Modeling the Oceanic Effects of Climate Change on Bivalves' Reproductive Behavior

Kenneth K. Choi

ABSTRACT

Global warming is a common topic of discussion, but the impact it has on ocean organisms is drastic and damaging without much media coverage. As temperatures begin to rise and fall drastically, ecological functions within these ecosystems degrade and change. To better understand and investigate this damage on bivalves, this paper investigates the reproductive dynamics between bivalves, taxonomic classification, and environmental variables using 435 scientific articles encompassing 542 entries. This study aims to model and learn the distinct differences in bivalve behaviors to better understand the effects of climate change. We found notable correlations between temperature and chlorophyll concentrations, particularly in cooler climates, while salinity demonstrated more modest associations with reproductive behavior. The study also examined the variability in spawning behavior across taxonomic families, prompting questions on the taxonomic framework's relativity to spawning patterns. By integrating these environmental data and taxonomic classifications, the study develops predictive models capable of forecasting spawn lengths with approximately 50% accuracy. We created another predictive model predicting the family each species belongs to performing with about a 31.579% accuracy rating. Using this model as a foundation, accurate predictions regarding the survivability of specific bivalve species are within reach. The potential to identify bivalve types with limited data presents another promising avenue.

KEYWORDS

random forests, temperature, chlorophyll, salinity, taxonomic family

INTRODUCTION

As of 2020, the United States alone released 5,222 million metric tons of carbon dioxide into the atmosphere, and this number continues to increase causing a greenhouse effect (US EPA, 2022). This radiation increases the ocean temperatures by "0.062 ± 0.013°C per decade over the last 120 years (1900–2019)" while the last decade (2010–2019) … has been 4.5 times higher than the long-term mean (Garcia-Soto et al 2021). Such a drastic increase in temperature within the past 10 years exemplifies the acceleration of the change in temperature in the ocean which dramatically affects biotic and abiotic factors. For instance, the excessive amount of carbon dioxide released into the atmosphere causes the carbon dioxide and water molecules to react creating hydrogen and carbonic acid in the ocean (Guinotte and Fabry 2008). This reaction increases the amount of hydrionium ions in the water ultimately decreasing the pH while damaging organisms that rely on the Hydrodium Gradient structure. Another example is that rising temperatures have increased plankton metabolic rates creating imbalances in photosynthesis and respiration rates in the ocean (Lewandowska et al 2014). Large fluctuations in the amount of plankton present in the water then change oxygen levels damaging the wildlife by increasing the organic matter, oxygen levels, and salinity (Lewandowska et al 2014). This destruction and restructuring of the environment pose grave threats to all species coexisting in the ocean as many cannot adapt to these new changes. Among these organisms, one of the most vulnerable classes is known as *Bivalvia*.

Bivalvia are severely impacted by rising ocean temperatures and are sensitive to low-oxygen environments (Li et al. 2019) making them a good indicator of ecosystem health. Bivaliva encompasses thousands of organisms that typically have two hard shells with some exceptions (UC Berkeley 2001) and are usually known as oysters, clams, and mussels. Due to their suspension feeding and broadcast spawning tendencies, bivalves need water to be a stable medium for the success of their species. For this reason, bivalves in drastically changing environments will suffer major population losses and biological changes. Such biologically damaging changes are already being observed today. In fact, multiple studies along the West Coast suggest the mussels (*Mytilus edulis*) showed signs of weakening shells when exposed to low-pH water often found near human-inhabited coasts. On top of this, shell creation was significantly slowed down in juvenile mussels causing more vulnerability to predators and underdeveloped weak shells when they reached adulthood (Zhao et al 2020). Such challenges will impact their survivability while damaging

