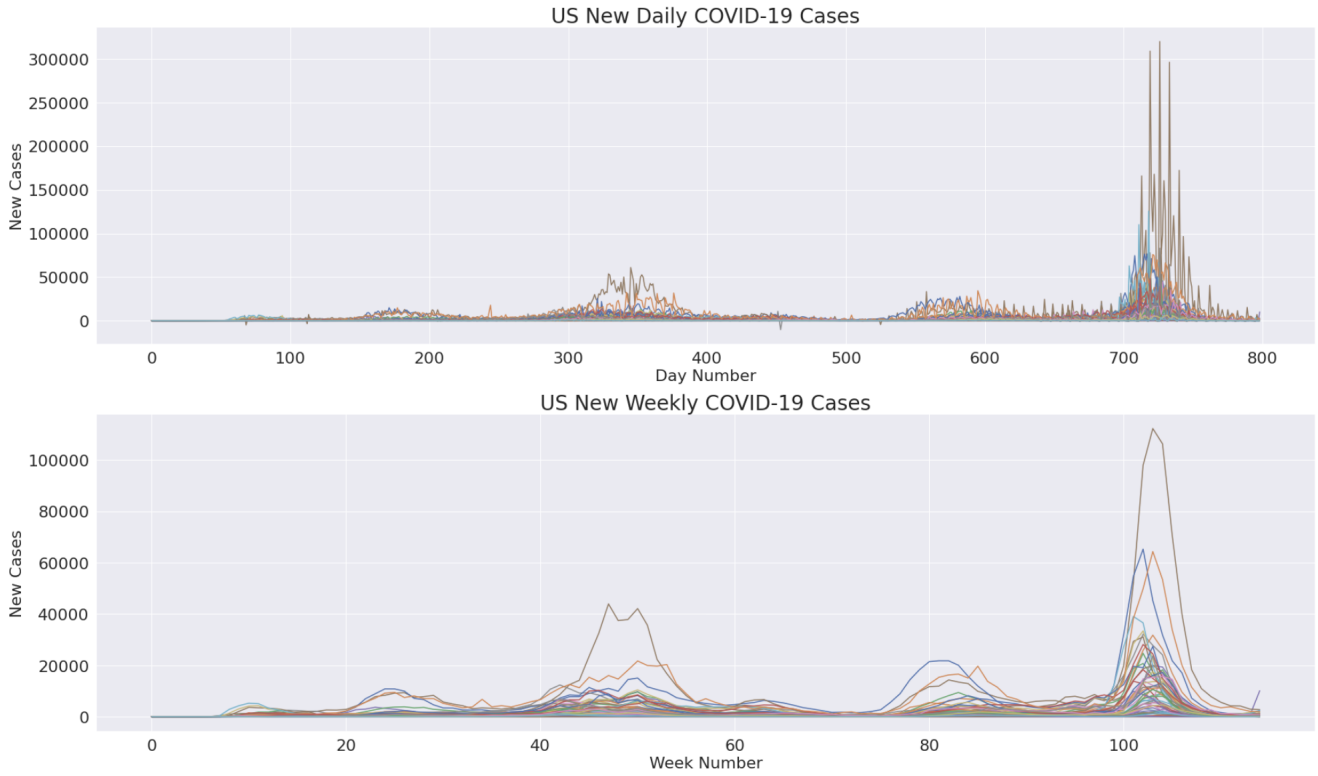

Figure 2: US daily and weekly COVID-19 case time series by state comparison. Note the reduction in noise.

