JIS Journal of Interdisciplinary Sciences
Volume 6, Issue 2; 53-76
November, 2022. © The Author(s)
ISSN: 2594-3405

Comparison of the Spread of Novel Coronavirus: Topological Data Analysis of 13 Countries

Shaun V. Ault* and Jia Lu
Valdosta State University, USA
svault@valdosta.edu*

Abstract: *Topological Data Analysis (TDA) is a recent rising method that provides new topological and geometric tools that can detect non-linear features, such as loops, in multidimensional data. As of now, most TDA studies are related to the biological structure of the SARS-COVID-2 virus and there is no literature on TDA application with country-level COVID-19 data. Thus, our study aims to fill the gap by applying this novel method to find data patterns of COVID-19 spreads in selected thirteen representative countries on six continents of the world and compare results among them. Briefly, TDA methods are useful for determining "features" in point-clouds, including clusters and loops. Furthermore, quantifiable differences in features of the data sets of different countries can suggest differences in public health policy among those countries. Our results suggest TDA can be a useful initial data tool to search for anomalies, which can then lead to a more comprehensive analysis combined with other techniques. Using TDA, we were able to identify three major groups of countries based on their virus data patterns. Australia, India, South Korea, and Taiwan are very similar, while Great Britain, Peru, and France have very different patterns from those of other countries. Next, the death-to-case ratio and death per million among countries were investigated. We also examined in detail the public policy and other reasons behind the similarities and differences of the TDA results and suggested possible successful public policies at national levels for a future pandemic.*

Keywords: COVID-19; topological data analysis; bottleneck distance; public health policy

Introduction

The Novel Coronavirus (COVID-19) pandemic has affected the world significantly since its first known case in Wuhan, China in the fall of 2019. As of May 2022, it has infected over 525 million people and killed over 6.2 million around the globe. Countries rushed to contain the virus, but policies implemented to combat COVID-19 differ from country to country and from region to region, and their results are very different as well. In order to better understand which countries' policies are the most effective at combating the spread of COVID-19, we need to understand the virus data pattern among the counties. In this research, we selected 13 countries in 6 continents that were deeply impacted by the coronavirus and

This work is licensed under a Creative Commons Attribution 4.0 International License

compared virus spreads with the novel application of Topological Data Analysis (TDA) for the case and death data of the outbreaks from January of 2020 to September of 2021.

TDA is a recent rising method that provides new topological and geometric tools to conclude relevant features for multidimensional data (Chazal and Michel, 2021). Regarding the current pandemic, most TDA studies are related to the biological structure of the SARS-COVID-2 virus. Chung and Ombao (2021) applied TDA to the topological characterization of the protein structures of the COVID-19 virus in order to demonstrate that there are topological changes during the conformational change of spike proteins. The study applied a persistent homology theoretical framework (Polterovich *et al*., (2020); Oudot (2015), which is the main tool used in topological data analysis. Mukherjee *et al.*, (2021) applied persistent homology to describe the shape and structure of the data representing immune cells in healthy donors and COVID-19 patients and conclude how these groups differ despite the noise and other variations in the data. Results are novel in their ability to capture the shape and structure of cytometry data, something not described by other analyses.

A study by Pérez-Moraga *et al.*, (2020) was the first time that a TDA-based strategy has been used to compare a massive amount of protein structures with the final objective of performing drug repurposing. Not many papers published are related to the pandemic outbreak's patterns through time. Chen and Volić (2021) applied this method, specifically the Mapper algorithm, to the U.S. COVID-19 data. The resulting graphs provide visualizations of the pandemic that are more complete than other more standard methods. The visualization allows for easy comparisons of the features of the pandemic across time and space and translates a variety of geometric features of the data created from geographic information, time progression, and confirmed number of COVID-19 cases. The resulting graphs reflect the spread of the pandemic across the U.S. and display hot spots and growth rates.

In addition, there is no literature on the TDA of country-level COVID-19 data and the comparisons among them. Thus, it is important for our study to apply this novel method to find patterns of COVID-19 spreads in representative countries in different parts of the world to identify patterns and search for possible reasons behind them, so better policies can be suggested for a future pandemic. It should be noted that we are using TDA in a different way than the above-mentioned papers. In our work, we use TDA methods (persistence homology and bottleneck distances) to discuss the similarities and differences among the abstract data sets arising from various countries. In this way, distinct responses to COVID-19 may be discovered and then further analysis may be applied in a more focused manner in future work. This paper begins with a TDA analysis and then explains topological similarities and differences in the data sets by way of policy decisions at the country level.

Our objective is to find the similarities and differences in terms of the virus data patterns in the studied countries so we can identify the possible reasons behind them and suggest better public policies for a future pandemic. Thirteen major countries from 6 continents (Asia, Africa, Europe, Oceania, North, and South America) are selected as our study area because of the importance, diversity, and representativeness of the countries in their virus outbreaks, as

This work is licensed under a Creative Commons Attribution 4.0 International License

well as their related public health policies. The countries included are India, South Korea, and Taiwan in Asia, South Africa in Africa, Great Britain, France, Germany, Italy, and Sweden from Europe, Australia from Oceania, the US from North America, as well as Brazil and Peru from South America. These countries include seven of the top eleven countries currently with the most cases, and India and South Africa are also the birthplaces for the two most infectious variants – Delta and Omicron. Our study also includes countries with very low cases per million (Taiwan, South Korea, and Australia). Taiwan and South Korea were generally considered successful in early control of the virus.

Our research questions are as follows:

1. What patterns were present with COVID-19 spread throughout these 13 countries?
2. What are the possible reasons behind the observed patterns?
3. Which related policies did countries use to control the spread and what were their effects on caseload and deaths?
4. What are the future public health policy recommendations?

Our time frame was the beginning of February 2020 to the end of September 2021. The data source was from Ourworldindata.org (Ritchie, *et al.* 2020) which is considered reliable in general.

Methods

First, we pulled data about COVID-19 from thirteen countries, Australia ('AUS'), Brazil ('BRA'), France ('FRA'), Germany ('DEU'), Great Britain ('GBR'), India ('IND'), Italy ('ITA'), Peru ('PER'), South Africa ('ZAF'), South Korea ('KOR'), Sweden ('SWE'), Taiwan ('TWN'), and the United States ('USA'), from Our World in Data (Ritchie, *et al*. 2020) up to December 2021. The specific countries were chosen for analysis based on the abundance of data available (there was about 600 days' worth of data for each country in this list), as well as to represent a broad and diverse selection of nations from around the world, whose COVID-19 policies were varied. Descriptive statistics of the raw data are provided in Figure 19, but our analysis is based on techniques of TDA, which further processes the data as described below.

Using Python (Van Rossum and Drake, 1995) to extract the relevant data from the database, we isolated the columns associated with *new (daily) cases* and *new (daily) deaths*. For each country, we processed the data as follows. First, the new deaths column was shifted 14 days earlier. This was done to account for the fact that when death does occur from COVID-19, it does so on average about two weeks from the initial diagnosis (depending on the region), as noted by Jin (2021) or Harrison *et al*., (2020). Next, each column was smoothed using a seven-day moving average. We opted to use seven days in order to reduce variance in reporting methods (e.g. some countries did not report data on weekends or holidays). Finally, the data were normalized by the population of each country to cases, resp. deaths, per million in order to make better comparisons among them.

This work is licensed under a Creative Commons Attribution 4.0 International License

The result of above-mentioned processing resulted in two-dimensional data sets of about 600 points associated to each country. We first produced scatterplots in which the *x*-axis represents new cases (scaled, smoothed) and the *y*-axis represents new deaths (scaled, smoothed). However, the ultimate goal was to use topological data analysis (TDA) techniques. TDA may be used to determine whether there are any non-linear *features* in a data set (point cloud). These features may then be used to distinguish data sets that are topologically distinct. While we are not the first to employ TDA in the study of COVID-19 statistics, we have done so in a novel way: to detect broad similarities and differences in various responses to the pandemic by country and then use these results to interpret the effects of national policies. In particular, we posit that different policy decisions may affect the shape of the data set that pairs COVID-19 daily cases to subsequent daily deaths; and TDA is precisely the tool that can easily detect the shape of multidimensional data.

Using the Python package Ripser (Tralie *et al.*, 2018), we produced a persistence diagram corresponding to the scatterplot of each country's data. Following the methodology of Chazal and Michel (2021), suppose we are given a set of *n*-dimensional data points, $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \ldots\}$, and a metric *d* on *X*. For our purposes, we use the standard Euclidean metric:

$$d(\boldsymbol{x_1}, \boldsymbol{x_2}) = \sqrt{\sum_{i=1}^{n} [(\boldsymbol{x_1})_i - (\boldsymbol{x_2})_i]^2}$$

Next, for any given real number $a \geq 0$, build the Vietoris-Rips complex, $\text{Rips}_a(X)$. The details of this construction fall outside the scope of this paper; however, one may think of joining points that are within distance *a* of one another into simplices, and then analyzing the resulting simplicial complex using techniques of algebraic topology; in particular, we may compute the homology of $\text{Rips}_a(X)$ (see Figure 1).

