K-Means Clustering of Abalone Populations Using an Incomplete Dataset

Michael Prendergast
mike.prendergast@ieee.org
July 31, 2021

1. Introduction

1.1. Objective

The objective of this report is to investigate the distribution of abalone data collected off of the coast of Tasmania, Australia in 1994 in order to determine whether or not the dataset includes groups which are unusual and possibly indicative of the abalone population collapse that has occurred in Tasmania since the data was collected. As a side effect, some of the abalone population distribution data that was reportedly collected in (**Nash, 1994**) but was not published in either (**Nash, 1994**) nor in the UCI data repository (**Nash, 1995**) has been reconstituted and is made available in the appendix.

1.2. Problem Domain

The Tasmania abalone industry has been declining for years. Abalone harvesting peaked at 4,500 metric tons in 1984. Although a thirty percent quota reduction in 1989 seemed to have had a temporarily stabilizing effect, the downward trend continued starting in 2001. By 2010, the Australian government had imposed a limit of 1,750 metric tons, 61% off of the 1984 high.

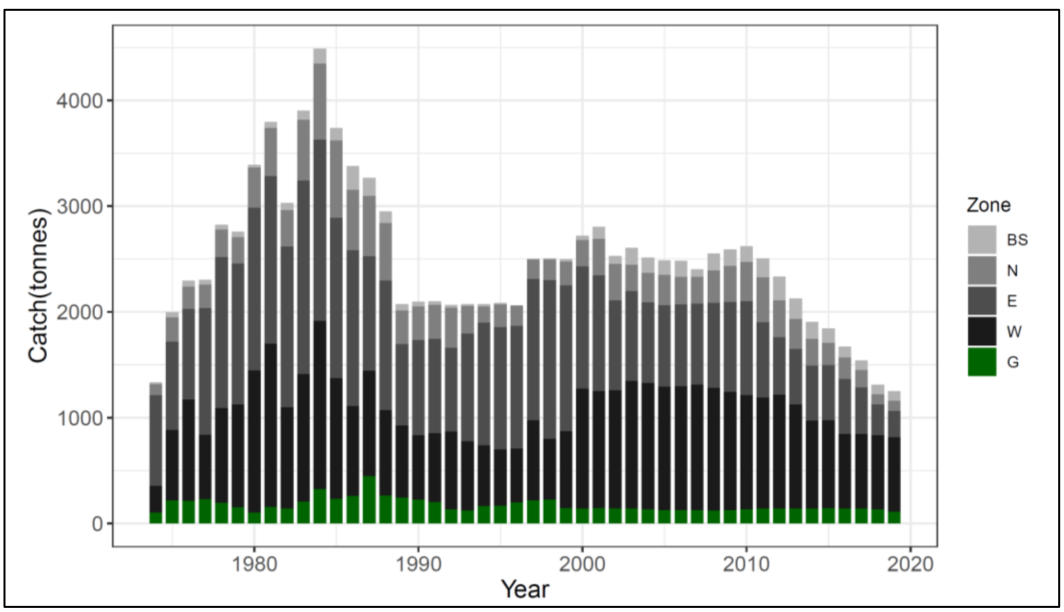

Figure 1: Abalone Harvest from 1974-2019 by Region, from (Mundy, 2020)

Figure 1, taken from (**Mundy, 2020**), shows that even that number was unsustainable, and the effect on commercial abalone divers since then has been devastating. As one diver in 2016 noted, “We've had downturns before but never this bad. It's the worst I've ever seen it.” (**Salmon, 2016**). By 2019, abalone yield was down to 1,250 tons, off 72% from the 1984 high.

Theories on the collapse cause include global warming (**Fears, 2019**), overfishing (**Salmon, 2016**) and the abalone viral ganglioneuritis (AVG) virus, discovered in 2005 (**Corbeil, 2012**).

Researchers have been studying this Tasmanian abalone population for some time. In 1994, the Australian government performed a survey of the population **(Nash, 1994)** to determine what limits should be placed on commercial fisherman. The resulting dataset is (**Nash, 1995**).

When the data was collected, abalone yield had already fallen by more than half, to 2000 metric tons. However, the yield appeared to be stabilizing. In hindsight, this stabilization was short-lived and the dramatic drop in abalone population was about to continue.

Is there anything in the abalone population study data that could have warned scientists that the abalone population collapse was about to resume?

1.3. Method Rationale

As noted earlier, the objective of this study is to look for data characteristics in a 1994 abalone study which might be useful in understanding the Tasmanian abalone population.

This study’s methodology is two-fold. First, a thorough exploration of the abalone dataset is performed to see if the un-clustered data provides interesting insights about abalone population characteristics. Then, a k-means clustering algorithm will be used to identify population clusters. These clusters are then analyzed for additional insight into abalone population traits.

An unsupervised clustering technique such as k-means clustering can be used to identify clusters when the cluster set labels or attributes are not known in advance. Since this study is exploratory and the characteristics that are being sought are not known in advance, the k-means clustering algorithm is appropriate and satisfies this report's purpose.

