Shipwreck Analysis:

Thunder Bay National Marine Sanctuary

16 December 2019

GEOG 6000

Melissa Gfeller

Abstract

In 2014, the Thunder Bay National Marine Sanctuary expanded by over 3,800 square miles to include an additional 100 known and possible historic shipwrecks.[1] The surveyed wrecks are from the 1800s and the early 1900s. It is difficult to gather the pertinent data because the wrecks need to be surveyed individually to assess the proper location, cause, and cargo along with many other factors. Out of the water, each of the wrecks need to be researched to identify specific information about the wrecked ship using the data from the survey. Some shipwrecks are more difficult to identify due to their condition or even a lack of historical data to cross reference. While there is a lot that can be learned from the analysis of historic shipwrecks getting the initial information takes time and effort. Even when information is gathered about a shipwreck the data may not always be complete. Working around this in an analysis can be challenging.

Introduction

Thunder Bay National Marine Sanctuary expanded, from 448 square miles to 4,300 square miles in 2014. More shipwrecks that had not yet been surveyed are now within the bounds of the Sanctuary. The 448 square miles only included the area of Alpena Bay, but now extends out to the US and Canadian border, north up to Cheboygan and Presque Isle county, and south to the lower border of Alcona county.[2] Based on current shipwreck research, the extended boundaries of the sanctuary include up to 100 known and suspected historic shipwreck sites. The cold freshwater conditions of the Great Lakes makes it possible for the shipwrecks to be preserved. The hunt for shipwrecks starts with wide range sonar to find anomalies on the sea floor. Then those anomalies are analyzed again with high resolution sonar imaging. If it is

[1] Appendix G
[2] (Sanctuaries 2014)

found to be a shipwreck, then a team of divers is sent down to get close up pictures and to survey the wreck for various information. The data collected by the divers allows researchers to identify the ship based on the length, beam, gross tonnage, cargo and type of vessel. Cross referencing the collected data and images with records of lost or abandoned ships makes it possible to identify them.

The surveyed data has to be obtained directly by divers or indirectly by remote sensing technology. Technology has made great strides with the use of advanced high resolution imagery and sonar. The improvements in technology allow for surveying of deep wrecks without requiring technical deep divers. Using technology to survey shipwreck sites pinpoints locations and helps make it safer when using divers. There are recent articles discussing the importance of historic shipwrecks and the technology used in mapping them. Marine archaeologists tested new 3D modeling techniques on the Defiance, a schooner, which sank in Lake Huron. It is the hope that 3D modeling of shipwrecks will become the standard for analysis and identification. With the improvements in 3D mapping, researchers want to use 3D printers to create models of the shipwrecks.[3] Additionally, scientists have been mapping the Great Lakes shipwrecks with lasers, sonar, photo sleds and robots in order to investigate shipwrecks. They used laser scanning to create data point clouds for 3D mapping. The wreck site they focused on was the Monohansett. If laser scan technology can be used for all the wrecks, it will help to better document Thunder Bay Marine Sanctuary resources.

I reviewed an article about chi-square testing. One of the problems I ran into was what to do after the chi-square test was significant. This article had a variety of different methods for post analysis of chi-square tests. The four methods they used were calculating residuals,

[3] (Lake Huron shipwrecks 3D imaged by marine archeologists 2015)

comparing cells, ransacking and partitioning. I was unable to use these based on the content of my data. I thought that calculating the residuals would be good but many of the variables I tested did not have an expected value they were unique to the specific observation. The other problem was that many of the variables were categorical values. Another challenge I faced was that most of my comparison tables were very large while the examples were symmetrical small tables generally 2x2.[4] It would be difficult to do the other methods with such large tables.

Based on additional research I also looked into Fisher's Exact Test of Independence. I was going to add this into my analysis of the chi-squared test until I reviewed the example coding. From the results it would only extrapolate the exact percentage of each comparison. Overall that does not improve my understanding of the statistical significance. So I opted to just leave the chi-square as a single analysis of the correlation between two variables.[5]

To expand the analysis of the k-means clustering I did additional research into cluster analysis. I used a variety of resources including the class notes and the labs that discussed interpretation of different methods. These resources were used to implement the code to run the analysis.[6] Another document I reviewed gave information on indexes and the best criteria to measure success. The interpretation for calinski-harabasz and silhouette index are similar. In the plots for both you are looking for a high value at a peak. If there are no peaks then any of the solutions could be used.[7] The last article that I read discussed the merits of silhouette analysis and the Davies-Bouldin Index. They concluded that even though the silhouette index is more accurate it is a more time consuming process. The Davies-Bouldin Index is less accurate it takes