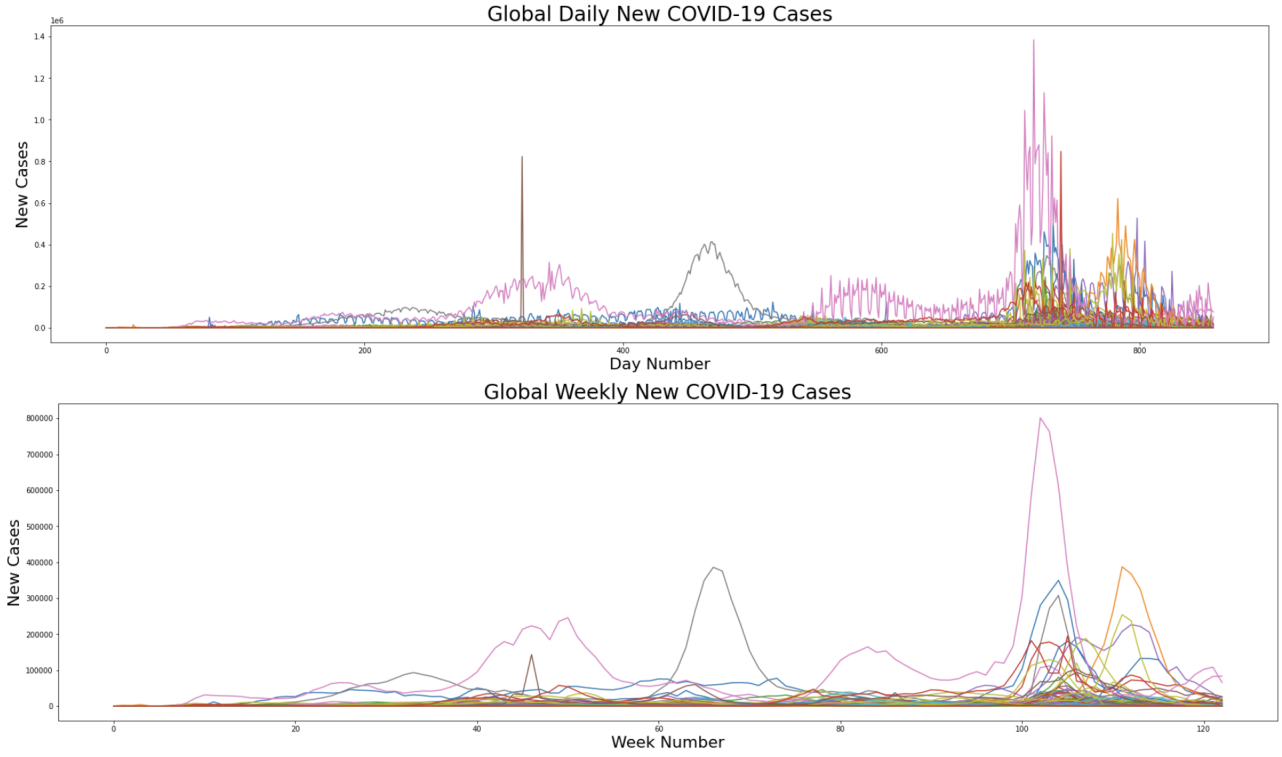

Figure 3: Global daily and weekly COVID-19 case time series country comparison. Note the reduction in noise.

The following set of equations represent the means by which the modes of DMD were calculated, based on the method originally developed by Peter Schmid[3]:

DMD aims to solve the following system for A:

$$X' = AX$$

In order to reduce this to a less computationally intensive form and analyze the dominant patterns of the data, singular value decomposition is used:

$$X = \tilde{U}\tilde{\Sigma}\tilde{V}^*$$

We can then approximate A as $\tilde{A}$:

$$A \approx \bar{A} = X'\tilde{V}\tilde{\Sigma}^{-1}\tilde{U}^*$$

We can then define Φ as the modes of the data:

$$\tilde{A} = \tilde{U}^* X'\tilde{V}\tilde{\Sigma}^{-1}$$
$$\Phi = X'\tilde{V}\tilde{\Sigma}^{-1}w$$

The following equation was used to calculate the frequency of oscillation for each mode.[4] Note that modes without complex eigenvalues do not oscillate.

