Polio Vaccination Inequity: A Persistent Curiosity

Kiri Strack, Jeff Uyekawa

December 11, 2023

1 Introduction

Polio is a disease that causes paralysis and in certain cases can prove fatal. Prior to the invention of the polio vaccine in the 1950s, polio caused paralysis in over 15,000 people in the United States each year. Now, due to the effectiveness of the vaccine, there have been no cases of polio in the United States since 1979. In 1988, the World Health Administration established the Global Polio Eradication Initiative, and since then, polio cases have decreased globally by more than 99 percent. As of today, polio has not been completely eradicated, and the CDC estimates that failure to eradicate the disease could allow for it to reappear with 150,000 to 200,000 new cases each year. To that end, it is of paramount importance to better understand the global polio vaccination effort in order to find potential locations with inequitable rates of polio vaccination. Utilizing data from the World Health Organization and techniques in Topological Data Analysis, we aim to identify patterns of vaccination inequity globally with the hope that our findings can further the effort for global eradication of the disease.

2 Data

We set out to explore the World Heath Organization's data set of polio vaccination coverage [5]. The primary data consists of a vaccination value representative of the percent of 1-year-olds vaccinated for polio. We filtered the data to include only values for the year 2022, collected latitude and longitude coordinates for each country, and then added data for each country's GDP, life expectancy, and median age.[2][3][4]

After merging data sets and dropping any countries containing null values, our data set represents 154 countries. To further prepare our data for use with persistent homology, we scaled the values of each column using the min/max scaling method shown below.

$$x_{scaled} = \frac{x - X_{min}}{X_{max} - X_{min}}$$

This scaling approach ensures that no one column dominates the calculation of persistent homology on our point cloud. We used the scaled data set for the weighted alpha complex filtration, and we used the original data set from the WHO for the sub-level set filtration.

3 Methodology

The main task in exploring our data with persistent homology was thoughtful design of a filtration. To that end, we attempted to explore our data with two different filtration designs.

3.1 Weighted Alpha Complex

In the initial exploratory stages, we compared each country in a "similarity space" where each country was assigned a point in $\mathbb{R}^6$ based on its latitude, longitude, GDP, median age, life expectancy, and vaccination value. Most of the patterns we observed seemed to be based on geographic factors, so we chose to try a weighted alpha complex with the hope of finding more meaningful patterns in the data.

An alpha complex is a subcomplex of a Delaunay complex which avoids sliver triangles in a triangulation and makes it more intuitive to map into a topological space. Though it is similar to a Čech complex, the basis in the Delaunay Complex is more favorable for this data set. A weighted alpha complex causes the radii of the individual balls to grow at different rates according to manually determined weights.

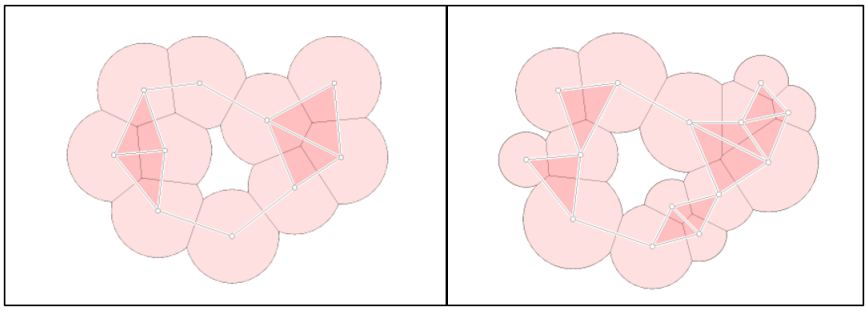

Figure 1: Illustration of an Alpha Complex on the left and a Weighted Alpha Complex on the right. Image taken from [1].

We decided to set the weights based on each country's vaccination rate in the hope that highly vaccinated countries would connect more quickly to their neighbors in the similarity space so that we might identify patterns in regions with potentially less equitable vaccination rates.

3.2 Sub-level Set Filtration

Often, after finding interesting patterns with persistent homology, it can be difficult to work backwards to identify where those patterns appear in the original data set. Currently, when computing persistent homology with many of the standard packages, there is not a simple way to nominate representatives of homology classes in the original data. One of the main reasons why we wanted to construct a sub-level set filtration on our data was because there is a reasonably easy method of assigning representatives to homology classes. In this way, we aimed to find countries that had lower vaccination values than other countries in their surrounding region.

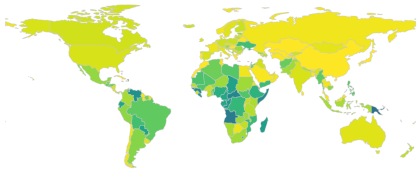

Figure 2: GeoPandas World map colored by vaccination value

To prepare our data for calculation, we colored a world map using the vaccination value for each country, and then padded the map with white pixels to make a square matrix.

The trick for assigning representatives of homology classes comes from the fact that birth and death times in a sub-level set image filtration are pixel values[6]. We introduced some random noise to the matrix to ensure that each pixel had a unique value. Next, we calculated persistence and plotted a birth-lifetime persistence diagram to choose a threshold lifetime value. For all points above the threshold, we collected birth times which could be assigned back to specific pixels in our original image.

4 Results

In this section, we discuss the results and interpretations obtained from each of our filtration techniques.

4.1 Weighted Alpha Complex Results

Choosing to use the vaccination rate as the weight caused countries with higher vaccination rates to connect and form higher dimensional simplices faster than countries with lower vaccination rates. Our hope was to look at the persistent homology and see what factors lead to either an increase or decrease of isolation of the countries with a low vaccination rate.