[4] (Sharpe 2015)
[5] (Mangiafico 2015)
[6] (Brewer 2019)
[7] (Desgraupes 2017)

significantly less time to run. For the case of my analysis I used the silhouette analysis because of the higher accuracy.[8]

Methods

The study area I chose was the Lake Huron region specifically in and around the Thunder Bay National Marine Sanctuary. This area does not encompass all the shipwrecks but it includes a large majority. The shipwreck data that I acquired included shipwrecks that had broken in to multiple pieces and scattered cargo. For the purposes of my analysis I consolidated the data to only "shipwreck sites" which means only one point per shipwreck is used. These points also have the most complete information for each shipwreck. One of the difficulties I encountered during my analysis was the amount a missing data values. Based off the condition of shipwrecks it can be impossible to determine more than its location and basic size data. This uncertainty accounts from most of the missing values.

Although plotting the data is not necessary I took the time to make overview plots. It was a challenge to overlay multiple layers to generate a good plot. One of the main struggles I avoided was using ArcPro and ArcCatalog to change the projections prior to exporting the shapefiles. Initially I attempted to change the bathymetry layer to the correct project within R but found that the projection did not match the other layers. Instead I used ArcCatalog to change the projection and the file is now in the correct position in relation to the other layers.

One method of analysis was the chi-squared analysis to compare categorical data. This worked fairly well but there were problems associated with this method. The majority of the variables within the shipwreck dataset were categorical data types. The chi-squared test allows the categorical values like hull type and loss type to be compared. Some of the analysis results

[8] (Petrovi´c n.d.)

were more useful than others. There were three possible outcomes that resulted from the chi-squared analysis. Either high significance, low significance and not applicable. I had hoped that much of the analysis would result in a high significance but many were not applicable. After reviewing the tables of these outputs I realized why they were "NA". The tables were all or mostly unique values. Without any common values there was no relationship between the variables. This happened for a number of variables including when and where it was built and the loss year. Combinations with too many unique values did not correlate. One last attempt I made was to try a chi-squared analysis between two separate features and found that they have to contain the same number of rows. Unsure of how to correlate the two features I used it as a visual reference.[9]

One of the additional analyses I conducted was a k-mean clustering in conjunction with calinski-harabasz and the silhouette index. This is a simple analysis but has a lot of steps which can get confusing. The initial k-means analysis was conducted then the centers were calculated for each cluster. At first I only ran a simple analysis based off one specific cluster value. I then expanded the analysis to test different number of clusters from 2 to 20 to see which resulted in the best fit.

There are some methods I tried that did not prove to be very informative. One such method was the linear model I generated. I did this mainly to see if any of my data variables could correlate. I found that many of the variables I wanted to compare were non-numeric which meant that they could not be used within the linear model. The set of variables that generated a model were gross weight correlated to length and beam. The significance between these values is obvious so it does not explain any additional relationships.

[9] Appendix F

Results

I did a range of different chi-square tests, some proving more significant than others. I will discuss the ones with high significance below. The benefits of using the chi-square test with this dataset is that it has a lot of categorical variables. The highest correlation, with a p-value of 8.671e-07, was between the year built and the hull type. I expected this relationship to have a high correlation. Having proved my expectations I reviewed the table and saw that only a small number of the vessels were built of material other than wood. Given the skewed sample it makes sense that the correlation was so high. If there was a more even distribution of steel and fiberglass ships then the analysis would be more compelling. The dataset would have to include more recent shipwrecks not simple historic shipwrecks. There are some recent wrecks but not as many to offset all the ships from the 1800s that are exclusively made of wood.

Another significant relationship, with a p-value of 0.02452, was between loss type and hull type. As with the previous hull type analysis most are wood so the loss type is pretty evenly distributed. Although if this correlation is associated with the year and loss type table then other statements can be extrapolated. All the abandoned shipwrecks were older ships. Within this dataset the last abandoned ship was in 1937. It is interesting that collisions are common in any year. Collisions are the most common loss type for steel ships. This makes sense because steel ships are more likely to be salvaged or sold for scrap then wooden ships. Owners would be less likely to simply abandon them. There are more strandings in earlier years but they persist up to present day. Of all the loss types the shipwrecks in this dataset are more likely to be stranded.