2. Analysis

2.1. Data

The abalone population dataset from (**Nash, 1995**) was first analyzed in a Ph.D. dissertation (**Waugh, 1995**) and represents 4,177 abalone specimens collected off of Tasmania, Australia. Each specimen is characterized by nine attributes. The abalone were collected in two regions, Bass Strait (1,617 samples) and St. Helens (2,560 samples). Bass Strait abalone were stunted and St. Helens abalone were faster growing (**Waugh, 1995**).

The numeric values in the dataset are scaled by a factor of 200, and the variable *Rings* is a proxy for age, with the age of an abalone in years approximated by: *Age* = 1.5 years + *Rings*.

This dataset was developed by (**Nash, 1995**) and then provided to (**Waugh, 1995**) who pre-processed it and used it in his Ph.D. research. Pre-processing included eliminating sample points with NULL values and removing the Site ID column which he chose not to use. Waugh provided the pre-processed dataset, with Site ID removed, to UC Irvine for the benefit of future researchers. The UC Irvine data repository currently cites 31 papers that reference this dataset.

Table 1: Available Dataset Attributes After Pre-Processing by (Waugh, 1995)

Attribute	Description
Sex	Sex of the abalone, M for adult male, F for adult female, and I for infant,
Length	Length of the abalone shell, measured in 200*millimeters,
Diameter	Diameter of the abalone shell, measured in 200*millimeters,
Height	Height of the abalone shell, measured in 200*millimeters,
Whole_weight	The weight of the abalone, measured in 200*grams,
Shucked_weight	The weight of the shucked abalone meat, in 200*grams,
Viscera_weight	The weight of the abalone gut, in 200*grams,
Shell_weight	The weight of the dried abalone shell, in 200*grams,
Rings	The number of rings found on the abalone shell.

The published dataset attributes are listed in Table 1 above.

However, the site identifier contains potentially useful population information since Bass Strait had stunted abalone and St. Helens had normal abalone (**Waugh**, 1995). To the best of this author's knowledge, the deleted site data has been unavailable since the dataset was first published in 1995. However, it has been successfully recovered and added back in this study.

2.2. Exploratory Analysis

Examining the dissertation (**Waugh, 1995**) yields several clues to help reconstruct Site ID data, provided in the appendix. Table 10 (Appendix) shows the dataset ring distribution by site (plus 4 rows, later discarded, for de-sexed abalone). Table 11 includes the minimum and maximum of each of the attributes by site. Table 12 shows the population by sex for each site.

Table 13 (Appendix) lists the 46 Bass strait abalone data constraints extracted from Tables 10-12. Any mapping of data to a SiteID of Bass Strait must meet all 46 of these constraints.

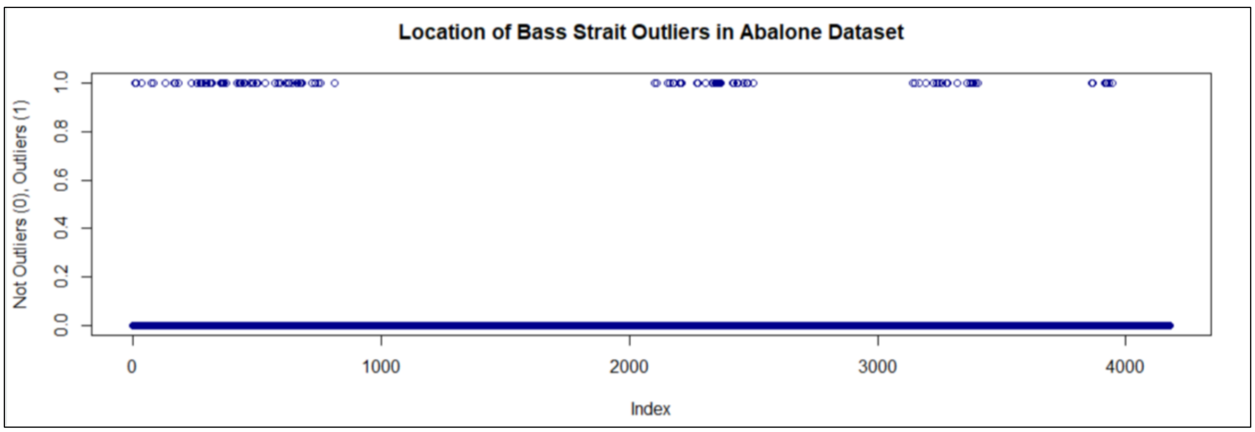

Figure 2: Bass Strait Outliers are in Four Clusters, Giving a Clue to Overall Distribution

From Table 10, one sees that only the Bass Strait site yielded abalone with 2 or fewer rings or more than 18 rings. Figure 2 plots the dataset indices of these outliers. All Bass Strait outliers are found in the following dataset index bands: [7-812], [2102-2500], [3141-3402] and [3864-3945].