We first created an unweighted alpha complex comparing all the factors, then deciding that geography was skewing the results we created a weighted alpha complex in an attempt to mitigate the geographical influence. We compared the two diagrams and found that they looked relatively similar. (Although the

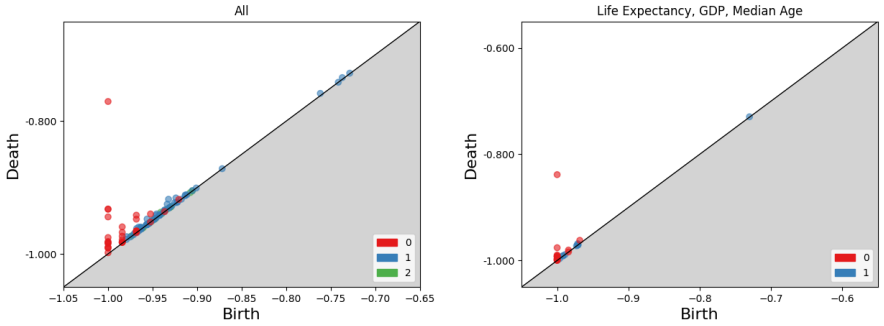

Figure 3: Comparison of persistence diagrams when using all values in the similarity space vs no geographic values.

unweighted complex does chart some 3-simplices.) There were two 0-simplices that persisted longer than the others, most of the other simplices died very close to the diagonal. Since this yielded a similar result, we then created a weighted alpha complex without using geographical factors. Thinking that the geography was still having too much of an influence on the homology. The two persisting 0-simplices still remained in this analysis of the data. So we concluded that these 2 main components must be caused by something other than geographical closeness.

We then created individual weighted alpha complexes for each unique pair of factors. Hoping to see if we could determine the cause of the 2 main components seen above. The pairwise comparison of our factors did not yield any insight. Unfortunately it is very hard to back track to see where the simplices originate. Thus making it very difficult to come to any definitive conclusion about which factors really contributed to the connected components we saw.

4.2 Sub-level Set Results

We were able to visually identify regions that were creators of 0-dimensional homology classes with reasonably high persistence. One might wonder whether this method adds any value beyond sorting our data by vaccination value from lowest to highest. The method of persistent homology does seem to add value over a simple sorting technique by identifying countries that have lower values than their surrounding region rather than just focusing on the overall lowest valued countries. If an entire region has relatively low vaccination values, we would identify the lowest value, which would quickly connect in the cubical complex with other low valued pixels. This can be visualized in figure 4.

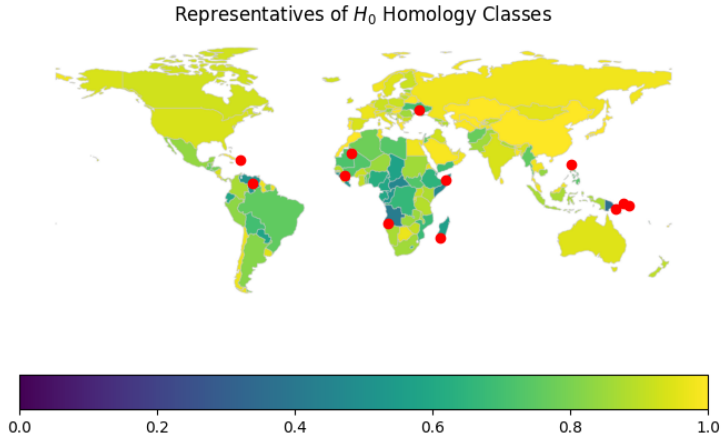

Figure 4: World map colored by vaccination rate with creators of 0-dimension homology classes plotted in red.

5 Future Work

5.1 Weighted Alpha Complex

Firstly, the use of weighted alpha complexes to study vaccination rates was limited in its explainability. In the future we would suggest the development of some way to trace the birth and death of persistence points back to points in the original data. One could spend more time exploring our original data set in an attempt to give meaningful explanation to the patterns observed in the persistence diagrams. Another possible avenue would be to use a cut off value for death times and to try and evaluate more in that specific environment. We also would suggest using different values for the alpha complex weights in search of more meaningful patterns. In the future we would also suggest continued work on this filtration using additional metrics for comparison.

5.2 Sub-level Set

In our current work, we manually change the names of countries in our data set to match those in the geopandas dataframe. We suggest developing an automated approach to this task. Next, our method allows us to identify creators visually, however it would be interesting to develop a method to link those creators to country names without having to look at the original map. Finally, one could utilize this sub-level set approach to explore more densely packed geographic areas. For example, the countries that we identified in our study could be studied by county or city vaccination rates.

6 References

References

[1] Alpha complex. https://courses.cs.duke.edu/fall06/cps296.1/Lectures/sec-III-4.pdf. Accessed: 2023-12-05.

[2] Gdp (current us$). https://data.worldbank.org/indicator/NY.GDP.MKTP.CD. Accessed: 2023-12-05.

[3] Life expectancy. https://www.worlddata.info/life-expectancy.php. Accessed: 2023-12-05.

[4] Median age. https://www.cia.gov/the-world-factbook/field/median-age/country-comparison/. Accessed: 2023-12-05.

[5] Polio (pol3) immunization coverage among 1-year-olds (%). https://www.who.int/data/gho/data/indicators/indicator-details/GHO/polio-(pol3)-immunization-coverage-among-1-year-olds-(-). Accessed: 2023-12-05.

[6] Lower star filtration documentation. https://ripser.scikit-tda.org/en/latest/notebooks/Lower%20Star%20Image%20Filtrations.html, 2022.