ecosystem food chains and the environment in the long run. Not only will the ocean lose 50 gallons of filtered water a day per clams lost, but runoff nutrition filtered by bivalves will cost $2.8-5.8 million a year when replaced with mechanical systems according to studies in Greenwich (Fisheries 2021). With a 1 billion dollar valuation in 2011, the market for bivalves is essential to consumers while tens of thousands of lives rely on the economic impacts as well as food sources that bivalves provide (US DoC, 2023). With bivalves being such an important species to the world, there is not much understanding of how climate change and ocean interactions affect bivalves' reproduction and survival. Most importantly, bivalves serve an important purpose as indicator species. They help scientists and biologists track bivalve behavior to better understand the health of the ecosystem due to their heightened sensitivity and reliance on the conditions of the water surrounding them. For instance, the use of razor clams to detect "changes in tissue and shell growth can be linked to increased temperature, pollution, ocean acidification, and changing nutrients in the ocean waters" (USNPS) shows how sensitive and crucial these organisms are to the monitoring of health in an oceanic environment.

The class *Bivalvia* consists of thousands of organisms ranging from sea slug-like clams to everyday hard-shell clams (*Mercenaria mercenaria*). With each organism having different environmental needs and nutrition, the collection and organization of such data will take decades to assess properly. To combat this issue, looking at spawning and reproduction displays the health of a species and whether or not changes in environmental factors have damaged the species' opportunity to reproduce. In bivalves, spawning happens in waves with gametes being released into the ocean using water as a medium (Cárdenas and Aranda 2000). For these reasons, being located near others and releasing gametes in certain periods is essential for the survival of the species. Therefore, the change in water temperature, pH, salinity, and plankton count can affect a bivalve's reproductive cycle and decision to release its' gametes. To better understand if bivalves are being affected by these factors, data collection on species, spawning period, spawn peak, location, salinity, chlorophyll-a, and study site. Using this data, creating correlations between bivalve species will help establish the current understanding of the different species and find more data to better understand the organisms.

To observe and see changes in bivalve reproductive patterns, we conducted a comprehensive data analysis, as well as a model on numerous bivalves throughout the world, to see the patterns of the changes in the class *Bivalvia*. The central research question was: Does

climate change affect bivalve behavior reproduction predictably? I predicted that climate change would affect bivalves to have changes in the period and frequency of spawning. Specifically, the first subquestion was does the bivalve reproduction period shift with temperature changes? Based on current information, my prediction was that the periods will shift later into the year as the bivalves have to deal with more stressors than before and cannot devote as much energy to reproduction. From there, the goal was to understand trends within each taxonomic Family of *Bivaliva*. Therefore, the second subquestion was does bivalve reproduction shift with the different families? This prediction was that different families have different frequencies and times to spawn, but they will generally all follow a trend within the family. Finally, the last subquestion was what is the model's accuracy score for predicting reproductive response based on environmental factors? Overall, with the presented data, the goal of the model was expected to classify the family and reproduction period at around 85% accuracy from a test-train split.

METHODS

Study site and organisms

With the help of Dr. Daniel Killam, we compiled the data from 435 different scientific articles with 542 entries. The bivalve data in this study contained organisms from all over the world including, Antarctica, the United Kingdom, India, and the United States of America. All of the organisms fell under the taxonomic class *Mollusca Bivalvia*, with each organism being bilaterally symmetric with flattened bodies and a hinge joint connecting the two shells at one point. The organism's data was only collected when an environmental variable (temperature, chlorophyll, salinity), a latitude and longitude location, study source, spawning period data, and at least a few months of data on reproduction. The majority of this data was collected in 2019.

Taxonomic data

We collected each species' data through the study. From there, we identified the taxonomic genus, family, and class and matched through multiple sources as there was no compiled list of species in the class *Bivaliva*. A total of 232 species were classified into 26 different families within

class *Bivaliva*. Considering the 324 families present, only 8 percent of the families were accounted for in this study as well as over ten thousand species being accounted for. The taxonomic genus was not considered in this study.