Using the hypothesis that the dataset was clustered into four index bands that were supersets of the bands above, a search was made to find a SiteID mapping that included the bands above and met all 46 of the constraints discovered and listed in Table 13.

The thesis also states that the data was split into a training set of 3133 rows and a test set of 1044 rows. From this, the hypothesis was made that the first band starts at index 1, and the third band starts at 3134. After extensive trial and error testing against all constraints, a SiteID mapping was found that *exactly matched all 46 constraints*. The Site ID is Bass Strait for abalone in [1-812], [2092-2502], [3134-3405] or [3831-3952]. Others are from St Helens.

The site identifier column, deleted 25 years ago, has been recovered.

The author is unable to prove that this is the *only* mapping that meets all constraints, though it is the only one discovered after thousands of attempts. However, this mapping exactly satisfied each constraint and fits the (**Waugh, 1995**) training/test split, and so its distribution should be similar to the original Site ID field. Hence it may be useful in studying abalone populations.

But does it actually contain useful discriminating information about the abalone population? Potentially yes. Figures 3 shows the length and ring count for Bass Strait and St Helens abalone. These figures have different shapes, and the abalone from Bass Strait appear to be smaller ("stunted"). Only Bass Strait has abalone with 1, 2, 18, or more than 18 rings.

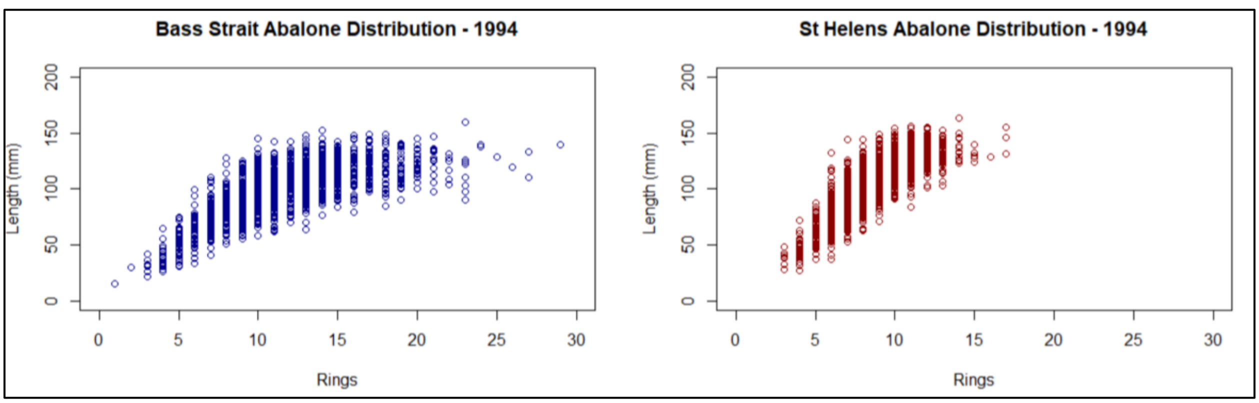

Figure 3: Abalone distribution by Ring count and Length at Bass Strait, 1994.

Also, from these diagrams one can see that neither site had many abalone longer than 150 mm in length. This may be partly due to the fact that, at the time this data was collected, the minimum size limit on Abalone in Tasmania was 132 mm (**Nash, 1994**).

Turning now to the other variables, Table 2 shows cross-correlations of attributes with *Length* and *Whole_weight*. All of the weight variables are very highly correlated. Similarly, the dimension variables (*Length*, *Diameter*, *Height*) are also correlated, although *Height* a little less so. *Height*, being smaller than *Length* and *Diameter*, may be measured less precisely. There are also two samples in the dataset with a Height of 0.

Table 2: Cross-Correlations for Weight and Size Variables

Cross-Correlations					
Correlations to *Length*		Correlation	Correlations to *Whole_weight*		Correlation
Length	*Diameter*	0.987	*Whole_weight*	*Shucked_weight*	0.969
Length	*Height*	0.828	*Whole_weight*	*Viscera_weight*	0.966
Length	*Whole_weight*	0.925	*Whole_weight*	*Shell_weight*	0.955
Length	*Rings*	0.557	*Whole_weight*	*Shucked_weight +Viscera_weigh t+Shell_weight*	0.995
			Whole_weight	*Length*Height*Diameter*	0.960

2.3. Pre-Processing

The reconstructed Site ID described in the previous section was added to the dataset. It was set to 1 if the sample came from Bass Strait, and to 0 if it came from St. Helens.

Because of the high correlation between all of the size and weight variables, all but one of the size and weight variables were removed. The one that was retained was *Length*. It was selected because abalone fishing restrictions in Tasmania are based on the length of the abalone.

No sample points were removed. The two points with a *Height* of 0.0 were retained because their other attribute values appeared reasonable, and also *Height* itself had been removed.