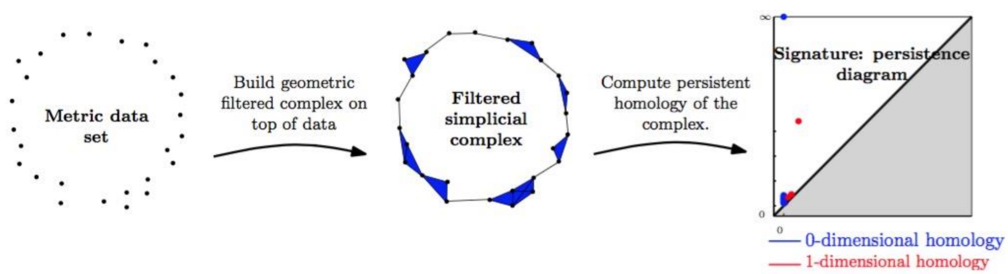

Figure 1: Illustration of Typical Workflow in TDA. *(Image courtesy of Wikimedia Commons, attributed to Larrysong. File:Illustration of Typical Workflow in TDA.jpeg.)*

The 0-dimensional homology essentially locates clusters at various levels, but we chose to focus on 1-dimensional homology instead, that is, the so-called Betti-1 numbers, or β_1. Briefly, β_1 counts the number of independent geometric cycles (loops) in multidimensional data, which is also called degree-1 persistence homology. However, the noisier the data, the

This work is licensed under a Creative Commons Attribution 4.0 International License

harder it may be to determine distinct cycles in it. The term persistence refers to the strength of the evidence for the existence of these topological features. Figure 1, for example, shows a data set with a strong persistence of a single cycle ($\beta_1 = 1$). The associated persistence diagram records these topological features in a concise way. The further away a dot is from the diagonal, the more prominent (persistent) that feature is. Thus, in the rightmost diagram of Figure 1, there is a single red dot off of the diagonal representing the strong signature of $\beta_1 = 1$.

Two data sets may be compared topologically via the so-called bottleneck distance between their persistence diagrams. Essentially, the more features of significant persistence that two data sets share, the closer they will be with respect to their bottleneck distance. Suppose *X*, *Y* are two data sets, and let D_X, D_Y, respectively, be their persistence diagrams. Then the bottleneck distance between the diagrams is given by:

$$d_b(D_X, D_Y) = \text{infimum}_M \max_{(\boldsymbol{x},\boldsymbol{y})\in M} ||\boldsymbol{x} - \boldsymbol{y}||_\infty$$

Here, *M* is a matching of points of D_X with D_Y (in which some points may be "matched" to the diagonal, meaning that they are effectively discarded), and $||\cdot||_\infty$ is the max norm (or uniform norm). The infimum is taken over all possible matchings *M*.

A rigorous and detailed explanation of TDA methods, including persistence homology and bottleneck distance, is beyond the scope of this paper; however, the authors recommend the excellent introduction by Chazal and Michel (2021) to the interested reader.

Analysis and Results

We ran a set of comparisons among the 13 countries, computing the bottleneck distances between each pair of data sets. A relatively lower bottleneck distance (e.g. 0.00--1.00 in this study) between two data sets indicates that the sets are topologically similar. This means that the data sets (as point clouds) share enough features of similar size and strength (i.e. persistence) that they are truly representative of similar statistics. A larger bottleneck distance (e.g. 1.01 or greater) between two data sets indicates that one has major features (loops in the data) that cannot easily be matched to the other. Indeed, the larger distances between the data sets of the two countries suggest that further analysis is needed to tease out the underlying reasons.

Figure 2 shows the matrix of bottleneck distances between the pairs of data sets of each country. The matrix is overlaid with a heat map that highlights closer data sets in greens and those that are further from each other in oranges and reds. In what follows we first examine the scatterplot of each country's COVID-19 data (*new cases*, *new deaths*) pairs. Then we make further connections based on bottleneck distances between each country.

This work is licensed under a Creative Commons Attribution 4.0 International License

	AUS	BRA	FRA	DEU	IND	ITA	PER	ZAF	KOR	SWE	TWN	GBR	USA
AUS	0.00	0.53	2.09	1.07	0.06	1.20	2.72	0.44	0.06	1.51	0.06	3.00	0.31
BRA	0.53	0.00	2.09	1.07	0.53	1.20	2.72	0.53	0.53	1.26	0.53	3.00	0.53
FRA	2.09	2.09	0.00	2.09	2.09	2.09	2.72	2.09	2.09	2.09	2.09	3.00	2.09
DEU	1.07	1.07	2.09	0.00	1.07	1.20	2.72	1.07	1.07	1.21	1.07	2.85	1.02
IND	0.06	0.53	2.09	1.07	0.00	1.20	2.72	0.44	0.09	1.51	0.09	3.00	0.31
ITA	1.20	1.20	2.09	1.20	1.20	0.00	1.58	1.20	1.20	1.51	1.20	2.26	1.20
PER	2.72	2.72	2.72	2.72	2.72	1.58	0.00	2.72	2.72	2.40	2.72	1.57	2.72
ZAF	0.44	0.53	2.09	1.07	0.44	1.20	2.72	0.00	0.44	1.51	0.44	3.00	0.44
KOR	0.06	0.53	2.09	1.07	0.09	1.20	2.72	0.44	0.00	1.51	0.03	3.00	0.31
SWE	1.51	1.26	2.09	1.21	1.51	1.51	2.40	1.51	1.51	0.00	1.51	2.15	1.23
TWN	0.06	0.53	2.09	1.07	0.09	1.20	2.72	0.44	0.03	1.51	0.00	3.00	0.31
GBR	3.00	3.00	3.00	2.85	3.00	2.26	1.57	3.00	3.00	2.15	3.00	0.00	3.00
USA	0.31	0.53	2.09	1.02	0.31	1.20	2.72	0.44	0.31	1.23	0.31	3.00	0.00

Figure 2: Bottleneck Distance Matrix.

Shades of green indicate distances between 0.0 and 0.6; shades of yellow indicate distances between about 0.6 and 1.8; shades of orange indicate distances between about 1.8 and 2.4; shades of red indicate distances greater than about 2.4.

Note that the scatterplot data is time-flattened. In other words, each point represents a particular day's data of new cases versus new deaths (shifted by 14 days and smoothed); however, the timeline of when each data point occurs is not preserved. Nearby points may or not be near each other in time. However, some temporal information can be inferred from the shape of the data, as linear features are likely to represent a progression of day-by-day data, including separate waves of infections and deaths.

In future work, we would like to examine each wave separately. Even more interesting are the "loops" in the data, representing not just a new wave, but the progression of a wave from start to finish. It is precisely these loops in the point-cloud that TDA techniques can detect. Before proceeding to the data, it is important to understand the meaning of a persistence diagram. A persistence diagram is generated from a point cloud by taking the homology (Betti numbers) of the Rips Complexes associated with the point-cloud (Chazal and Michel (2021). Each point (p_{birth}, p_{death}) of the persistence diagram specifies when a feature (such as a loop) appears (p_{birth}) and when it disappears (p_{death}) in the process of computing persistence homology. These *birth* and *death* times do not correspond to actual dates associated with the data, nor does the term *death* in this context refer to actual deaths due to COVID-19. These are simply terms used in the field of TDA. The longer the time period between p_{birth} and p_{death}, the more significant that feature is. Graphically, the more significant features correspond to points that are further away from the diagonal. (It is helpful to think of points close to the diagonal as "noise" in the data, which do not actually correspond to significant features, while those points further from the diagonal are the important ones.)

This work is licensed under a Creative Commons Attribution 4.0 International License

Australia had three distinguished lines in Figure 3. It is probably due to the three waves before October 2021 (peaked in March and August of 2020, as well as September of 2021, respectively), and the first two waves were very small compared with other countries.

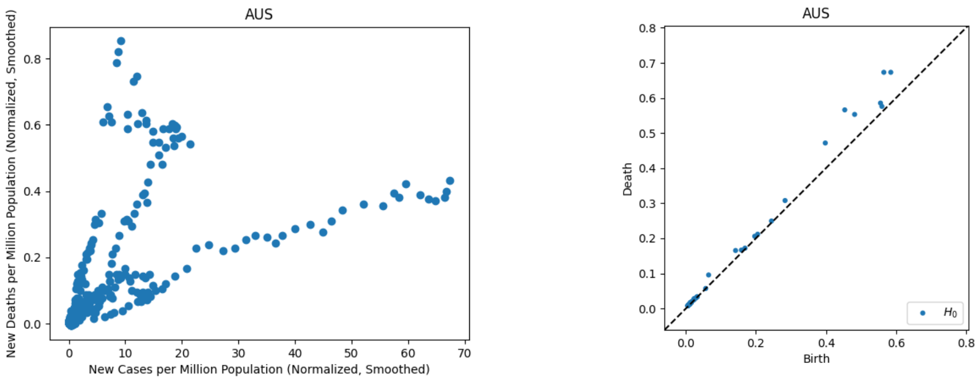

Figure 3 Australia. Left, a scatterplot of new cases vs. new deaths. Right, the persistence diagram generated from the scatterplot

As shown in Figure 4, Brazil had a unique pattern compared with other countries. It has two half loops and two small flat loops. It is probably due to the four waves it had before October 2021 which peaked in August of 2020, January, April, and June of 2021, respectively.