$$freq_j = \frac{imag(\frac{\log(\lambda_j)}{\Delta_t})}{2\pi}$$

3 Results

The primary output of the DMD analysis is the modes which correspond to the columns of A, however an eigenvalue spectrum is a good way of analysing the dynamics that they correspond to. Figure 4 shows the eigenvalue spectrums for the US time series, comparing the daily and weekly versions. Recall that for a longitudinal system like this, the real components of eigenvalues can be interpreted as follows: Positive real components denote growing features, zero real components denote stable features, and negative real components denote a decaying feature. Furthermore, eigenvalues with a nonzero imaginary component are oscillating, and eigenvalues with a zero imaginary component are non-oscillating.

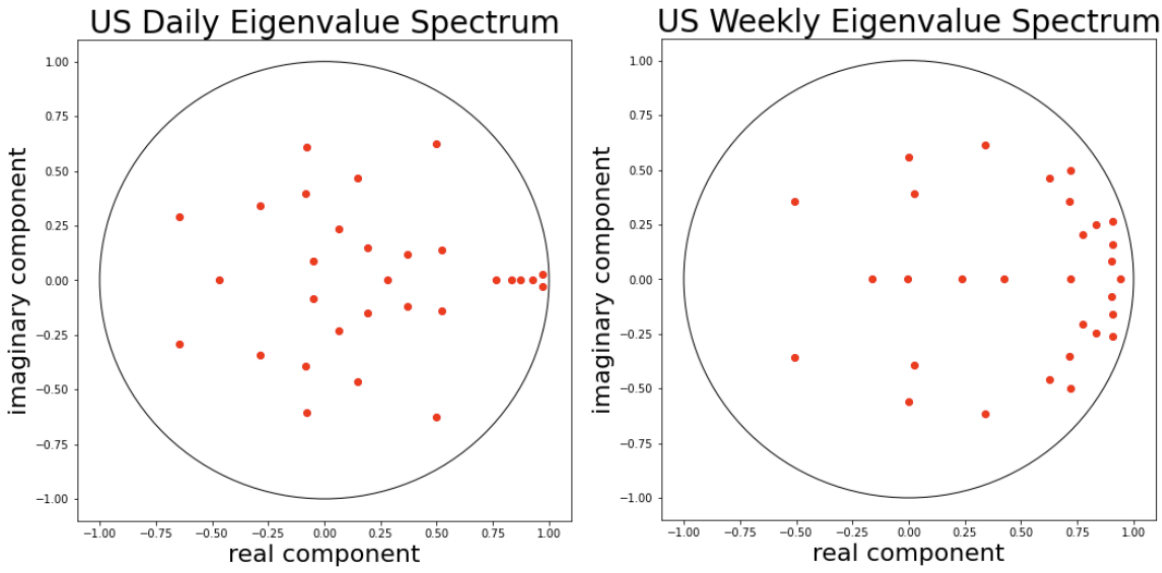

Figure 4: Eigenvalue spectrums for the US daily and weekly time series, respectively. Note the greater proportion of imaginary components in weekly data.