The categorical variable *Sex* was removed and replaced with two numeric variables, *SexM* (set to 1 if the sex was male, 0 otherwise) and *SexF* (set to 1 if the sex was female). If both *SexM* and *SexF* are set to 0, the data sample represents an immature abalone.

After pre-processing, five variables remained: *SexM*, *SexF*, *Whole_weight*, *Rings* and *Site*.

2.4. Algorithm Intuition

The expectation was that the k-means algorithm would cluster appropriately based on the five variables listed above. Since longer-lived abalone occur more often in Bass Strait, intuition suggested longer-lived Bass Strait abalone may represent one of the clusters.

Since Sex represents two of the five variables (*SexF* and *SexM)*, it was thought that sex might also place a dominant role in the clustering.

The optimum number of clusters is determined during the analysis that follows. However, intuitively, it was thought that there might be four cluster sets, one for older abalone in Bass Strait, two for adult abalone based separated by length or sex or both, and one for infant abalone.

2.5. Model Fitting

Clustering models were fit against the data using k-means clustering for k sizes of 2, 3, 4, 5 and 6. The results of this analysis are shown in Table 3 below and plotted in Figure 4.

Table 3: Summary of K-Means Results for Different Values of K

K	Number of instances in each cluster	Between clusters sum of squares	Within clusters sum of squares	Number of iterations
2	2649, 1528	5661.14	15218.86	3
3	1307, 993, 1877	9344.18	11535.82	11
4	1295, 752, 1240, 890	12977.41	7902.59	12
5	622, 752, 683, 890, 1230	14728.65	6151.35	14
6	570, 608, 683, 875, 1002, 439	16023.68	4856.32	32

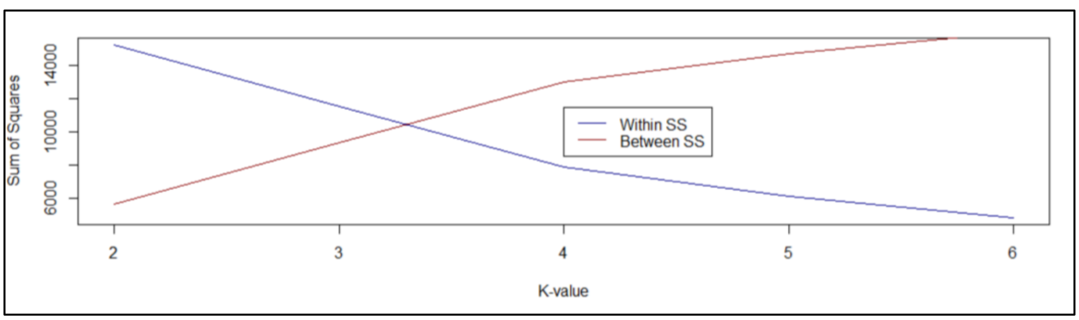

Figure 4: Within-Cluster and Between-Cluster Sum of Squares for Varying K values

From Figure 4, the sum of squares data has an inflection points at k=4 and 5, so these might be good choices for k. A test using the optimum average silhouette width criteria (R function pamk) found k=5 to be the best fit. Figure 5 depicts a cluster plot with k=5. The data appears to fall into 3 clusters, although close inspection shows that each of the three clustered points are actually two parallel lines. Hence, k=3, 4, 5 or 6 all are possible. Given that the within-sum of squares value drop off for k=6 was small, k=5 was chosen for the final model.

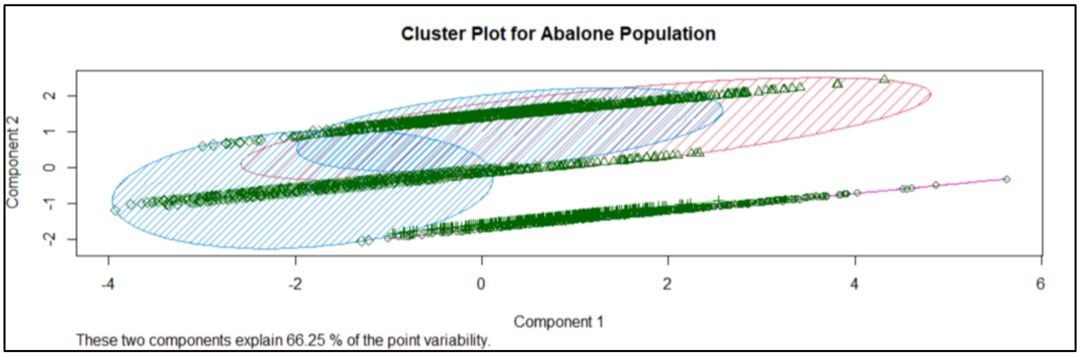

Figure 5: Cluster Plot for Abalone Population Using K-Means Clustering with K=5.

3. Results

3.1. Output

Our objective was to identify abalone dataset characteristics and clusters that can be useful in understanding Tasmanian the abalone populations. This objective was satisfied – both with this study's data exploration activities as well as with the discovered population clusters.