Data analysis tools

Data sets and processing code packages

We used basic Pandas and NumPy data frame processing methods to process the majority of the coding processes. We used Anaconda Navigator to launch a server for Jupyter Notebook to process the data in Python 3.12.3 using an IPYNB file to edit the Comma-Separated Values (CSV) files provided. We imported data frames using OS and glob directory imports to navigate MacOS systems. Importing of the CSV files was done through pandas on a local server. We used the math package to perform basic mathematical functions along with floor and ceiling functions to round the numbers. I used Regular Expressions (RegEx) to identify English words and patterns. For graph generation, the datasets were run through MatPlotLib's Pyplot and Patheffects to generate a base environment for the graphs to be built on. The plots are then created through Seaborn to produce the regression line with scatter, box and whisker, line, and bar plots. Seaborn is used to stylistically change the color, hues, and axes to make the data more easily digestible.

I created machine-learning algorithms and models through SciKit-Learn (Sklearn). Sklearn's sub-packages: Train Test Split, Neural Networks, Random Forest Classifier, Accuracy Score, Precision Matrix, One Hot Encoder, and Scientific Python (SciPy)'s Statsmodel were used. One Hot Encoder was used to convert categorical variables to number form for Random Forests and Neural Networks to be able to digest the information. Accuracy score and Precision Matrix were used to calculate the accuracy of the models through randomization. Using Train Test Split, I implemented a randomized group of data points using 80% of the data for a training set and 20% of the data for a testing set. SciPy was used to create Pearson Correlation Coefficients between two independent variables to show dependence with a number between 0 and 1. I also used SciPy's randint module to create a random distribution for randomizing the test set to reduce overfitting and prevent fitting the model incorrectly.

There were 27 tables for chlorophyll, 27 for salinity, and 91 for temperature. The data points on environmental data sum to 8860 values and the dates range from the mid-1950s to the late-2010s. The next dataset contains 233 lines with the taxonomic species name and family the species belongs. The final dataset contains the Site ID, Study Site, Location data, and Spawn data on the different species. The dataset contains 565 lines with 18 features.

Dataset combiner

I created a combiner file to compile the data into a large dataset. All the datasets mentioned above were pivoted and grouped to initially represent the data containing Site ID, location, time, environmental data value, type of data, spawn data, Species, Family, and Normalized Dates. This initial representation created more lines than necessary, so I restructured the data to contain all 3 environmental variables to decrease the number of rows along with decrease complexity without losing information. The combiner also served as a time normalizer. I converted all the different types of time variables to a year scale where the number one amounts to 1 year after the start of the study date.

Data engineering

(Figure 1) I started by using the Normalized Dates to create the number of days, months, and years the data was collected. The data did not contain the date, therefore, I used RegeX to identify the start date found in the Study feature. This allowed me to create the Accurate Normalized Date. The date was then converted into NumPy's datetime64 through RegEx formatting and NumPy permutations. These conversions allowed for the data to be logged in chronological order and ordered for graphing and analytic purposes.

- **SiteID**: ID number correlated to the study the information was extracted from.
- **Study**: The Author's last name and the year the study was done.
- **Species**: Contains the taxonomic *Genus* and *Species* name of the bivalve.
- **Locality**: Unoffical location the study has taken place.
- **LatDeg, LatMin**: The Lattitude Degree and Minute of the study's location in respective order
- **LongDeg, LongMin**: The Longitude Degree and Minute of the study's location in respective order
- **spawnstart, spawnstart2, spawnstart3**: The starting month of each spawn cycle (2/3 exist for species that spawn multiple times a year).
- **spawnend, spawnend2, spawnend3**: The ending month of each spawn cycle (2/3 exist for species that spawn multiple times a year).
- **peak1, peak2, peak3**: The peak month of the corresponding cycle (2/3 exist only for some species that spawn multiple times a year).
- **Normalized Date**: The date value where 1 is considered a year.
- **Chlorophyll**: The milligrams concentration of Chlorophyll in a Liter of water
- **Salinity**: The Practical Salinity Units (PSU) in parts per thousand (ppt).
- **Temperature**: The Degrees Celcius of the water.
- **Family**: The taxonomic family that each species belongs in.
- **Dropped columns (Notes, Unnamed: 18)**: Notes with small descriptions