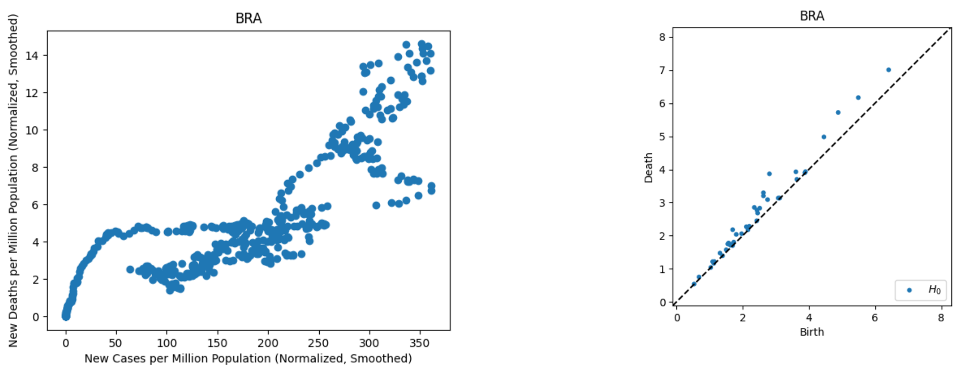

Figure 4 Brazil. Left, a scatterplot of new cases vs. new deaths. Right, the persistence diagram generated from the scatterplot

Germany had a very interesting pattern in Figure 5, with roughly three lines and a circular feature (loop) in the middle. We suspect that is because Germany had four waves before October (peaked in March of 2020, January, April, and August of 2021, respectively) and the first wave is very small compared with the later three waves; thus, the scatterplot shows three major lines. Also note the relatively strong signature of cycles in the data, which are picked up by TDA and suggested by the dots in the persistence diagram that are further from the diagonal.

This work is licensed under a Creative Commons Attribution 4.0 International License

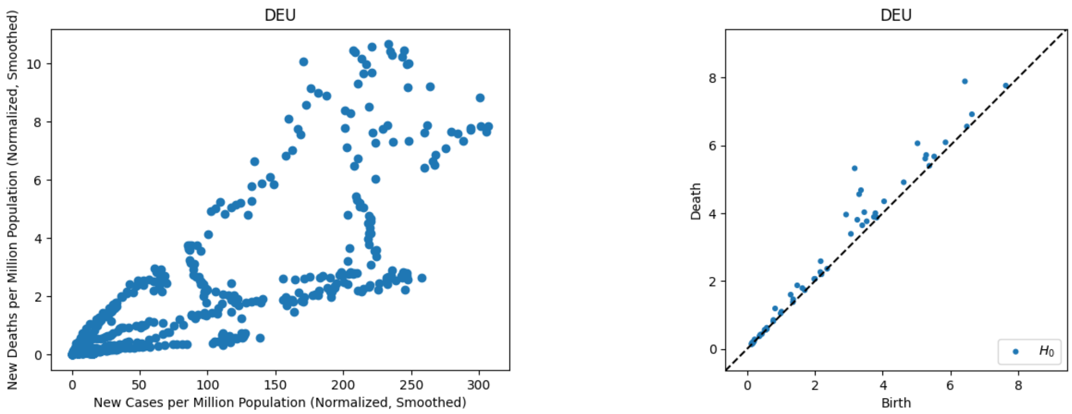

Figure 5 Germany. Left, a scatterplot of new cases vs. new deaths. Right, the persistence diagram generated from the scatterplot

France had four waves up to the end of September (peaked in March and November of 2020, April and August of this year, respectively), and the first one is very small compared with the latter three. That is probably why there are three distinct lines in Figure 6.

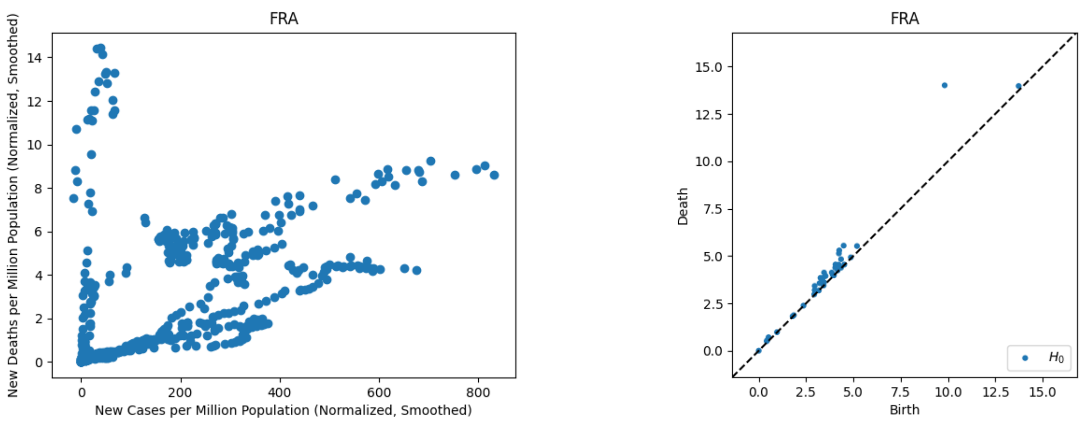

Figure 6 France. Left, a scatterplot of new cases vs. new deaths. Right, the persistence diagram generated from the scatterplot

Similar to France, the UK had four waves (peaked in March and November of 2020, and January and September of this year, respectively.) But different from other countries, its fourth wave is the smallest and can almost be ignored. Thus, the patterns in Figure 7 show only three major flat loops.

This work is licensed under a Creative Commons Attribution 4.0 International License

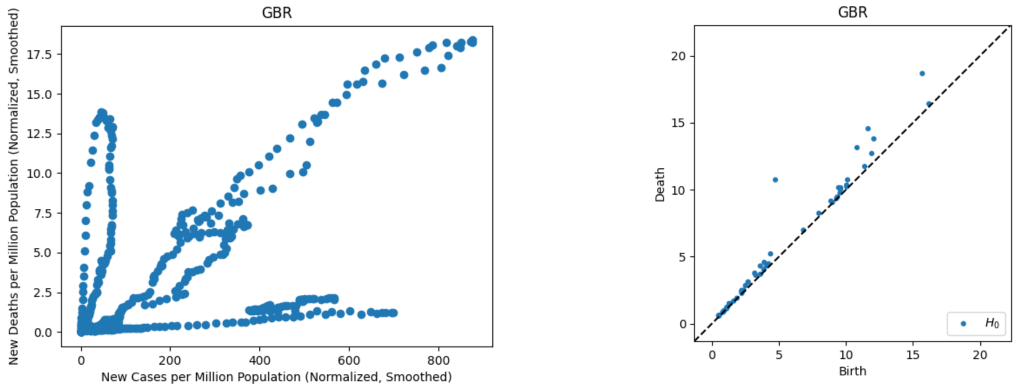

Figure 7 United Kingdom- Great Britain. Left, a scatterplot of new cases vs. new deaths. Right, the persistence diagram generated from the scatterplot

India is unique since it only had two major waves (one peaked in September of last year and one in May of this year, respectively). It has no waves after that, and the cases are continuing to decline even with the Omicron variant spreading around the globe. Since its first wave was relatively small, Figure 8 resembles a mostly curvy line with no major loops.

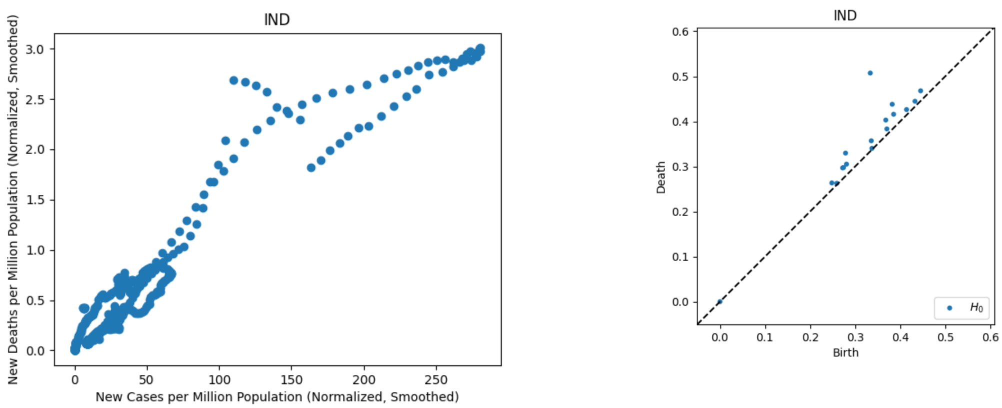

Figure 8 India. Left, a scatterplot of new cases vs. new deaths. Right, the persistence diagram generated from the scatterplot

Similar to other European countries, Italy had four waves (peaked in March and November of 2020, and March and August of this year, respectively.). However, its first wave is the smallest and can almost be ignored. Thus, the patterns in Figure 9 show only three prominent flat loops.