During data exploration is was discovered that there were no abalone three years or younger, or more than 18 years old from St. Helens. The lack of young abalone from St. Helens may be a statistical anomaly or a biproduct of the data collection process. However, young abalone have a number of predators, including crabs, lobsters, starfish and octopus (**Fishtech, 2020**). It would be useful to know whether or not any of these invasive species has infiltrated the abalone beds.

The lack of older abalone at St. Helens may be due to them being harvested, while the older abalone in Bass Strait, being stunted, are less likely to reach the minimum 132 mm size limit.

Abalone size is very important because larger females can produce many more eggs – possibly a million or more. Hence, a certain stock of larger female abalone may be necessary to maintain larger populations. Increasing the abalone size limits would increase the population of large females, leading to more eggs and possibly more abalone.

Other interesting information came from data clustering. The clustering algorithm identified five clusters which had the following overall features:

Table 4: Characteristics of the Five Abalone Clusters

Cluster	Count	Site ID	Rings	Sex	Length (cm)
1	622	All Bass Strait	Adult to Aged (5-29)	All Female	58-160
2	752	All Bass Strait	Adult to Aged (5-27)	Mostly Male	41-152
3	682	All St Helens	Adult (6-16)	All Female	75-163
4	890	All St Helens	Adult (6-17)	98% Male	77-156
5	1230	Mostly St Helens	Young (1-13)	98% Immature	15-126

Mature Adult Abalone from Bass Strait are Smaller (Clusters 1 and 2)

Bass Strait is known to host stunted abalones (**Waugh, 1995**), and we see this in Table 4, where the clusters 1 and 2 minimum lengths are smaller than for cluster 3 and 4.

There are a number of theories about what causes stunted abalone growth, including boring worms (**McDiarmid, 2004**), lack of oxygen (**Boch, 2018**), and an inadequate food supply (**Saunders, 2009**). Stunted abalones cannot usually be harvested, because they are too small. Solving the problem of stunted abalone would be a boon to the abalone fishermen.

Observation 2: Some older abalone are still sexually immature (Cluster 5)

According to (**United Nations, 1990**), lack of adequate food supply or cold habitats can prevent abalones from maturing. A study to determine if either of these conditions are present could increase the stock of sexually mature abalone, resulting in increased abalone populations.

Stunted abalone, immature abalone, and harvesting size limits all impact the Tasmanian abalone industry. Perhaps one or more of them holds the key to understanding why it is dying.

3.2. Model Properties

The cluster model characteristics can be found in Table 5 below.

Table 5: Abalone Attribute Variation by Cluster

	Length (mm)				Rings				Site	SexM	SexF	Sex Infant	Total
Cluster	Min	Mean	Max	Standard Dev.	Min	Mean	Max	Standard Dev.					
1	58	107.8	160	17.58	5	12.6	29	3.66	1	0	1	0	622
2	41	102.7	152	19.61	5	12.41	27	3.58	1	0.843	0	0.157	752
3	75	123.3	163	12.7	6	9.8	16	1.54	0	0	1	0	682
4	77	121.1	156	13.62	6	9.77	17	1.6	0	0.982	0	0.018	890
5	15	82.49	126	21.39	1	7.26	13	1.73	0.198	0.016	0.002	0.982	1230

The sum of squares components of the final fitted model can be found in Table 6 below.

Table 6: Sum of Squares Components for the Final Fitted Model

Within SS by Cluster					Total Within SS	Between SS	Total SS
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5			
1134.85	1855.36	346.62	571	2243.51	6151.34	14728.65	20879.99

Between Sum of Squares is 70.5% of total sum of squares. From Figure 5, 66.3% of the total point variability is described by the first two principal components of the clustering model.

3.3. Model Evaluation

The unsupervised classification model did differentiate the categorical variables (Site, SexF and SexM) well, but was not as effective at differentiating the numerical variables.

For example, the mapping between clusters and sites and between male abalone and sites is in Table 7 below. Here we see that assigning clusters 1 and 2 to Bass Strait, and the other clusters to St. Helens would result in an error rate of only 243/4777, or 5.1% (an accuracy of 94.9%).

Table 7: By Cluster Counts of Site ID, Male Population and Female Population

Site \ Cluster	1	2	3	4	5
0	0	0	683	890	987
1	622	752	0	0	243

SexM \ Cluster	1	2	3	4	5
0	622	118	683	16	1210
1	0	634	0	874	20

SexF \ Cluster	1	2	3	4	5
0	0	752	0	890	1228
1	622	0	683	0	2

Similarly, an assignment of “Male” to clusters 2 and 4 and “not Male” to the other clusters is accurate 96.8% of the time. Assigning clusters 1 and 3 results to “Female” in a very high accuracy of 99.96%. There are only 2 misclassifications, all in Cluster 5.