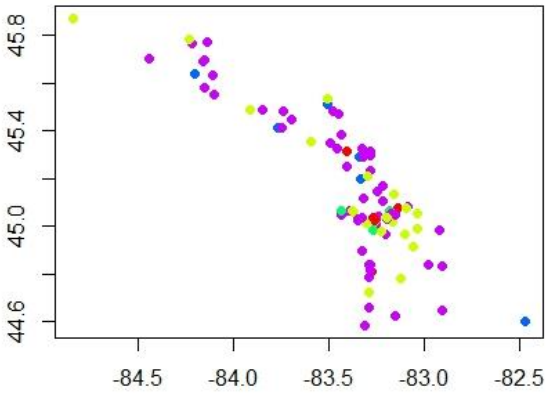
The chi-square test for loss month and loss type returned a p-value of 0.03297. The correlation would have been better if the collision type was also focused in the winter months. Instead, collision is the only type of loss that is not concentrated in the winter months. They are

more evenly distributed across all the months. For the rest, the loss type did not seem to matter as much as the loss month. The losses seemed to happen mostly in the winter months. This correlates to the worst storms associated with lake effect snow. During the winter, large portions of coast are covered in ice which makes traversing the waters very dangerous. Between the ice and the severe storms I expected the most ship casualties would occur during the winter months.

The p-value is 0.01199 for the chi-square test between loss month and the county. The significance is fairly high for this combination. The largest number of shipwrecks occurred in the shallows of Alpena Bay. One of the reasons Alpena County has so many shipwrecks is because it is one of the central port cities along the eastern coast of Michigan. Not only is it a central port it is also the headquarters of the Thunder Bay National Marine Sanctuary. They are in charge of identifying, surveying, and preserving historic shipwrecks. Alpena Bay is the best surveyed area because it was the initial bounds of the Marine Sanctuary. The location of numerous shipwrecks in Alpena Bay are in the shallows anywhere from 15 to 30 feet deep. It is a lot easier to survey shallow wrecks than deep wrecks.

A combination that I thought would be interesting was the correlation between the loss type and the builder. Unfortunately the p-value was high suggesting no significant correlation. If there was more data to evaluate then maybe a correlation could be found. The same was true for the year lost and the loss type. There was no year that had more shipwrecks than others.

The next type of analysis that I did was to test the clustering with k-means and indexing. I started by reordering the table to place all the numerical variables next to each other making them easier to call later. The k-means analysis requires the initial values to be scaled. The initial analysis ran into problems because of missing data values. To correct these errors the na.omit() function was added to remove the missing values from the analysis. The next step was to plot the k-means distribution. The plot does not seem to have any meaningful distribution but the clusters were by no means even. I additionally plotted the centers to see the distribution based off the initial clusters.[10] The next step is to reset the original values and plot the month and depth centers.[11]



The final analysis is to conduct two clustering tests the Calinski-Harabasz and the Silhouette index. This type of analysis tests different numbers of clusters from 2 clusters to 20 clusters to find the best solution. The results of the calinski-harabasz index showed a semi-parabolic shape with no specific peaks.[12] Based on the shape of the plot for the calinski-harabasz index there was no solution that fit best. The silhouette index was different there was a peak at 5 clusters. This means for this distribution 5 clusters will best reflect the data.[13]

Discussion

After running all the analysis I would have liked a more robust dataset. Until I encountered errors trying to run various analytical methods I did not realize how much of the

[10] Appendix A & B
[11] Appendix C
[12] Appendix D
[13] Appendix E

data was incomplete. I have done visualizations of the data in ArcPro and did not notice all the incomplete values within the data. Unfortunately this causes a number of issues when running numerical data analysis. The missing data values cannot simply be ignored unless specifically told to which is why I had to use the "omit na" function. Again this allows the analysis to run but also eliminates a great number of data points in the process. It was frustrating that a lot of the categorical data was unable to be analyzed further. The dataset includes a lot of variables with unique values. If the dataset was larger maybe there would be more cross over with factors like builder and the cargo. Both factors tended to have almost exclusively unique values. Trying to use the analysis we learned from the labs was also a challenge because this dataset has some unique challenges. With only one dataset I was trying to fit analyses to data instead of fitting the data to an analysis.