This work is licensed under a Creative Commons Attribution 4.0 International License

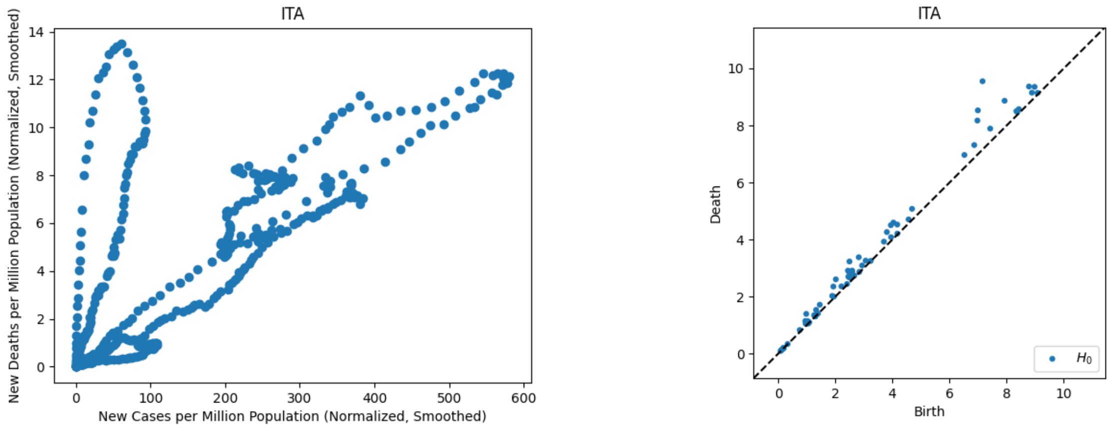

Figure 9 Italy. Left, a scatterplot of new cases vs. new deaths. Right, the persistence diagram generated from the scatterplot

South Korea is a country considered relatively successful in controlling the spread and Figure 10 shows four loops and a one-half loop. It had five waves (peaked in March, August, and December of last year and the other two peaked in September and May of this year, respectively.). However, the case numbers are relatively smaller compared with other countries, which contributes to a very low bottleneck distance from the baseline.

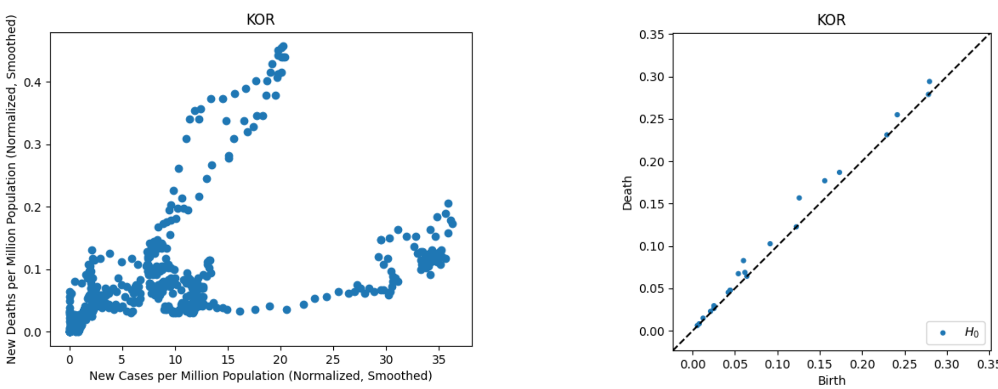

Figure 10 South Korea. Left, a scatterplot of new cases vs. new deaths. Right, the persistence diagram generated from the scatterplot

Figure 11 shows a super unique pattern with Peru. Peru had four major waves which peaked in May and August of last year, and February and April of this year, respectively. Thus, Figure 11 shows three major lines and a loop in the middle. Moreover, the data set seems to be noisier than most. Peru's COVID-19 statistics were reported in a rather haphazard manner, in which many days showed zero cases and deaths followed by days with much

This work is licensed under a Creative Commons Attribution 4.0 International License

larger numbers to catch up. Seven-day averaging did mitigate this discrepancy somewhat, but the noise was not entirely eliminated.

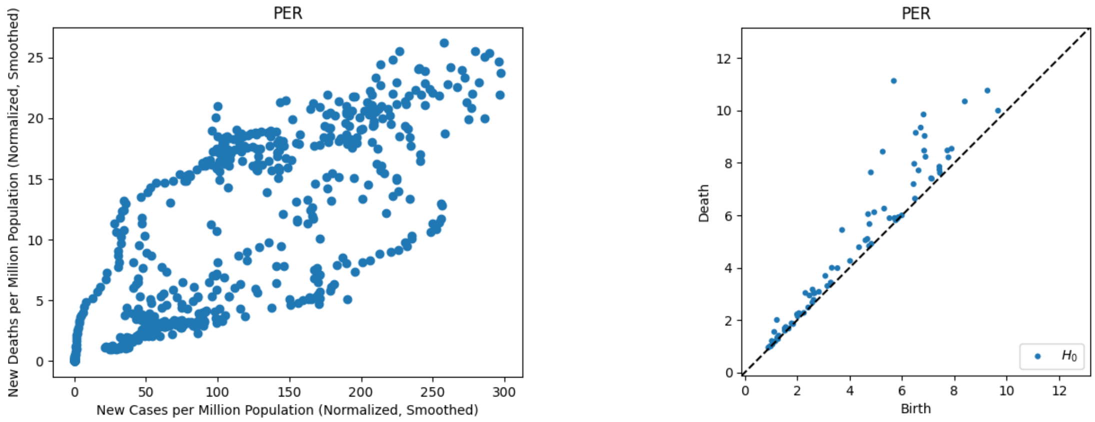

Figure 11 Peru. Left, a scatterplot of new cases vs. new deaths. Right, the persistence diagram generated from the scatterplot

Figure 12 displays a unique pattern for Sweden. It had four waves before October of this year, which peaked in June and November of last year and April and September of this year, respectively. Thus, it shows roughly three lines on the right and one loop on the left in Figure 12.

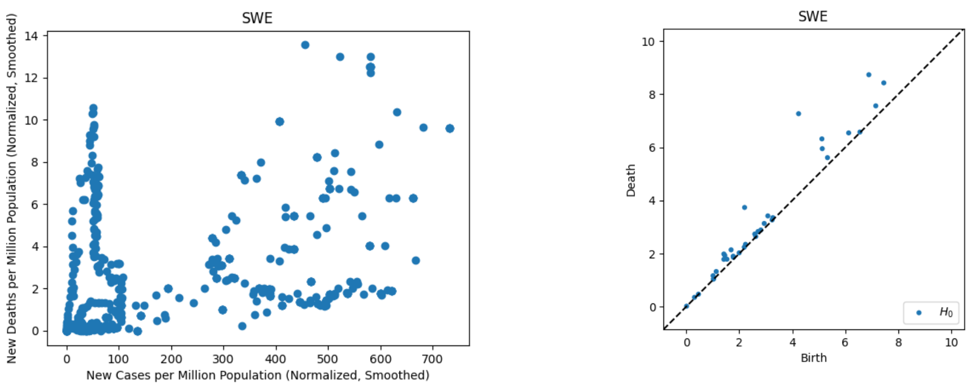

Figure 12 Sweden. Left, a scatterplot of new cases vs. new deaths. Right, the persistence diagram generated from the scatterplot

Taiwan has another distinguished pattern. It has two lines in Figure 13. It makes sense since it only had two waves, one peaked in March of last year and one in May this year. However, their case numbers are very low compared to other countries. Its highest case number was 543 in May of 2021.