Figure 1. **A description of all the provided features in the original big dataset.**

In terms of locational data, I created 4 extra features. Using RegEx again, I created categorical variables by dissecting the Locality feature of the dataset. Although there was a lot of variability and inconsistencies in the data, I manually corrected them in the CSV files. Through this, I then categorized each of the data points into countries creating a new feature Country. Using the latitude data, I created two new features. The Hemisphere is then created by identifying if the latitude value is positive or negative. Similarly, I categorized each species into a Latitude Zone. The 4 zones were Tropical, Subtropical, Temperate, and Boreal where the latitude restrictions were between each zone separated by 22.5-degree increments starting from the equator respectively.

For the spawn length data, I created a feature for the length of the spawn length by finding the distance between the start and end months. With a few extra simple adjustments, I created the features Number of Peaks, Average Time Step, etc. The spawn data was compressed to create a new data frame just to look at the species and spawn data along with environmental factors and Family.

Statistical analysis tools

Statistical comparisons

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

Figure 2. Pearson Correlation Coefficient (Turney, 2022).

I selected the Pearson Correlation Coefficient due to its simplicity and ability to find correlations with any 2 variables. The coefficient quantifies the strength of the linear relationship between 2 continuous variables. The coefficient's visualization is a scatter plot against each other to show a linear relationship. (Figure 2) The equation above can be summarized as the covariance of x and y over the standard deviation of the x and y variables. The correlation coefficient is represented by a value between -1 to 1. Correlation increases at the value strays away from 0 as a negative correlation suggests that -x correlates to y and vice versa. In our case, a coefficient above 0.2 shows a slight correlation where multiple slight variables would be good for a model while a score of more than 0.4 shows a significant enough correlation for modeling. The prior distributions are normal and independent to make them independent variables. Although the environmental data are affected by each other. For the simplicity and rudimentary comparisons for this study, they were considered independent. Moving forward, the variables are to be considered conditional and need to be properly distributed.

$$\text{Cramer's V} = \sqrt{\frac{\chi^2}{n \cdot \min(r-1, c-1)}}$$

Figure 3. **Chramer's V equation (Soula, 2024).**

For the categorical variable of Family, I used Cramer's V to calculate the Family and spawn length correlation. The Cramer's V equation is based on the Chi-Squared statistic utilizing the number of observations and minimum value between the number of rows and columns in the

contingency table. It creates an effect size for the chi-squared test allowing for larger and more general categorical variable comparisons (Soula, 2024). The Cramer's V value ranges from 0 to 1 representing complete independence and dependence respectively. A value above 0.3 is considered to have a significant enough correlation for modeling in this specific case.

Models

Preprocessing for machine learning

The models utilized the main data frame containing SiteID, Study, Species, LatDeg, LatMin, spawnend, peak1, spawnstart2, spawnend2, peak2, spawnstart3, spawnend3, peak3, Normalized Date, Chlorophyll, Salinity, Temperature, Family, Month, Day, Year, NumPeaks, Country, Hemisphere, Spawn Len 1, Spawn Len 2, Spawn Len 3, Lat Zones, Start Year, Accurate Noramalized Date, Salinity Fluctuation, Chlorophyll Fluctuation, and Temperature Fluctuation. I then grouped the data to have each species contain exactly one row of data. Each row of data then contained the first value it encountered when grouping. The only exceptions were Normalized Date, Chlorophyll, Salinity, Temperature, Month, Day, Year, Start Year, Accurate Noramalized Date, Salinity Fluctuation, Chlorophyll Fluctuation, and Temperature Fluctuation. These values took the mean of the data since it normalized the data. I converted the categorical data variables into numerical values that machine learning algorithms can interpret using the OneHotEncoder. One Hot Encoder automates the changing of each categorical value to a binary column. The data was then split into X variables containing all the features except the y variable. The first model used y as Family while the second model used Spawn Length 1. Then, the train-test-split was created on both X and y variables where 80% of the randomized data was placed in the training set and 20% was placed in the test set. This test set was set aside and the file remained locked to remove any source of bias or overfitting due to knowledge of the desired outcome.