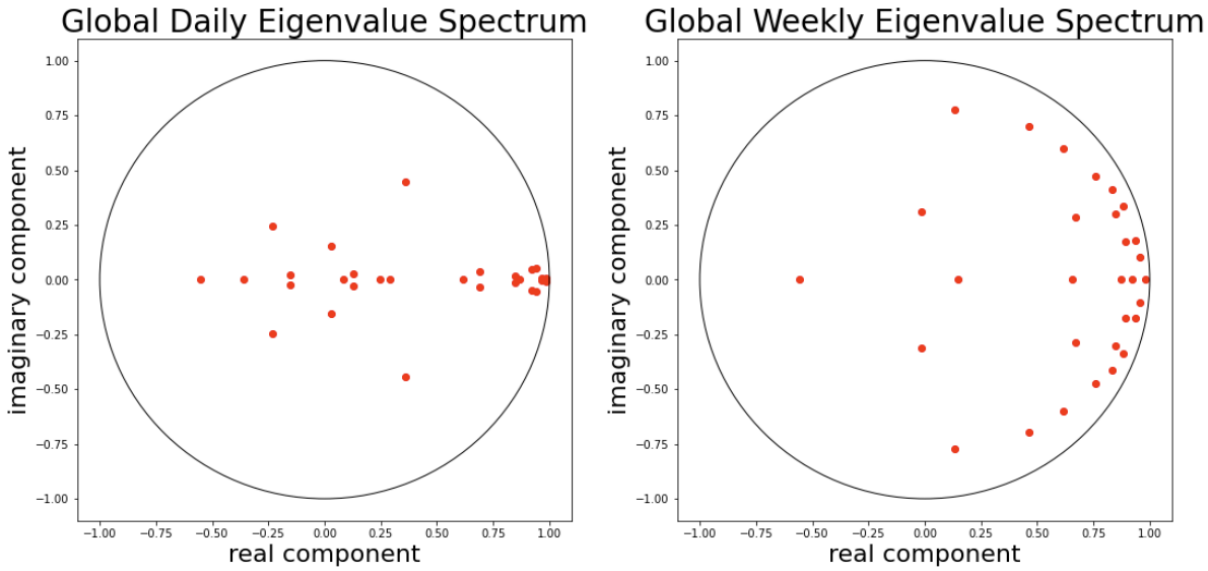

Figure 5: Eigenvalue spectrums for the Global daily and weekly time series, respectively. Note the greater proportion of imaginary components in weekly data.

By observing figures 4 and 5, we see that the change from daily to weekly data allowed more oscillating modes to be recovered by DMD, especially in the global data. This makes sense since the consistent and artificial weekly decline in cases that we saw in the daily case time series was acting as a sink for the pattern recognition of DMD. Another difference is that the eigenvalue spectrum of weekly data, both for the US and global time series, had consistently greater real components, indicating a greater ability of DMD to pick up on growing modes after the removal of artificial noise. All of this is to say that the weekly time series is a better choice for this analysis.

While there are no hard conclusions that can come from observing the eigenvalue spectrums, they allow us to get a general picture of the range of dynamics for the temporal patterns. For example, we can conclude that in the global time series, nearly every oscillating mode was growing. This is consistent with separate data that variants have become more infectious, especially in the case of the omicron variant.

In figure 6, the frequency of each mode is plotted against the Frobenius norm of that same mode. This allows us to select modes based on their dominance in the DMD analysis, higher norms corresponding to a more prominent mode. As such, we observe that the US analysis yielded dominant modes of frequency 0.22 years per oscillation and 0.4 years per oscillation, while the global analysis had more modes with large norms, ranging from 0.08 years per oscillation to 0.275 years per oscillation. This analysis allows us to get a sense of which frequencies, a type of temporal pattern, were strongest in the data.