This work is licensed under a Creative Commons Attribution 4.0 International License

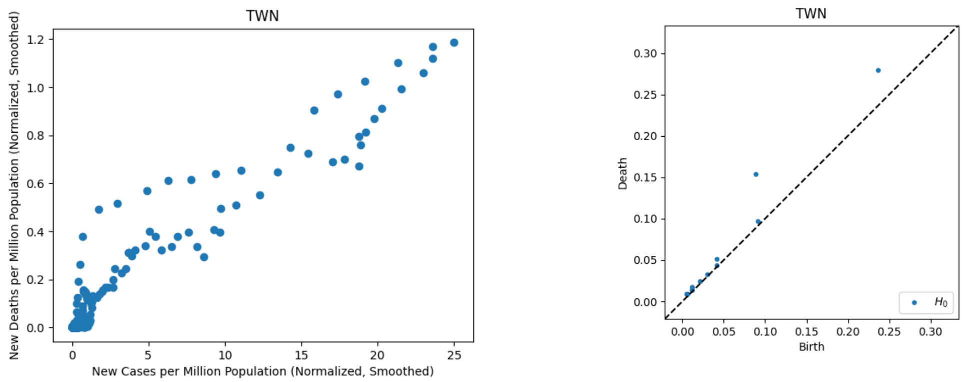

Figure 13 Taiwan. Left, a scatterplot of new cases vs. new deaths. Right, the persistence diagram generated from the scatterplot

Figure 14 shows a small loop on the left, along with three major lines. More than most other countries, the U.S. had five waves before October of this year which peaked in April, June of last year, and January, April, and August of this year, respectively. Compared with the later waves, the first wave could be ignored due to its size. Figure 14 corresponds to this trend with four major lines.

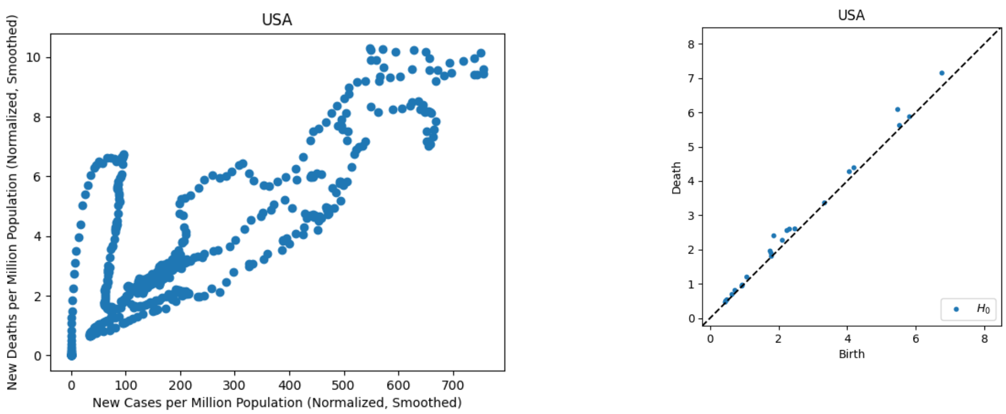

Figure 14 USA. Left, a scatterplot of new cases vs. new deaths. Right, the persistence diagram generated from the scatterplot

South Africa's pattern is displayed in Figure 15. It had four waves before October, and they peaked in July of last year, January, July, and August of this year, respectively. However, the last two waves are so close to each other, that we could also consider them as one large wave. Thus, Figure 15 has three major lines and some noise below the third line from the top, corresponding to the wave patterns.

This work is licensed under a Creative Commons Attribution 4.0 International License

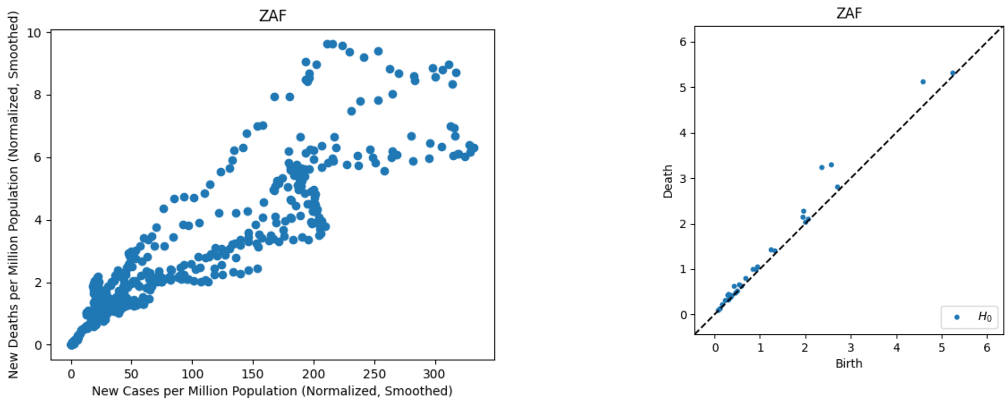

Figure 15 South Africa. Left, a scatterplot of new cases vs. new deaths. Right, the persistence diagram generated from the scatterplot

Finally, what follows are some observations based on the bottleneck distance matrix (Fig. 2). First, the COVID-19 data sets of three countries are "far away" from all of the others in terms of their bottleneck distances, Great Britain, Peru, and France. They are also far from one another. Second, there is a cluster of countries that are all relatively close to one another: Australia, India, South Korea, and Taiwan. Of these four countries, the daily cases and deaths (per million) were very low in three of them (AUS, KOR, and TWN) throughout the period under consideration, in comparison to the other countries. Thus, compared to the other data sets, these three data sets were relatively "featureless."

In other words, the AUS, KOR, and TWN data sets were close (in bottleneck distance) to a data set with no features at all ($\beta_1 = 0$). The reason that India (IND) is also so close to AUS, KOR, and TWN, likely has to do with the linearity of India's scatterplot. There are simply no large loops to detect by TDA methods (so, $\beta_1 = 0$ is predicted by its persistence diagram). In other words, TDA did not distinguish between India and countries with low case and death loads, even though India's experience with COVID-19 was indeed much different than that of AUS, KOR, or TWN. These relationships are displayed schematically in Figure 16.

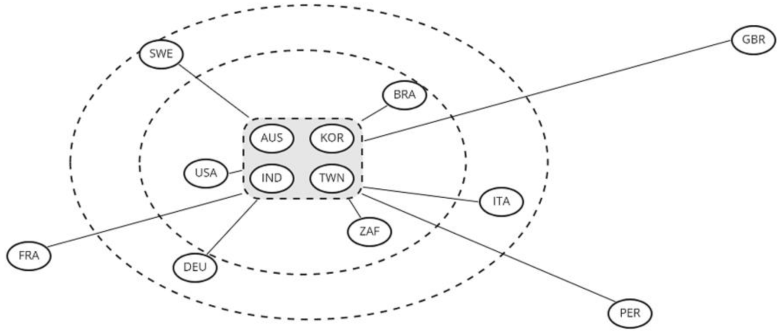

Figure 16. Schematic showing the relative closeness of data sets (in bottleneck distance) from the various countries under consideration

This work is licensed under a Creative Commons Attribution 4.0 International License

Brazil, South Africa, and the United States have bottleneck distances of about 0.31-0.53 from the core group of four. What this signifies is that there are some significant features in the data sets from those countries, which could indicate real differences in COVID-19 responses. Germany, Italy, and Sweden make up the third tier of distance from the core group. Finally, the three countries whose data was least similar to the others or to each other (in bottleneck distance) were Great Britain, Peru, and France. Figure 16 represents these bottleneck distances, especially with respect to the core group of four.

In summary, TDA methods point to major distinctions in COVID-19 data among the European countries, the United States, South Africa, and Brazil, compared to a few distinctions among Australia, India, South Korea, and Taiwan. Further analysis will be required to explore these apparent differences.

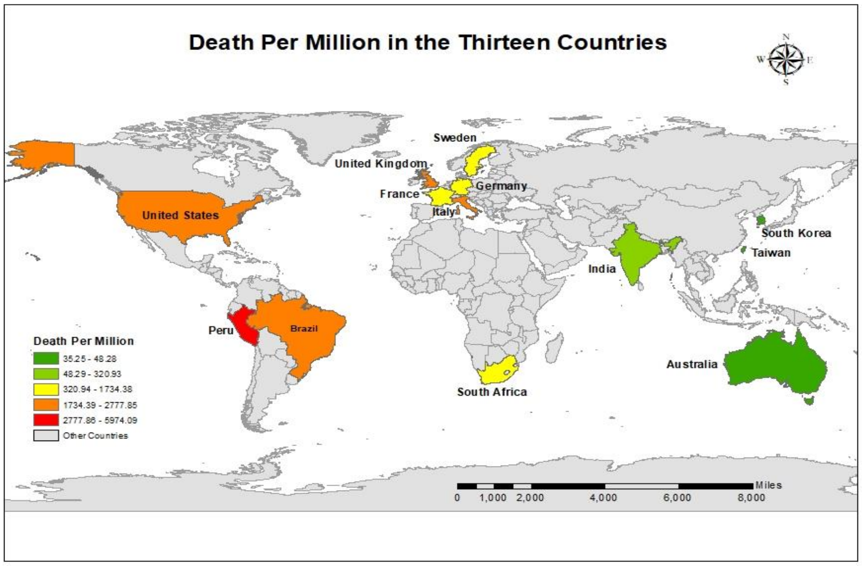

Figure 17. Deaths from COVID-19 per million population in the 13 countries

To further compare our data, we investigated the death per million for all countries. As displayed in Figure 17, the death per million is the highest in Peru, followed by Brazil, Italy, the US, and Britain. The lowest is Taiwan, followed by South Korea, Australia, and India. Thus, the grouping by TDA analysis makes sense from this aspect as well.