Random forests

Simply put, Random Forests "utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees" (IBM). Utilizing the supervised learning model of decision

tree classifiers, Random Forests mitigates the overfitting and bias through bootstrapping in aggregation. With aggregate bootstrapping, multiple versions of a dataset are created with subsets of validation and training. This reduces overfitting by introducing a randomizer. By using multiple iterations of the bagged decision trees, the model can procure a supervised machine learning model. By introducing multiple features, the Random Forest classifier produces a guess of the test results in which an accuracy score is calculated.

Neural networks

"A neural network is a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain." (AWS). In layman's terms, the network contains a set number of nodes that run linear regression models using specific data, weight, bias, and eventually output. Using initial values, weights are assigned over multiple iterations of nodes to optimize each variable. This output variable is then run through an 'activation function' which creates a threshold and only allows effective variables through. This constitutes one layer of the network. (IBM) Multiple layers of this are run until a solution is reached the data is not able to send through enough nodes through a layer. Manually setting the layers and node size is crucial to the network, and in our case, not a lot of data was able to pass through the neural network's Relu activation function. Creating a Mean Squared Error for the data would have been the most effective way to fix this problem, but other issues arose. Therefore, I created a cost function and reduced it to the best of my ability with the available data to optimize the neural network. Much like the Random Forest, a test dataset was produced, and an accuracy score was calculated.

RESULTS

Temperature's effect on spawning

Looking at a specific species, *Laternula Ellitica*, the spawning period lined up with the warm season created an initial indication that temperature directly impacts the spawning period (Figure 5). As *Ellitica* is a species residing in Antarctica, the temperature fluctuations are minor not going above 2°C and below -2°C with temperatures peaking in the winter and fall. I realized

that species in different hemispheres will behave based on their seasonal cycle (Figure 4). Therefore, I decided to separate the two groups in future graphs. Focusing on a singular feature shows that the spawning data is extremely varied throughout species.

Figure 4. (Left: Spawning start) & (Right: Spawn end) months data for each species colored by Family.

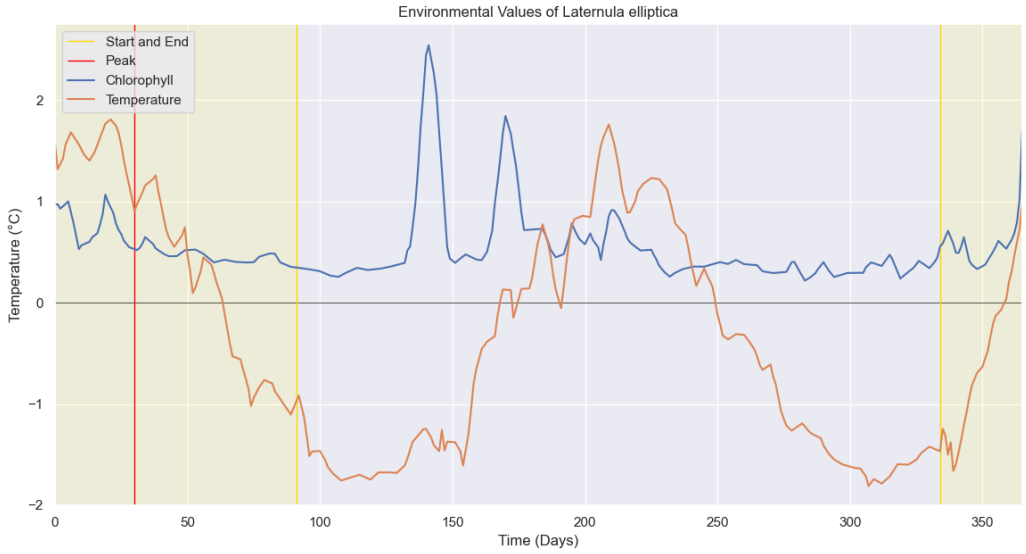

Figure 5. A closer look at Laternula Elliptica's env. data plotted over time with the spawning period shown shaded in yellow.