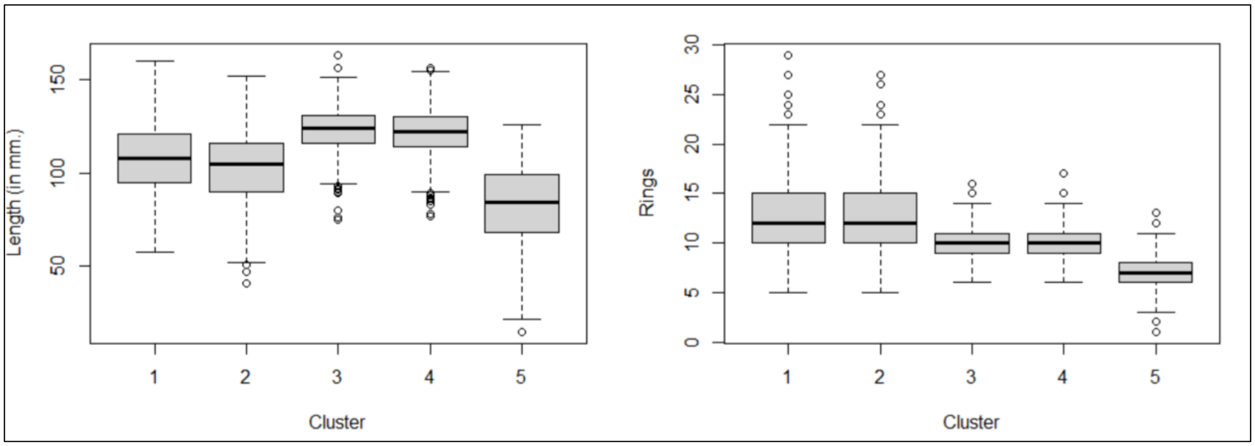

Figure 6: Length and Ring Count Distribution by Cluster

The separation by cluster along the other two attributes is not as clear, however. From Figure 6, clusters 1 and 2 do appear smaller and older then cluster 3 and 4. However, all clusters have numerous abalone from 7 to 11 rings, and the ring count distribution in cluster groups (1,2) and (3,4) are almost identical. All clusters also have abalone in the range of 90 mm. to 120 mm.

However, the model does differentiate abalone by a combination of their *site* and *length* attributes. Note that the abalone in clusters 1 and 2, all from Bass Strait, are 10-20% smaller on average than those in Clusters 3 and 4 (from St. Helens), despite being an older population. This confirms the assertion that the abalone from Bass Strait are stunted (**Waugh, 1995**).

Another interesting combination is sex and *Rings*. Cluster 5, the largest cluster, is comprised of sexually immature abalone with a mean ring count of 7.26 (which equals 8.76 years old). What is surprising about this is that the age of abalone sexual maturity is reported to be 4 to 7 years (**Prince, 1988**). Yet this data indicates that there are a large number of abalone at both sites that were over 7 years of age but had not yet reached sexual maturity.

In summary, the fitted model differentiates samples by categorical variables well. It also provides interesting differentiation of samples using variable combinations of site-length and sex-age. One site has smaller but longer-lived "stunted" abalone. Both sites have a surprisingly large counts of abalone that have not reached sexual maturity, yet from their ages should have.

4. Conclusion

4.1. Summary

This report investigated abalone population data at two sites in Tasmania, Australia from 1994. For those involved in the harvesting, packaging, distribution and sales of abalone in Tasmania's abalone industry, this topic is of the utmost importance. The abalone industry, though still valued at over $100 million dollars, has been collapsing over the last 35 years.

The original abalone dataset was analyzed by (**Waugh, 1995**) for his PhD thesis, and he removed the harvest site information from the dataset before publishing it, as this information was not relevant to his research. Believing that site data would be useful in understanding abalone demographics, this study successfully reconstructed the Site ID data that had been deleted 35 years ago.

The results from adding the site information back in and the follow-on data clustering were illuminating. As noted by (**Waugh, 1995**), Bass Island abalone were stunted, not growing to a harvestable size – a fact confirmed by the enhanced dataset. Bass Island adult abalone were typically 10-20% smaller than St Helens adult abalone. Also, only the Bass Strait abalone lived for 20 years, probably because few ever reached the minimum harvesting size of 132 mm.

Clustering helped identify another anomaly. Both sites (Bass Strait and St. Helens) had many immature abalone much older than 7 years, a typical age for abalone maturity.

The underlying causes of these observations are unknown. Suspects for the stunted abalone issue include boring worms, lack of oxygen in the water, lack of food supply, and global warming (**McDiarmid, 2004**), (**Boch, 2018**), (**Saunders, 2009**). Late sexual immaturity can be caused by water temperature or inadequate food supply (**United Nations, 1990**) or possibly other factors as well.

Perhaps one of these findings holds a key to understanding the Tasmania abalone collapse.

4.2. Study Limitations

This analysis was based on abalone data collected in Tasmania, Australia during 1994. The results obtained are only applicable to those regions, and only for the 1994 timeframe.