This work is licensed under a Creative Commons Attribution 4.0 International License

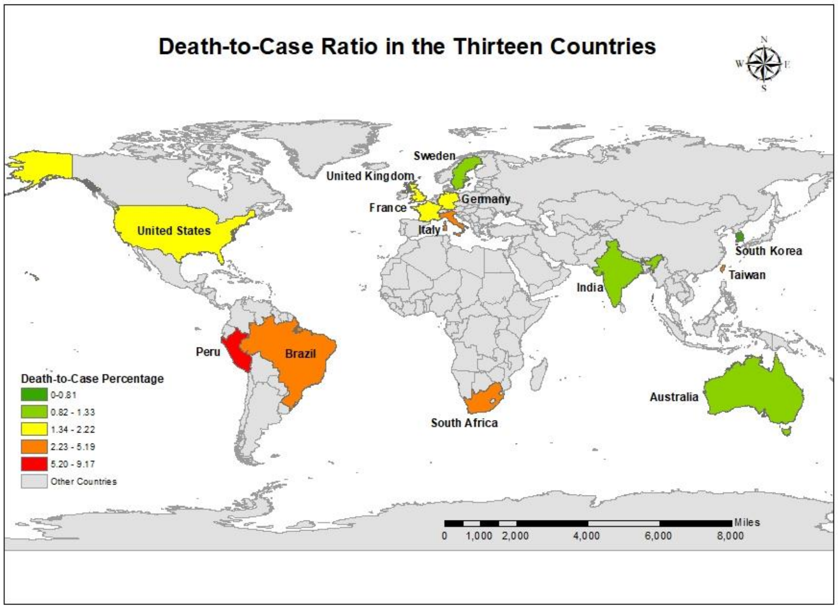

Figure 18. Death-to-Case Ratio in the 13 countries

Next, we also look at the death-to-case percentage (ratio), as displayed in Figure 18. One country that stands out as the highest is Peru (9.17%). It is followed by Taiwan (5.19%), South Africa (3%), Italy (2.8%), and Germany (2.22%). The lowest is South Korea (0.81%), followed by Australia (1.26%), Sweden (1.29%), India (1.33%), and the US (1.60%). It makes sense since South Korea was the first to start widespread testing and it was able to detect more cases, and therefore the death-to-case ratio is the best among the countries. On the other hand, we suspect that some of the countries (such as Italy and South Africa) were under-testing. Many Covid-19 cases are asymptomatic, and its percentage varies from 40.5% to 58.3%; it could be higher in pregnant women (54.11%), air or cruise travelers (52.91%), and nursing home residents or staff (Gruskay et. al., 2020; Ma et. al., 2021). This could be the reason that some countries have exceptionally high death-to-case ratios.

Public Policy and Other Causes

To further understand the reasons behind these data patterns, it is necessary to review the public health policies of these countries. First, countries with high similarities from TDA analysis are reviewed. South Korea was the best in early testing, and it was the first to initiate innovative on-street testing sites, drive-through testing, and large-scale testing of the general public (Lee and Choi, 2020). It also legislated new laws to protect healthcare workers to

This work is licensed under a Creative Commons Attribution 4.0 International License

prevent human capital loss. Similar to Taiwan, they are very efficient with extensive contact tracing and quarantine of populations that had close contact with the virus. The public preferred transparent information over privacy which enabled the government to implement innovative movement tracking systems. South Korea also has a high full vaccination rate (82.81%) among the population by the end of 2021. Taiwan also was the first place to stop all flights from much of China, stop cruise ships from docking, quarantine travelers from other areas, and execute widespread testing and mandatory mask-wearing. Taiwan's border restriction stands as one of the strictest in the world, requiring vaccinated entries, including citizens, to undergo 14 days of hotel quarantine, which is just a little shorter than China's mandatory 21-day hotel confinement (Davidson, 2021). On the other hand, Taiwan provided intensive economic support to patients with COVID-19 and the people they encountered. This helped to increase compliance among citizens with policies. The vaccination process started slowly, but it ended with 61% of the population fully vaccinated by the end of 2021. In summary, both Taiwan and South Korea were quick in their early actions. Those could be the reason that the two countries have similar data patterns.

At the start of the pandemic, Australia determined to force out COVID with lockdowns and travel restrictions and it responded to the Delta variant by instituting "circuit breakers": any detected cases would result in a lockdown to stop the spread at the beginning. More than half of Australians were locked down during the Delta surge, often in response to a smaller number of cases compared with other countries. For example, Melbourne enacted a curfew that limits people to leave their homes, along with hefty fines for non-compliance. Travel has been cut off between states, and many overseas Australians could not go back home because of monthly limits on how many people can return. The Australian response to the pandemic includes imposing tight restrictions on external borders and strict limits on social distancing and interaction.

The pandemic policy within a decentralized federal system has been most effective in flattening the curve for infections and containing the number of deaths. Political leaders at the federal and state/territory levels applied advice from health experts and committees with collective leadership. Public services have adapted well in responding to crisis conditions (Halligan, 2020). The vaccination process went well and about 76% of the population are fully vaccinated by the end of 2021. Similar to Taiwan, Australia is an island nation with a natural barrier to human movement; thus, it had considerable success in controlling the virus and it was the last of the G-20 countries to hit 1,000 total deaths.

India had faired relatively well with low deaths per million, only behind Taiwan, South Korea, and Australia. India was quick to react to the pandemic and initiated very strict lockdown policies at the beginning, and their outbreaks were well under control until the emergence of the Delta variant in the spring of 2021 when the case and death numbers skyrocketed. However, the short notice of the lockdown harshly impacted millions of low-income migrant workers. Moreover, the pandemic has placed an extra burden on the already overstretched healthcare system. In the second wave of the summer of 2021, India's infected cases and deaths grew exponentially and overwhelmed the healthcare facilities and patients

This work is licensed under a Creative Commons Attribution 4.0 International License

had trouble getting necessary treatments. The vaccination process was slow, and its population's full vaccination rate remains low at 43% by the end of 2021. Gauttam et. al., (2021) suggested India needs to expand the public healthcare system and enhance expenditure. The private healthcare system was not reliable during the emergency and the public health system is necessary for a large country with much of its population being rural and poor. India has a relatively young population and media age, compared with other countries. However, when we ran the correlation between median age and mean daily death per million, there was no strong correlation.

In general, the Asian countries and Australia did a decent job in controlling the spread at the early stage of the virus, which is probably the reason that all these four countries have similar patterns. The geographic isolation of most of these countries also helped. Australia and Taiwan are island states, and South Korea is mostly in isolation except for a border with the hermit state of North Korea and population movement was always strictly controlled. All the above four countries closed their borders and required travelers to mandatory quarantine in government-run or government-approved hotels. All these appeared to be effective in controlling the spread of the virus.

Compared with its counterparts in Asia, the U.S. was relatively slow in responding at the early stage. At the federal level, it was fast in shutting down the flights from China which caused some controversy in politics; however, it was slow in testing at the early stage and did not detect the spread in the community at early stage. On the plus side, the U.S. was very successful in developing several vaccines with the U.S. Food and Drug Administration (FDA)'s fast approval of emergency use under the federal government program "Operation Warp Speed." It started mass vaccination on December 14 of 2020 and ended in 2021 with about 66% of the population fully vaccinated.

Different from European countries, the U.S. public policy was mostly controlled at the state level under the federal system, which varies dramatically from state to state. Some states are stricter than others and some states had long lockdowns (such as California), while others opened up quickly (such as Georgia and Florida) (Hallas et. al., 2021). The disputes among states regarding their policies also highlighted the political division and shortcomings of the state system in the U.S. For example, in March 2020, New York Governor Cuomo threatened to sue the Rhode Island governor regarding the quarantine policy of New Yorkers entering Rhode Island, and he issued the same warning to President Trump's consideration of a short-term quarantine of hot spots in parts of the tri-state area of New York, New Jersey and Connecticut (Moreno, 2020). As a result, Rhode Island had to revise its policy and the federal government never quarantined hotspots in the tri-state. In addition, the relatively low population density helped to mitigate the spread outside of major metropolitan areas. The local policy variations might be the reason that the pattern from the U.S. is different from its European counterparts, but more resemble the three Asian countries and Australia.

Similarly, the policy reactions at the beginning were slow for the European countries. After the virus spread out to the communities, most of these countries had a hard time keeping the

This work is licensed under a Creative Commons Attribution 4.0 International License

spread under control, even with lockdowns and stay-at-home orders and a high amount of financial support for businesses and individuals. Some countries had multiple lockdowns based on the waves of outbreaks.