To investigate trends, I decided to look into a specific family containing enough time data for the species to complete at least half a year cycle. I chose *Mytilidae* due to it having 8 species containing plenty of time and temperature data to create rudimentary comparisons. (Figure 5) The

temperature data of the species divided up into their respective hemispheres. The data varies drastically with some species seeming to ignore their seasonal cycles.

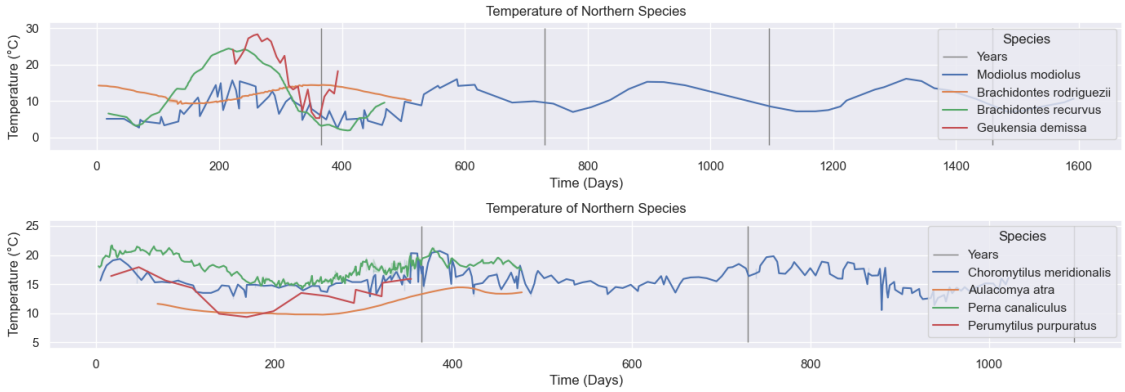

Figure 6. Temperature data for the Family *Mytillidae* divided up into their respective hemispheres (North top, South bottom).

The large contrast in temperatures required a simpler feature. Therefore, correlating the length of spawn with temperature fluctuations would create two independent variables that can be checked to see for correlation. (Figure 7) To exemplify some correlation that can be seen between the hemispheres, the graphs' spawn period was created, but no significant correlation was noticed.

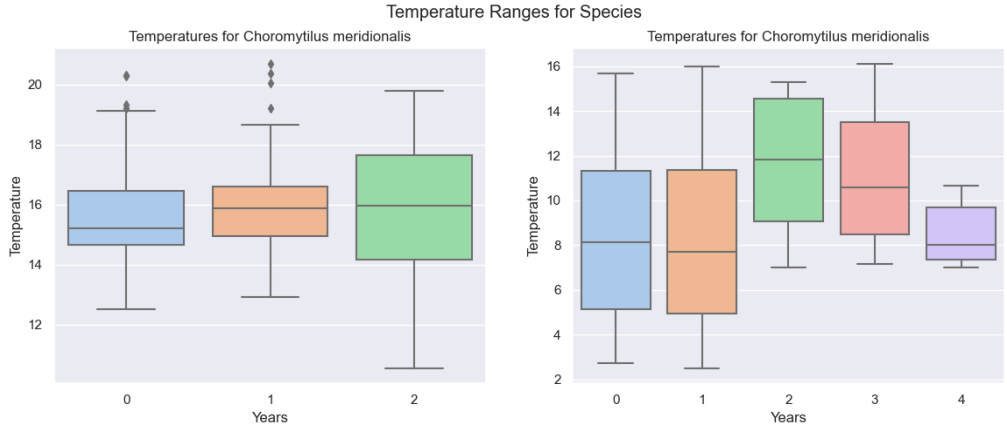

Figure 7. Temperature ranges of Mytillidae species split up by Southern (left) and Northern (right) hemispheres. Each Box and Whisker plot represents a separate species. The box missing in the Southern column did not contain enough data for a plot.