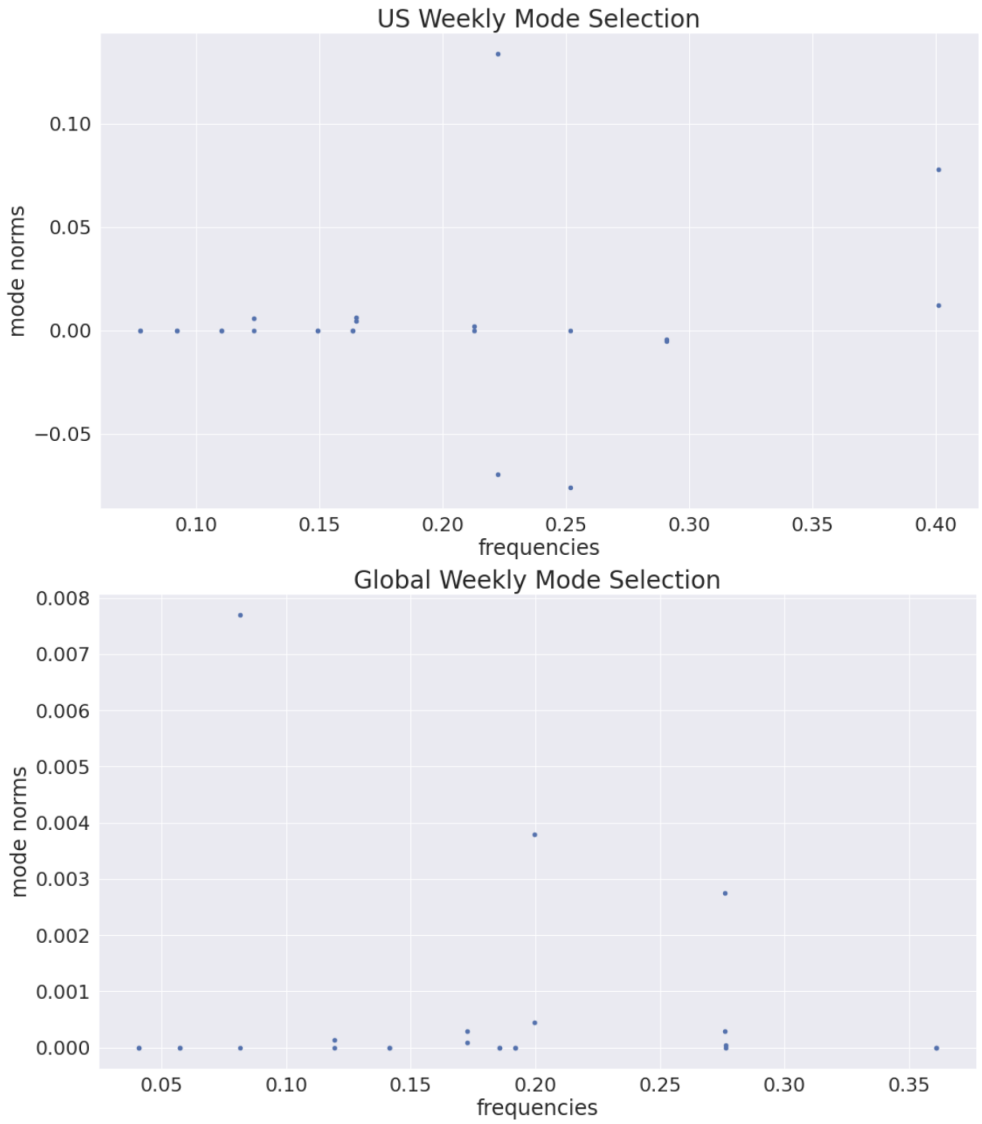

Figure 6: Mode selection for the US weekly data. Frequencies are in units of years per oscillation, so we note that the dominant frequency for US weekly modes is 0.22 years per oscillation and 0.08 years per oscillation for Global weekly data.

After finding a dominant frequency, figure 7 displays and example of the efficacy with which that mode captured the frequency of the data. Since it was done on a reduced form of the data using SVD, the frequency does not correspond to one specific country's temporal dynamics but is instead capturing the frequency dynamics of a larger part of the system.