For example, Great Britain is the country with the least similarity to other studied countries. The country at the beginning had a herd immunity policy suggested by its scientific advisors. When the death toll went up, the government reversed its policy and ordered a complete lockdown on 23 March 2020, two months after the government's Sage committee of scientific advisers' first meeting about the crisis. This slow and gradualist approach was deliberate, proposed by official scientific advisers, and adopted by all governments of the UK (Sample and Walker, 2021). The government was unable to share vital information and scientific advice was impaired by a lack of transparency, input from international experts, and meaningful challenges. It rejected South Korea's approach and stopped mass testing in March 2020. The government provided poor protection for people in care homes, minority groups, and people with disabilities.

Overall, the biggest issue was the UK government's dismissing the precautionary principle in favor of not taking decisive action until definitive evidence emerged as the truth. On the positive side, the UK was the first Western country to roll out mass vaccination, and by late 2021, it has the fourth-highest full vaccination rate in Europe (73% of the population). The various policy decisions explain a large number of topological features in the data set, and hence the relatively large bottleneck distance from the other data sets.

Peru is another country that is not similar to the other countries. It was praised early for imposing swift pandemic measures (lockdown) 10 days after its first case, but its death-to-case ratio remains high, and its TDA result is very different from other countries. When the initial lockdown was imposed in March 2020, the financial pain caused by Peru's national lockdown pushed thousands of families to flee the capital of Lima for their rural hometowns. The government hoped that the lockdown pain suffered by its citizens would be short-term and justified by a triumph over the virus later. Instead, the mass migration of Peru's families symbolized its failure to help its people.

In addition, the high death ratio can also be explained by its underfunded health system with poor primary care, outdated hospitals, underpaid health professionals, and extremely low numbers of intensive care beds (Taylor, 2021). On top of that, it has diverse populations and geographies, deep poverty, crowded multigenerational housing, and a lack of testing facilities. Three-quarters of people work in informal jobs, and with government support coming later than expected, they had to risk either contracting covid-19 or starving during the lockdown. Like in other South American countries, most people eventually opted to work during the lockdown. Moreover, the government support check caused long queues outside banks and the inexperience of government food delivery to low-income families did not help to keep people at home. The country has also seen political turmoil during the pandemic, with its president ousted followed by three new presidents and massive protests. Vaccination

This work is licensed under a Creative Commons Attribution 4.0 International License

rollout was also criticized due to political scandals, and by the end of 2021, it has about 64% of the population fully vaccinated.

France is another interesting case because it is also different from other countries. The policy reflects the structural weaknesses of the health system, including its governance and decision-making process. It responded to the first wave with a full lockdown as an emergency response with a low level of preparedness for pandemics and a hospital-centered hhealthcaresystem. When the second wave hit, this response became a more level strategy trying to reconcile health needs from a broader perspective with socio-economic considerations but failed to construct an effective health strategy (Oz et. al., 2021). In order to achieve the balance, France will have to strengthen its health system capacity and improve the cooperation between shareholders at central and local levels with greater participatory decision-making that considers local-level realities and the range of needs. On the positive side, France has the third-highest full vaccination rate among its population in Europe by late 2021 (73%).

Sweden is a very unique case in the world. It was the first country to adopt a herd immunity policy. As a result, Sweden registered a high level of excess deaths in the first wave of the pandemic. Because it has a considerable population of vulnerable elderly with an elevated risk and non-pharmaceutical interventions were more efficient compared to those implemented in its neighbor Denmark. Rizzi *et al.,* (2021) identified the failure of senior care, housing, and the Swedish strategy of shielding the elderly which contributed to the high number of deaths at the beginning. However, the vaccination program was successful and about 71% of the population was fully vaccinated by the end of 2021.

Italy was the first developed country to be overwhelmed by the pandemic and the first country outside of China to apply lockdown measures. It has the highest death-to-case ratio in Europe. According to Vicentini and Galanti (2021), the crisis was in coordination with shortcomings typical of the Italian political and institutional system. The coalition government had considerable difficulties in facing the serious challenge and reacted with an incremental reaction. However, the situation deteriorated with a system already handicapped by capacity and the country still recovering from the 2012 economic crisis. On the positive side, Italy has the second-highest full vaccination rate (74%) in Europe by late 2021, only behind Portugal.

Germany has been relatively successful at preventing many of the negative pandemic impacts across the rest of Europe. Its policy responses were able to address some of the most important factors in containing an infectious disease outbreak, such as contact tracking, testing and containing cases, and sufficient state capacity to quickly implement policy. It also tried to reduce economic damage and ease restrictions and returned to normal economic activities earlier than other European countries (Desson et. al., 2021).

Brazil is another South American country plagued by the disease. Its fragmented authority and overlapping functions in a decentralized federal system make them more susceptible to

This work is licensed under a Creative Commons Attribution 4.0 International License

coordination problems than centralized systems. The pandemic requires national health system stewardship. Using sub-national NPI data, Knaul et. al., (2021) concluded that Brazil suffers from protracted, high levels of COVID-19 mortality and inadequate pandemic responses, including little testing and disregard for scientific evidence. Kmietowicz (2021) suggested the government adopt evidence-based public health measures to control the disease and treatment guidelines for those infected. On the other hand, Brazil's vaccination program ended in 2021 with over 67% of the population fully vaccinated.

The Covid-19 TDA pattern in South Africa is more similar to that of Australia and Brazil. The pandemic arrived in South Africa later than in other countries. However, this is a nation plagued by social issues. The strategies to control the spread of pandemic-social distancing and frequent handwashing—are not easy for people living in highly dense communities, with poor sanitation and access to clean water. South Africa has 15 million people living in places with high HIV infection rates and the immunocompromised populations are at greater risk of the pandemic (Lancet, 2020). To make matter worse, South Africa has the lowest full vaccination rates (26%) of our 13 studied countries. Poverty and poor sanitation make the pattern similar to that of Brazil. That is why the two countries have similar TDA results.

Discussion

During our research, we encountered some issues with data. Data collection among different countries is not uniform. Different countries count case and death numbers differently. For example, a positive postmortem COVID test will be counted into the case and death totals in the U.S., but it would not be included in other countries, such as Germany. In addition, the U.S. coronavirus relief legislation created a 20% premium, or add-on, for COVID-19 Medicare patients. Even though there have been no public reports that hospitals are exaggerating COVID-19 numbers to receive higher Medicare payments, there is an incentive to do so. Cases can be greenlighted for Medicare payment and eligible for the 20% add-on allowed under the relief act. For example, most states (such as Minnesota and California), list only laboratory-confirmed COVID-19 diagnoses, but others, specifically New York, list all presumed cases. The latter was allowed under guidelines from the Centers for Disease Control and Prevention as of April 2020, and that could result in a larger payout. Due to the time limitation, we were not able to dig deep into more TDA analysis. For future research, we would like to examine each wave in each country separately with the TDA approach.

Conclusion

The patterns of COVID-19 spread throughout these 13 countries are different, but some countries are more similar than others. We identified three groups out of the 13 countries, with Taiwan, South Korea, and Australia as a similar group. Peru, Britain, and France are vastly different from other countries. The U.S. is more similar to the Asian countries and Australia than the European countries. In addition, we also looked at the policy and other possible reasons behind the pattern. In general, the island countries or the countries with the least amount of population flow, combined with early actions and successful policies did the

This work is licensed under a Creative Commons Attribution 4.0 International License

best in controlling the spread, such as the case with Taiwan, Australia, and South Korea in the top group.

TDA can be a useful initial analysis of the data and it could tease out anomalies, which can then lead to a more comprehensive analysis using other techniques. In this case, we followed the TDA analysis with an examination of death per million and death-to-case ratio in 13 countries and found that the geographic locations of the countries are also important in explaining their data patterns. In order to find out if the age of the population impacts the death rate, we ran the correlation between median age and mean daily death per million; however, there was no strong correlation.

Lastly, policies of countries were investigated for each country, and suggestions can be made to national-level governments so they can be better prepared for future pandemics. Learning from the countries with better control of the spread, we conclude that early action in widespread testing, contact tracing, quarantine of close-contact populations and travelers, border control, mask usage, and vaccination are all effective ways to control the virus outbreaks.

References

Balmford, B., Annan, J.D., Hargreaves, J. C., Altoè, M., Bateman, I. J. (2020) Cross-Country Comparisons of Covid-19: Policy, Politics and the Price of Life. *Environmental Resource Economics* 1-27. doi: 10.1007/s10640-020-00466-5.

Chazal, F. and Michel, B. (2021) An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists. *Frontiers of Artificial Intelligence.* Available at: https://hal.inria.fr/hal-01614384 (accessed 25, December 2021)

Chen, Y. and Volić, I. (2021) Topological data analysis model for the spread of the coronavirus. *PLOS ONE.* Available at: https://doi.org/10.1371/journal.pone.0255584 (accessed 25, December 2021).