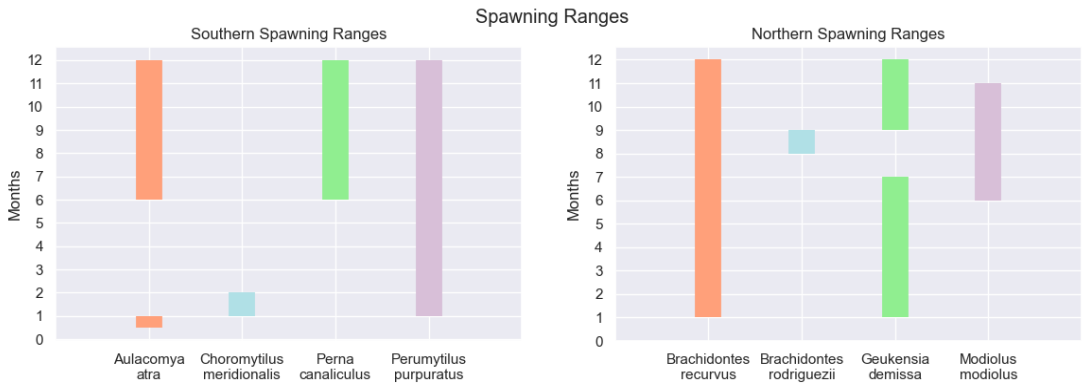

Figure 8. Spawning months of Mytillidae split up by Southern and Northern hemispheres. Species are highlighted by hue.

I decided that looking at the spawn start and end would create bias and not encapsulate the majority of the data. The peak would have too much variation and not always be in the middle of the spawning period. Therefore, I utilized the spawn length of the data for each species' first cycle. I used the first spawn length as there were only 4 species with 2 spawn periods while one of those species had a third spawn. The Pearson's Correlation Coefficient for Temperature and Spawning Length was -0.21437.

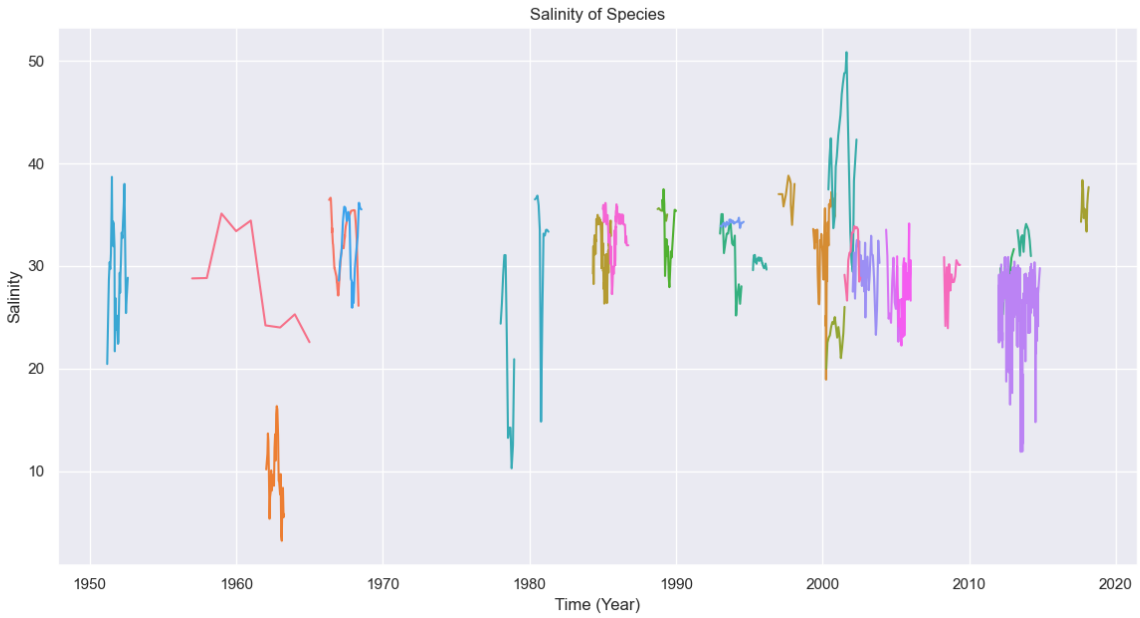

Figure 9. A look at all the differently colored species and their salinity readings over time.