Conclusion

The analysis produced by the current data set was too small when the unknown values were taken out. There are a couple different ways to solve this problem. The most difficult method would be to fill the missing values of the current dataset. If the data is not filled there was probably no information found for that variable. Maybe additional research could someday fill those missing values but it would no doubt take a lot of time and effort. The other way to expand the dataset would be to include all the shipwrecks throughout the great lakes. The current dataset is only a portion of Lake Huron shipwrecks. This would add a great number of additional values. Adding in all the great lakes would hopefully help generate a more robust model for a more general shipwreck model. Given the addition of other datasets the biggest question would be formatting. If the formatting is identical to the first dataset then there would be no problem combining them. Although I could foresee formatting and even the quality being

different. Even though there are a lot of question in adding in other data it would only help to improve the model.

Bibliography

Brewer, Simon. 2019. "GEOG 6000 Lab07 Multivariate Analysis." *Class notes.*

Desgraupes, Bernard. 2017. "Clustering Indices." *University Paris Ouest.*

2015. *Lake Huron shipwrecks 3D imaged by marine archeologists.* September 10. http://www.cbc.ca/news/technology/lake-huron-shipwrecks-3d-imaged-by-marine-archeologists-1.3219021.

Mangiafico, Salvatore S. 2015. *Fisher's Exact Test of Independence.* Accessed November 2019. https://rcompanion.org/rcompanion/b_07.html.

Petrovi´c, Slobodan. n.d. "A Comparison Between the Silhouette Index and the Davies-Bouldin Index in Labelling IDS Clusters." *Gjøvik University College* 12.

Sanctuaries, NOAA. 2014. "NOAA expands Thunder Bay National Marine Sanctuary." https://sanctuaries.noaa.gov/news/press/2014/pr090514.html.

Sharpe, Donald. 2015. "Your Chi-Square Test is Statistically Significant: Now What?" *Practical Assessment, Research & Evaluation* Vol. 20, 10pg.

Data Sources

Shipwreck Data (John Bright) Note: Given as .shp and converted to .csv

Additional Boundaries from Natural Earth: https://www.naturalearthdata.com/downloads/

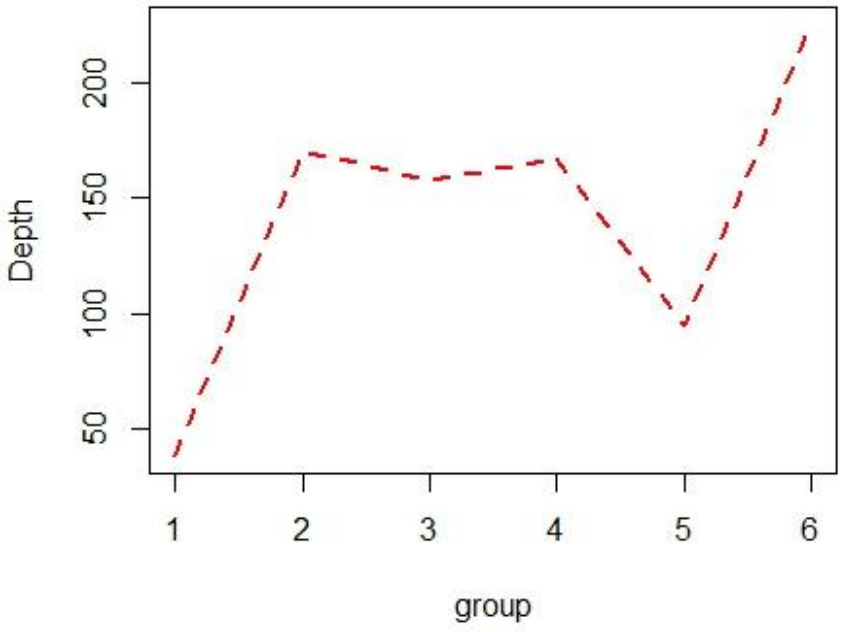
Great Lake Boundary States: http://maps.glin.net/data/635d4cef-cb52-4620-9f02-1197b5c3afd1

Canadian Great Lakes Provinces Boundary: http://maps.glin.net/data/e1fa466e-7baf-40f3-a81b-f278230c16a2

