Assessing the similarity of countries' COVID-19 experiences via time-series clustering

Mark Raney
Denison University – Data Analytics
Dr. Zhe Wang

Introduction

Through the past two years, the COVID-19 virus and the ensuing pandemic inspired staunch politization regarding a nation's efforts to minimize the number of new positive cases.

Now, over two years past the WHO's initial declaration of a global pandemic, we have public access to historical data detailing of new positive cases for most countries around the globe, allowing insight into the effectiveness of the actions of national governments relative to one another.

Understanding this, the purpose of this project is to construct a quantitative assessment of similarity in COVID-19 experiences between countries via clustering methods, a "hands off" approach that searches for patterns inherent within data to make these insights. Doing so entails carefully examining the quality of the data used, applying it in a manner that both produces a meaningful result and accurately represents the patterns inherent to the data, and finally assessing the viability of our approach.

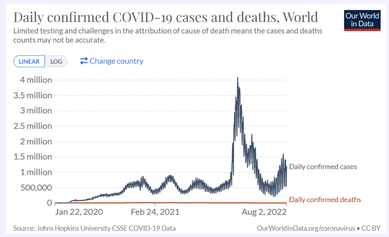

Figure 1. Example of a COVID-19 time series, demonstrating the daily confirmed cases reported anywhere in the world. Data via Our World in Data

Methods

To conduct our research, we utilize Our World in Data's (OWID) publicly available dataset that contains the daily confirmed cases of every country.

The first decision to be made is defining similarity regarding a COVID-19 experience. To accomplish this, we reference the overarching goal of this project - creating a quantitative metric that allows for meaningful assessment of the effectiveness of government intervention in slowing the spread of the virus. In practice, this takes shape as minimizing the heights and frequency of peaks.

To quantify this, we measure the distance between two time series as their Euclidean distance. In this, we align two time series, and for all aligned dates, calculate the difference in values between the two and take the average of them all. To account for varying first exposure dates between countries, the dates need not be perfectly aligned.

Methods - Continued

Next, we assess the quality of our data through examining structural issues and limitations in collection. For structural issues such as missing values, we perform imputation to estimate and insert the expected value. Furthermore, we smooth the data by setting each day's value to be the two-week rolling average to overcome disparity in reporting and non-reporting days. Regarding limitations in data collection, we recognize that varying testing availability between countries inevitably skews the data. To work around this limitation, in addition to the raw data we also utilize locally normalized time series (meaning for each time series, the maximum value is 1 and the minimum is 0) and ranked time series (each day is assigned its position in the series if sorted in ascending order).

Using these three sets of time series, we begin clustering. In the scope of this project, we found that a hierarchical agglomerative approach (A method of repeatedly combining the two most similar items until a set number remain) best fit the type of result we are looking to achieve. To quantify similarity between clusters of time series, we utilize average linkage (the average difference of every pairwise combination of time series between two groups) as our error metric.

Next, we identify the optimal number of clusters to use. To do this we combine two methods, the "elbow point" and silhouette coefficient. The elbow point method involves utilizing an interesting characteristic of our average linkage error metric - the fact that it is guaranteed to decrease as the number of clusters increases. Knowing this, we do not necessarily want to take the minimum error value, but rather the point at which the error value no longer meaningfully decreases as the number of clusters increases. Using this, we identify the optimal number of clusters as the point where the silhouette coefficient, a measurement of how well samples fit within their clusters and are distinct from other clusters, is maximized and is not too far from the elbow point.

Finally, with an optimal clustering generated, we perform early data analysis to identify similarities within clusters that may be correlated with their shared experience.

Results

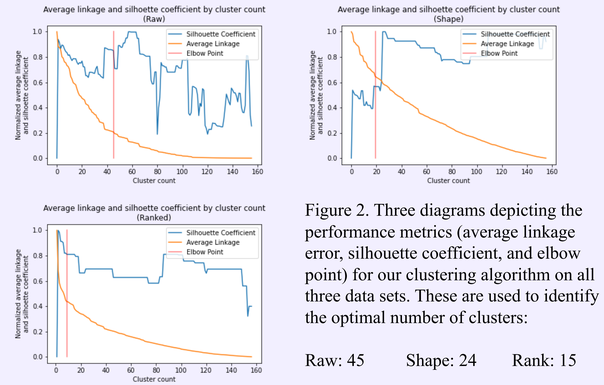

Figure 2. Three diagrams depicting the performance metrics (average linkage error, silhouette coefficient, and elbow point) for our clustering algorithm on all three data sets. These are used to identify the optimal number of clusters:

Raw: 45 Shape: 24 Rank: 15

Of these, the ranked time series presented quality results with a high silhouette coefficient and relatively low average linkage error, thus analysis is conducted on the results of this clustering, seen in figure 3.