The samples used in this study may not be indicative of the actual population distribution of abalone at the two collection sites. For example, the collection process may have resulted in the collection only of specimens of a certain size or weight. Also, as some abalone are known to hide (**Fishtech, 2020**), abalone in hiding may be underrepresented in the sample dataset.

This report used an unsupervised clustering algorithm (K-means) to group the 1995 abalone samples into clusters. This study does not try to explain why the abalone populations fell into these clusters, however. For example, although the population of abalone in Bass Strait is stunted, and this study confirms that fact, it does not definitively explain why the population is stunted.

Since this study performed unsupervised classification, the population clusters identified may not all be significant or may not all be important for the study of Tasmanian abalone populations.

4.3. Improvement Areas

To better understand the abalone situation in Tasmania today, another sample collection at multiple abalone harvesting sites is needed. This collection should include representative samples of various sizes and ages. It should also include estimates of juveniles too small to collect. To observe changes over time, samples should be collected regularly, such as every year.

This study used unsupervised learning to identify population clusters. With proper domain expertise, semi-supervised learning can be applied to ensure that, at a minimum, important abalone population subgroups, called out by shellfish domain experts, are separately clustered.

These experts can identify important population groups using a smaller sample set, and a clustering algorithm can be applied to find other fish that are like ("close to") the representative samples.

5. References

C. A. Boch, et. al. (2018). Local oceanographic variability influences the performance of juvenile abalone under climate change. Sci Rep. 2018; 8: 5501. Published online 2018 Apr 3. doi: 10.1038/s41598-018-23746-z. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5882807/

Corbeil, S., et. al. (2012). Abalone viral ganglioneuritis: Establishment and use of an experimental immersion challenge system for the study of abalone herpes virus infections in Australian abalone. Virus Research, Volume 165, Issue 2, May 2012, Pages 207-213. Abstract retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0168170212000652?via%3Dihub.

Fears, D. (December 27, 2019). 2°C: BEYOND THE LIMIT. The Washington Post. Retrieved from https://www.washingtonpost.com/graphics/2019/world/climate-environment/climate-change-tasmania/

Fishtech. (2020). Facts About Abalone. Retrieved from http://www.fishtech.com/facts.html.

McDiarmid, H., R. Day, and R. Wilson. (2004). The Ecology of Polychaetes that Infest Abalone Shells in Victoria, Australia. Journal of Shellfish Research, Vol. 23, No. 4, 1179–1188, 2004. Downloaded from https://www.researchgate.net/publication/249967992_The_ecology_of_polychaetes_that_infest_abalone_shells_in_Victoria_Australia

Mundy, C. and Jaime McAllister. (2020). Tasmanian abalone fishery assessment 2019. Institute for Marine and Antarctic Studies, University of Tasmania. Retrieved from https://www.imas.utas.edu.au/__data/assets/pdf_file/0017/1342061/AbAssessFY2019Final.pdf

Nash, W., et. al. (1994). The Population Biology of Abalone (Haliotis species) in Tasmania. Technical Report Number 48. Sea Fisheries Division, Marine Research Laboratories – Taroona, Department of Primary Industry and Fisheries, Tasmania. Retrieved from https://www.researchgate.net/publication/287546509_7he_Population_Biology_of_Abalone_Haliotis_species_in_Tasmania_I_Blacklip_Abalone_H_rubra_from_the_North_Coast_and_Islands_of_Bass_Strait

Nash, W. (1995). Abalone Data Set. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Prince, J.D., Sellers, T.L., Ford, W.B. et al. Recruitment, growth, mortality and population structure in a southern Australian population of Haliotis rubra (Mollusca: Gastropoda). Marine Biology 100, 75–82 (1988). https://doi.org/10.1007/BF00392957

Salmon, G. (May 1, 2016). Tasmania's abalone industry overfished and heading for collapse, veteran divers warn. Australian Broadcasting Corporation. Retrieved from https://www.abc.net.au/news/2016-05-02/tasmania-abalone-industry-heading-for-collapse-say-divers/7374966

Saunders, T.M., S. D. Connell & S. Mayfield. (March 11, 2009). Differences in abalone growth and morphology between locations with high and low food availability: morphologically fixed or plastic traits? Marine Biology, volume 156, pages 1255–1263 (2009). Abstract retrieved from https://link.springer.com/article/10.1007/s00227-009-1167-4.

United Nations. (1990). Training Manual on Artificial Breeding of Abalone (Haliotis discus hannai) in Korea DPR, Part 1, Chapter 1. Food and Agriculture Organization of the United Nations. Retrieved from http://www.fao.org/3/AB731E/AB731E01.htm.

Waugh, S. (August, 1995). Extensions to the Cascade-Correlation architecture and benchmarking of feed-forward supervised artificial neural networks. Ph.D. Dissertation in Computer Science. University of Tasmania. Retrieved from https://eprints.utas.edu.au/21965/.

6. Appendix

The following table, built using data from (Waugh, 1995), represents the distribution of abalone by ring count off of Bass Strait and St Helens in Tasmania, Australia. Counts include 4 trematode (de-sexed) abalone from Bass Strait and 22 trematode abalone from St Helens, which have been removed from the dataset.