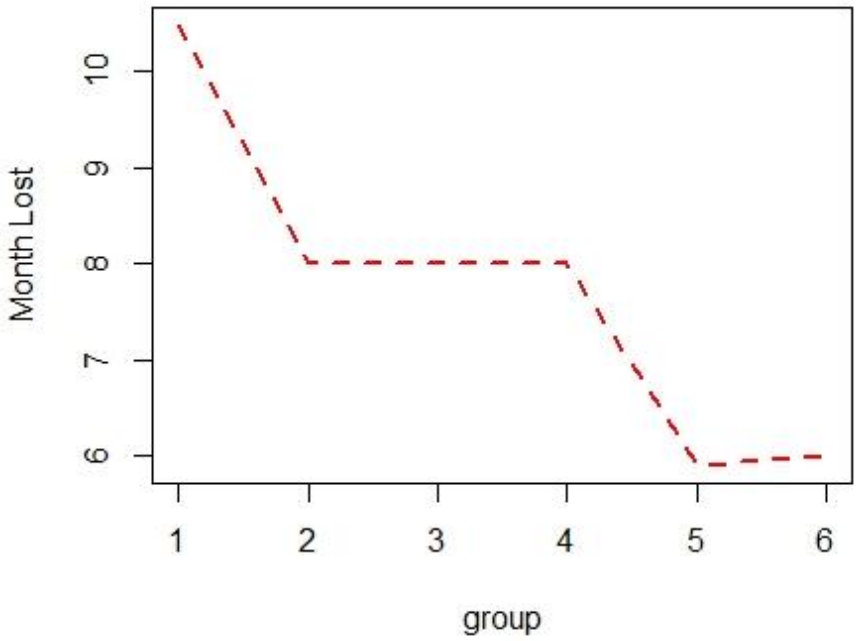
U.S. Maritime Limits & Boundaries: https://www.nauticalcharts.noaa.gov/csdl/mbound.htm