Chung, M. and Ombao., H. (2021) Lattice Paths for Persistent Diagrams. Available at: https://export.arxiv.org/pdf/2105.00351 (accessed 25, December 2021).

Davidson, H. (2021) 'How long can you maintain it?' Cost of Taiwan's pursuit of Covid zero starts to show. *The Guardians*. Available at: https://www.theguardian.com/world/2021/nov/08/how-long-can-you-maintain-it-cost-of-taiwans-pursuit-of-covid-zero-starts-to-show (accessed 25, December 2021).

Desson, Z., Lambertz, L., Peters, J. W., Falkenbach, M., & Kauer, L. (2020) Europe's Covid-19 outliers: German, Austrian and Swiss policy responses during the early stages of the 2020 pandemic. *Health policy and technology* 9(4): 405-418.

Hallas, L., Hatibie, A., Majumdar, S., Pyarali, M., & Hale, T. (2021) *Variation in US States' Responses to COVID-19*. Oxford: University of Oxford.

Halligan, J. (2020). Australia's Response to the COVID-19 Pandemic. *Good Public Governance in a Global Pandemic* 233.

Harrison, E., Docherty, A. and Semple, C. (2020) COVID-19: time from symptom onset until

This work is licensed under a Creative Commons Attribution 4.0 International License

death in UK hospitalised patients. *Scientific Advisory Group for Emergencies*. Epub ahead of print 2020. Available at: https://www.gov.uk/government/publications/co-cin-covid-19-time-from-symptom-onset-until-death-in-uk-hospitalised-patients-7-october-2020.

Jin, R. 2021. The lag between daily reported COVID-19 cases and deaths and its relationship to age. *Journal of Public Health Research* 10(3):2049.

Gauttam, P., Patel, N., Singh, B., Kaur, J., Chattu, V. K., & Jakovljevic, M. (2021) Public health policy of India and COVID-19: Diagnosis and prognosis of the combating response. *Sustainability* 13(6): 3415.

Gruskay, J. A., Dvorzhinskiy, A., Konnaris, M. A., LeBrun, D. G., Ghahramani, G. C., Premkumar, A., & Ricci, W. M. (2020). Universal testing for COVID-19 in essential orthopaedic surgery reveals a high percentage of asymptomatic infections. *JBJS* 102(16), 1379-1388.

Kmietowicz Z. (2021) Covid-19: Failed response in Brazil has led to humanitarian catastrophe, says MSF. *BMJ* 373 :n1002 doi:10.1136/bmj.n1002

Knaul, Felicia Marie et al. (2021) Punt Politics as Failure of Health System Stewardship: Evidence from the COVID-19 Pandemic Response in Brazil and Mexico. *The Lancet Regional Health – Americas* 4: 100086.

Lee, D. and Choi, B. (2020) Policies and innovations to battle Covid-19 – A case study of South Korea,*Health Policy and Technology* 9(4): 587-597.

Lancet, T. (2020) Redefining vulnerability in the era of COVID-19. *Lancet* 395(10230): 1089. Ma, Q., Liu, J., Liu, Q., Kang, L., Liu, R., Jing, W., ... & Liu, M. (2021) Global Percentage of Asymptomatic SARS-CoV-2 Infections Among the Tested Population and Individuals with Confirmed COVID-19 Diagnosis: A Systematic Review and Meta-analysis. *JAMA Network Open 4*(12), e2137257-e2137257.

Mukherjee, S., Wethington, D., Dey, T., Das., J. (2021) Determining clinically relevant features in cytometry data using persistent homology. *bioRxiv*. DOI: 10.1101/2021.04.26.441473.

Moreno, J. Edward. (2020) Cuomo threatens to sue RI over new policy to find New Yorkers in the state. *The Hill*. Available at: https://thehill.com/homenews/state-watch/490019-cuomo-threatens-to-sue-ri-over-new-coronavirus-policy-targeting-ny (accessed 25, December 2021).

Naidu, T. (2020) The COVID-19 pandemic in South Africa. *Psychological Trauma: Theory, Research, Practice, and Policy* 12(5): 559–561.

Or Z, Gandré C, Durand Zaleski I, Steffen M. (2021) France's response to the Covid-19 pandemic: between a rock and a hard place [published online ahead of print, 2021 Mar 5]. *Health Econ Policy Law* 1-13. doi:10.1017/S1744133121000165

Oudot, S. Y. (2015) Persistence Theory: From Quiver Representations to Data Analysis. In: Guralnick et al, (eds) *AMS Mathematical Surveys and Monographs* 209.

Pérez-Moraga, R.; Forés-Martos, J.; Suay, B.; Duval, J.; Falcó, A.; Climent, J. (2020) A COVID-19 Drug Repurposing Strategy Through Quantitative Homological Similarities by using a Topological Data Analysis Based Formalism. Available at: 2020120281 (doi: 10.20944/preprints202012.0281.v1) (accessed 25, December 2021).

This work is licensed under a Creative Commons Attribution 4.0 International License

Polterovich, L.; Rosen, D.; Samvelyan, K.; Zhang, J. (2020) Topological Persistence in Geometry and Analysis in the AMS University Lecture Series, Vol. 74.
Ritchie, H.; Mathieu, E.; Rodés-Guirao, L.; Appel, C.; Giattino, C.; Ortiz-Ospina, E.; Hasell, J.; Macdonald, B.; Beltekian, M.; Roser, D. (2020) Our World in Data. Available at: https://ourworldindata.org/coronavirus (accessed September 30, 2021).
Rizzi S, Søgaard J, Vaupel JW. (2021) High excess deaths in Sweden during the first wave of COVID-19: Policy deficiencies or 'dry tinder'? *Scandinavian Journal of Public Health.* Available at: doi:10.1177/14034948211027818 (accessed December 25, 2021).
Sample, I. and Walker, P. (2021) Covid response "one of UK's worst-ever public health failures." Available at: https://www.theguardian.com/politics/2021/oct/12/covid-response-one-of-uks-worst-ever-public-health-failures (accessed September 10, 2021).
Taylor L. (2021) Covid-19: Why Peru suffers from one of the highest excess death rates in the world. *BMJ* 372 :n611 doi:10.1136/bmj.n611.
Tralie, C., Saul, N., Bar-On, R. (2018) Ripser.py: a lean persistent homology library for python. *The Journal of Open Source Software* 3(29):925.
Van Rossum, G.; Drake Jr., F. L. (1995) *Python Reference Manual.* Amsterdam: Centrum voor Wiskunde en Informatica Amsterdam.
Vicentini, G., & Galanti, M. T. (2021) Italy, the sick man of Europe: policy response, experts and public opinion in the first phase of covid-19. *South European Society and Politics* 1-27.

Paper Received February 1, 2022; Accepted July20, 2022; Published November 2, 2022

Appendix: Descriptive Statistics of the Raw Data

Figure 19 summarizes the descriptive statistics of the new cases and new deaths data for each of the 13 countries under consideration in this paper, obtained from Our World in Data (Ritchie, *et al*. 2020). The data was first smoothed using a 7-day moving average.

		New Daily Cases				New Daily Deaths			
	N	Mean	STD	Med.	Max.	Mean	STD	Med.	Max.
AUS	610	155.27	356.97	16.57	1765.43	2	4.05	0.14	22
BRA	583	36857.66	20670.74	36788.93	77327.57	1025.72	696.07	957.07	3123.57
FRA	612	12217.7	11861.25	10259.71	56233.14	192.8	200.14	111.14	975.14
DEU	612	6923.02	7185.4	3518.71	25757	154.43	211.43	49.86	894.43
IND	606	55894.98	79851	35900.43	391232	742.65	974.81	473.71	4190
ITA	605	7749.38	8606.28	4249.43	35072.57	217.56	226.67	94.07	814.29
PER	635	3841.37	2492.83	3379.14	9928.29	352.48	252.61	313.86	874
ZAF	598	5108.3	5118	2632.21	19955.71	153.15	137.57	99.36	577.57
KOR	615	486.65	534.42	384.79	2288.14	4	4.55	2.79	23.43

This work is licensed under a Creative Commons Attribution 4.0 International License

SWE	604	1916.69	2062.84	770.57	7441.71	24.72	29.59	13.71	137.57
TWN	620	26.59	88.03	2.43	596.57	1.38	4.38	0	28.29
GBR	605	12667.84	14312.05	4919.79	59828.57	226.86	305.22	83.43	1253
USA	614	70165.09	62855.17	51239.79	251863.7	1124.04	853.82	868.36	3425.57

Figure 19 Summary Statistics by country for new daily cases and new daily deaths (smoothed data).

Paper Received February 1, 2022; Accepted July20, 2022; Published November 2, 2022

This work is licensed under a Creative Commons Attribution 4.0 International License