Examining figure 3, it is evident that adjacent countries are likely to belong to the same cluster. To assess geography's influence on these results, we calculated the between cluster variance to make up 91% of the total variance between country capitals, suggesting adjacent countries are highly influential in a countries COVID-19 experience.

Finally, two separate ANOVA tests suggest that for at least one cluster, both GDP per capita and population density influenced the cluster's unique COVID-19 experience in comparison to the other clusters.

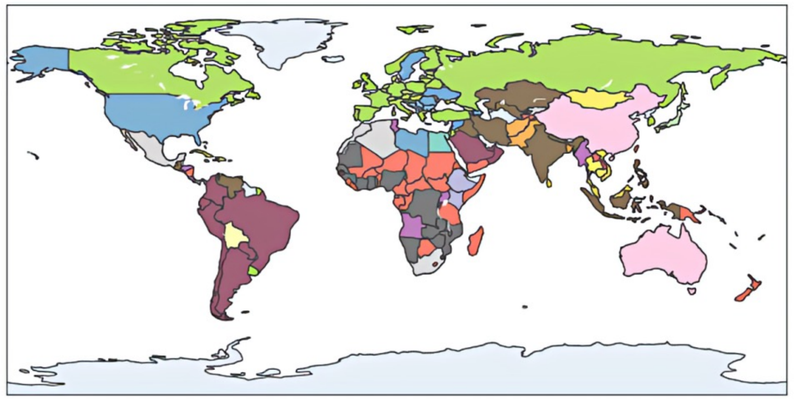

Figure 3. Map of the world with each country colored according to the cluster that it belongs to. This specific clustering is derived from the ranked time series data set. Our model identified fifteen clusters as the optimal cluster count.

Discussion

These findings suggest that while political action for the sake of mitigation may certainly have been influential in shaping a countries' COVID-19 experience, we cannot ignore the influence of external factors such as a countries' geography, population density, and population wealth.

GDP per capita for example likely influences the spread of the virus within a country for a multitude of reasons. First, it may be used as a proxy variable for mask, hand sanitizer, and other general health product availability; a country with a higher GDP per capita likely has more access to physical resources that reduce the likelihood of spreading the virus. Furthermore, GDP per capita may also act as a proxy variable for the type of work that the population does, namely differentiating between majority blue-collar or majority white-collar work. We might expect this to have a meaningful influence on a countries' COVID-19 experience through determining the viability of a transition towards working from home.

As it pertains to population density, we understand that this factor motivates more people encountering one another, providing more opportunity to spread the virus as population density increases. Similarly, a countries' geography with respect to adjacent countries would expectantly influence its experience, as certain borders are simple to cross and motivate interaction among countries' populations.

It is important to recognize that these factors are not necessarily unrelated to the actions of these national governments; GDP per capita might denote the viability of a work-from-home mandate, and it is possible that the actions of one country is influenced by the actions of its neighbors. With this, we are provided an area to focus initial political analysis into through examining the outliers within clusters according to these factors. For example, we might examine why Bolivia belongs to a different cluster from its neighbors, who all belong to the same cluster. Conversely, we might examine why distant countries such as Mexico, Algeria, and South Africa all belong to the same cluster.

Acknowledgements

This project was funded by the Lauri & David Hodgson Faculty Support Endowment at Denison University. Without them this would not have been possible!

Works Cited

Liao, T. Warren. "Clustering of time series data—a survey." *Pattern recognition* 38.11 (2005): 1857-1874.

Satopaa, Ville, et al. "Finding a" kneedle" in a haystack: Detecting knee points in system behavior." *2011 31st international conference on distributed computing systems workshops*. IEEE, 2011.

Rousseeuw, Peter J. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." *Journal of computational and applied mathematics* 20 (1987): 53-65.

Van Wijk, Jarke J., and Edward R. Van Selow. "Cluster and calendar-based visualization of time series data." *Proceedings 1999 IEEE Symposium on Information Visualization (InfoVis' 99)*. IEEE, 1999.