Table 8 : Abalone Distribution by Ring Count by Site

Number of Rings	Bass Strait	St Helens
1	1	N/A
2	1	N/A
3	8	7
4	24	33
5	49	67
6	60	199
7	96	297
8	101	471
9	165	530
10	185	451
11	155	336
12	153	116
13	155	50
14	113	15
15	97	6
16	66	1
17	55	3
18	42	N/A
19	33	N/A
20	26	N/A
21	14	N/A
22	6	N/A
23	9	N/A
24	2	N/A
25	1	N/A
26	1	N/A
27	2	N/A
28	0	N/A
29	1	N/A

The following table, built using data from (Waugh, 1995), represents the distribution of abalone in the abalone dataset by sex. This includes trematodes, which were removed from the dataset.

Table 9: Distribution of Abalone Sex by Site

Sex	Bass Strait	St Helens	Total
Male	653	875	1528
Female	624	683	1307
Infant	340	1002	1342
Trematode	4	22	26

The following table, built using data from (**Waugh**, 1995), includes the minimum and maximum values for each of the numeric attributes in the abalone dataset. Data in the dataset has been scaled down by a factor of 200, so this data needs to be divided by 200 to match the data in the dataset.

Table 10: Minimum and Maximum Numeric Values in Abalone dataset by site

Attribute	Bass Strait		St Helens	
	Min	Max	Min	Max
Length	15	160	27	163
Diameter	11	126	21	130
Height	2	50	0	226
Whole_weight	0.4	510	2.8	565.1
Shucked_weight	0.2	214.1	1.1	297.6
Viscera_weight	0.1	118	0.1	152
Shell_weight	0.3	201	0.8	179.4

The following table lists the 46 Bass Strait constraints that each proposed mapping of abalone samples to sites must satisfy, together with how the selected mapping met those constraints. Some constraints are denoted by an asterisk (*). These constraints have an ambiguity of up to 3 because the (**Waugh, 1995**) PhD dissertation analysis deleted four data points during pre-processing, and the ring count of three of them could not be determined from the dissertation.

Table 11: Bass Strait Constraints Satisfaction with Selected Mapping

Constraint	Bass Strait Site Abalone Distribution Constraints	Reconstituted Values
1	Exactly 1 abalone with 1 ring.	1
2	Exactly 1 abalone with 2 rings.	1
3	Exactly 8 abalone with 3 rings.	8
4	Exactly 24 abalone with 4 rings.	24
5	46, 47, 48 or 49 abalone with 5 rings. (*)	49
6	Exactly 60 abalone with 6 rings.	60
7	94, 95 or 96 abalone with 7 rings. (*)	96
8	98, 99, 100 or 101 abalone with 8 rings. (*)	100
9	162, 163, 164 or 165 abalone with 9 rings. (*)	165
10	182, 183, 184 or 185 abalone with 10 rings. (*)	185
11	152, 153, 154 or 155 abalone with 11 rings. (*)	154
12	150, 151, 152 or 153 abalone with 12 rings. (*)	153
13	152, 153, 154 or 155 abalone with 13 rings. (*)	155
14	110, 111, 112 or 113 abalone with 14 rings. (*)	112
15	Exactly 97 abalone with 15 rings.	97
16	Exactly 66 abalone with 16 rings.	66
17	Exactly 55 abalone with 17 rings.	55
18	Exactly 42 abalone with 18 rings.	42
19	Exactly 32 abalone with 19 rings.	32
20	Exactly 26 abalone with 20 rings.	26
21	Exactly 14 abalone with 21 rings.	14
22	Exactly 6 abalone with 22 rings.	6
23	Exactly 9 abalone with 23 rings.	9
24	Exactly 2 abalone with 24 rings.	2
25	Exactly 1 abalone with 25 rings.	1
26	Exactly 1 abalone with 26 rings.	1
27	Exactly 2 abalone with 27 rings.	2
28	Exactly 0 abalone with 28 rings.	0
29	Exactly 1 abalone with 29 rings.	1
30	Exactly 653 male abalone.	653
31	Exactly 624 female abalone	624
32	Exactly 340 infant abalone	340
33	Minimum Length of 15 mm	15
34	Maximum Length of 160 mm	160
35	Minimum Diameter of 11 mm	11
36	Maximum Diameter of 126 mm	126
37	Minimum Height of 2 mm	2
38	Maximum Height of 50 mm	50
39	Minimum Whole weight of .4 grams	0.4
40	Maximum Whole weight of 510 grams	510
41	Minimum Shucked weight of .2 grams	0.2
42	Maximum Shucked weight of 214.1 grams	214.1
43	Minimum Viscera weight of .1 grams	0.1
44	Maximum Viscera weight of 118 grams	118
45	Minimum Shell weight of .3 grams	0.3
46	Maximum Shell weight of 201 grams	201