Bathymetry: https://www.ngdc.noaa.gov/mgg/greatlakes/

Appendix A: Depth Centers



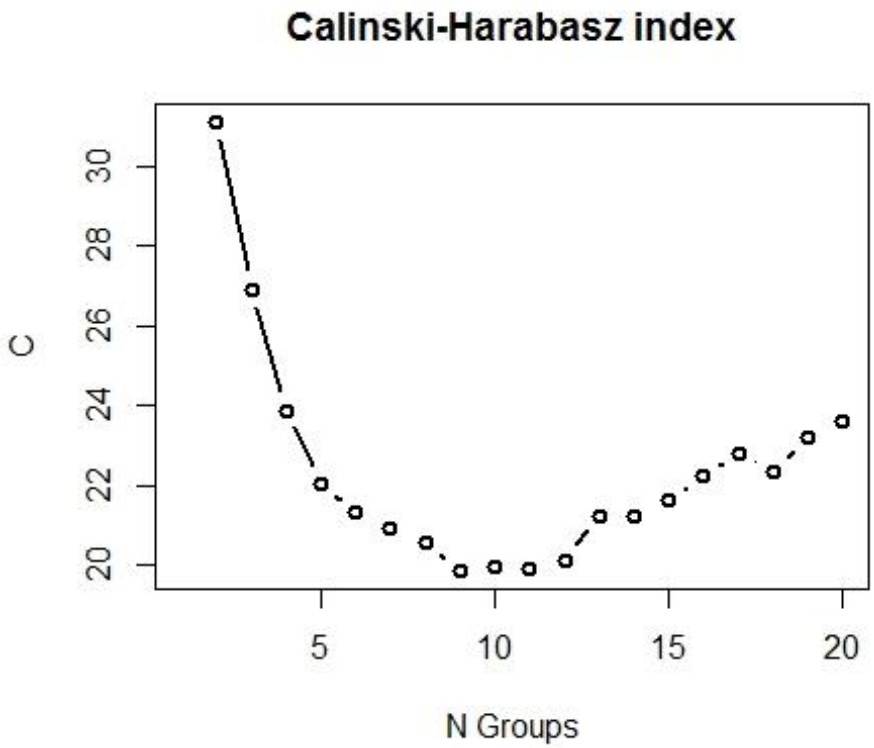
Appendix B: Month Centers



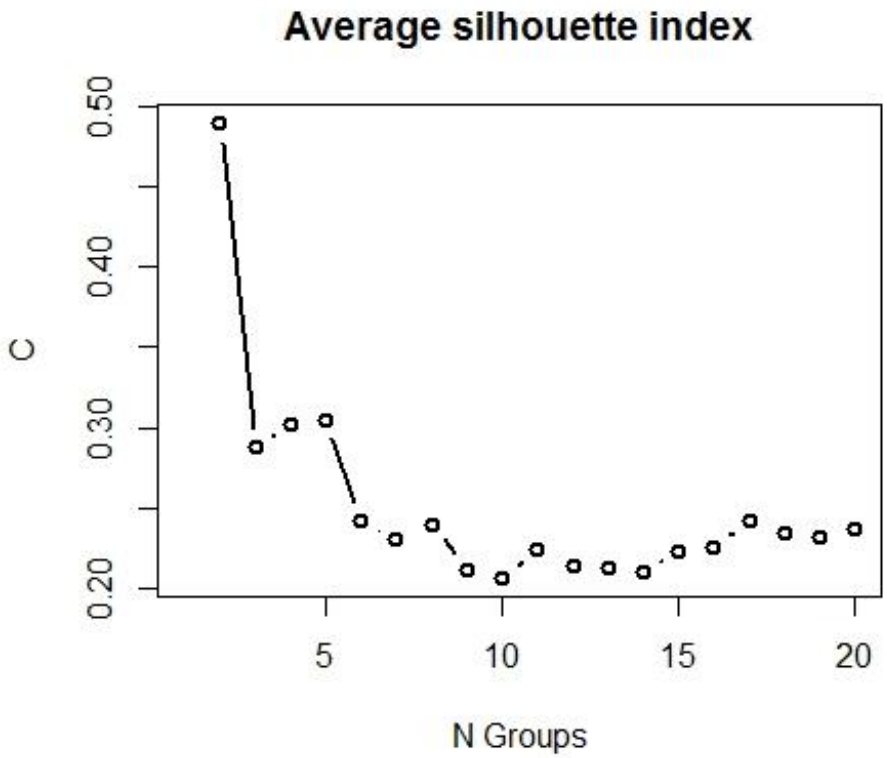
Appendix C: List of Centers with Original values

```
> wk.centers
  Group.1    Built   Length     Beam      DoH     Gross       Net     Lost      Lives     Depth Loss_Month
1       1 1871.217 153.2087 29.40870 11.09565  189.6617  413.5674 1897.913  0.1739130  38.39130  10.478261
2       2 1863.000 198.5000 30.60000 12.50000  979.0700  738.7900 1865.000 35.0000000 170.00000   8.000000
3       3 1907.500 518.0000 55.00000 30.50000 6697.5000 5165.5000 1916.000 14.0000000 157.50000   8.000000
4       4 1888.462 271.8769 39.32308 17.87692 1869.2592 1441.3638 1911.615  0.5384615 166.69231   8.000000
5       5 1870.455 147.6800 28.19636 11.55818  203.4309  372.0555 1897.545  1.9090909  94.90909   5.909091
6       6 1922.500 431.0000 53.00000 25.00000    0.0000 4623.5000 1932.000  0.0000000 225.00000   6.000000
```

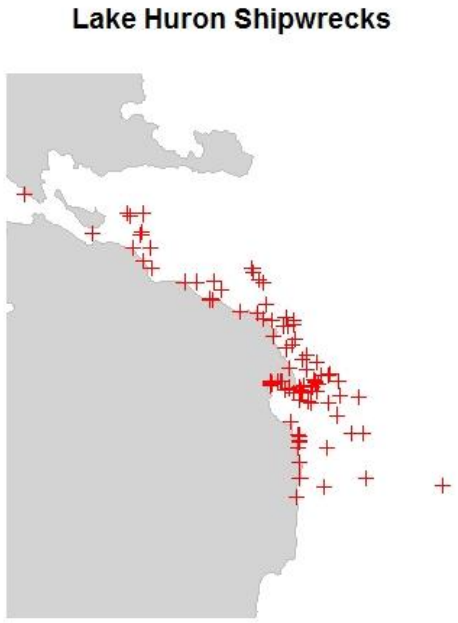
Appendix D: Calinski-Harabasz Index

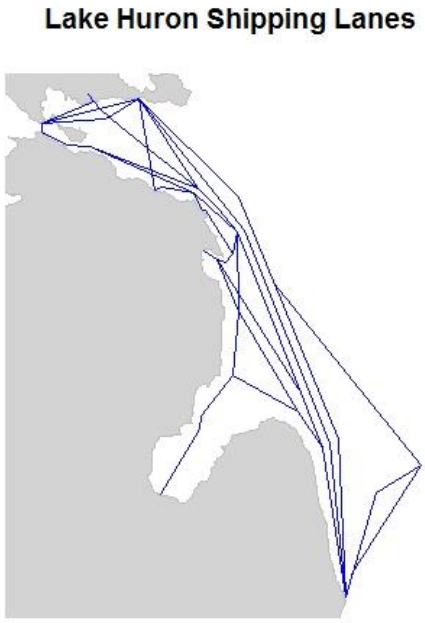


Appendix E: Average Silhouette Index



Appendix F: Overview Maps





Appendix G: Shipwreck Locations