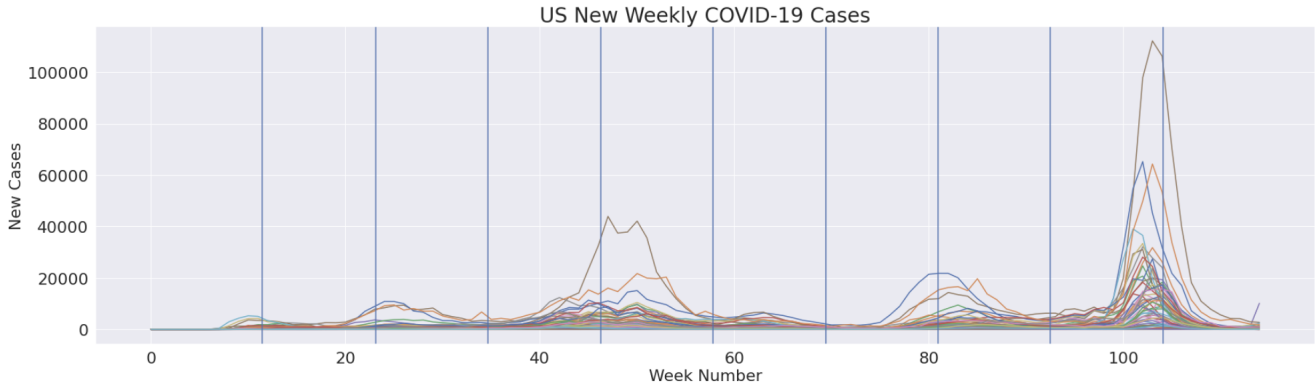

Figure 7: The dominant frequency of US weekly data overlaid onto its respective time series.

4 Discussion

This analysis provided quantifiable insight into the frequency and dominance of different temporal patterns of COVID-19 data, however there are several improvements that we recommend be implemented in a further iteration. As we noted in the methods sectioned, we summed the data by weeks to produce a less noisy time series at the cost of resolution. This is not ideal, and it is possible that it could be accomplished without the sacrifice in resolution. Exploration of the use of independent component analysis on the basis of two source signals would be our first step.

This analysis used raw case counts, whether by week or day, for the input into DMD. This method grants larger weight to areas with larger populations since they are likely to have inherently higher case counts. While this is not necessarily a erroneous path to take, it is worthwhile to re-analyse the data normalized for against the size of the total population. To accomplish this we would scale the data to cases per capita per week.

A similar issue to that of population normalization is of varying degrees of testing protocol integrity. Some countries and states test more than others and so even with a lower population may have higher case numbers. Simply normalizing for population would not mitigate this bias, so we propose that a future iteration of this project run the analysis on a time series of the proportion of positive tests out of total tests per time unit (the unit could be days or weeks).

Furthermore, we observed that almost all countries with oscillating modes had growth factors of varying degrees. We would like to investigate this phenomenon. By gathering more relevant metrics for each country such as the robustness of their lockdown or their testing protocols we could look for correlations between those factors and the magnitude of oscillatory growth. Significant findings in this would have implications for understanding how to cull the pandemic, especially if we can collate a large ensemble of variables to test the correlation with.

5 Conclusion

In this analysis we explored DMD as a method for gaining quantified insight into the temporal patterns of COVID-19 data. Using the visualisations produced we were able to broadly understand the varying dynamics of this data without any knowledge of the underlying system. Understanding these patterns and their frequency of occurrence can be useful to policy makers or to vaccine producers, and so we believe that further investigation and improvement of this analysis could be beneficial for anticipating the dynamics of future pandemics and coping with the current one.

References

[1] Centers for Disease Control and Prevention. United States COVID-19 Cases and Deaths by State over Time. Centers for Disease Control and Prevention. Retrieved April 18, 2022, from https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36

[2] COVID-19 Tracking. Johns Hopkins Coronavirus Resource Center. Retrieved May 1, 2022, from https://coronavirus.jhu.edu/data

[3] Peter Schmid. Dynamic mode decomposition of numerical and experimental data. Jour- nal of Fluid Mechanics, Cambridge University Press (CUP), 2010, 656 (August), pp.5-28. 10.1017/s0022112010001217 . hal-01020654

[4] Proctor, J. L., & Eckhoff, P. A. (2015). Discovering dynamic patterns from infectious disease data using dynamic mode decomposition. International health, 7(2), 139–145. https://doi.org/10.1093/inthealth/ihv009