Using a very similar principle, I first graphed the data for salinity to look at the variability in the data over time. I wanted to look at the variation in salinity hoping it would increase over time. Although there is a little bit of an increase, there is not enough visual change to dig deeper into the hypothesis. I then created a new feature by finding the lowest and highest salinity levels and subtracting them. This created the salinity range being a continuous variable. The spawn range and salinity range's continuous nature alone allowed for Pearson's Correlation Coefficient to be effective. Using SciPy's Pearson's Correlation Coefficient, I got -0.061159. Due to the low value and correlation, I went back to try a correlation coefficient on the salinity values with the spawn ranges. This produced a much more desirable coefficient of 0.251682.

(Figure 10) Finally, I graphed chlorophyll against time. The peak chlorophyll value was 72.2 mg/L while the average was around 10.2 mg/L. Each species experienced at least one spike and had high variance scores throughout their study period. However, using the knowledge from salinity, I decided to continue with the data points. The Pearson's Correlation Coefficient was around 0.352838.

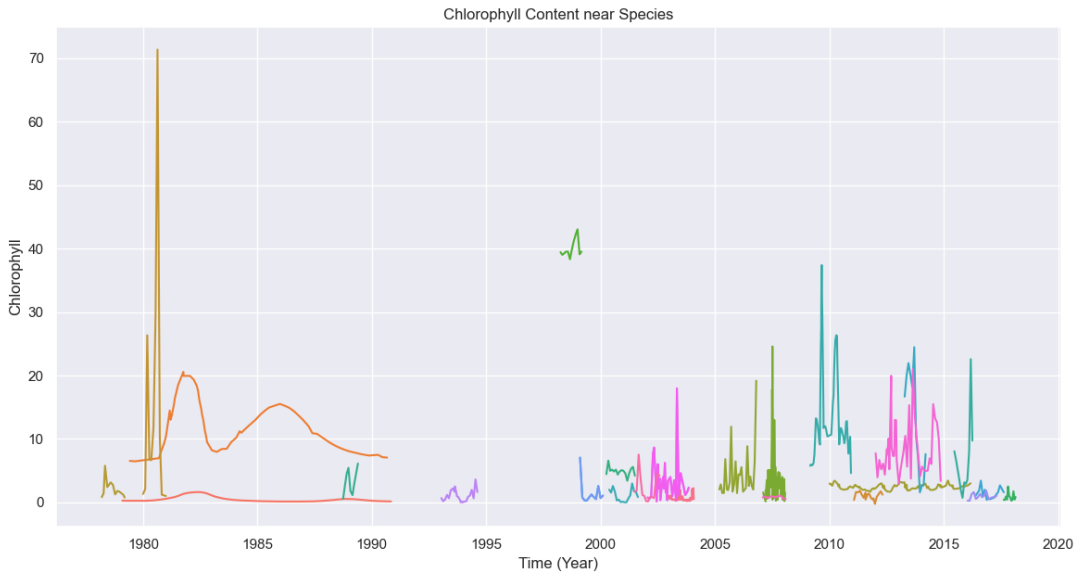

Figure 10. The chlorophyll values shaded by species graphed over time in years. Each color represents a species.

Effect of family on variables

Using the classification of taxonomic family as a basis, I hypothesized that species categorized into these families will have similar features and traits which ultimately should affect how they are affected by their environment. However, without a direct causal relationship studied, finding out trends was the biggest probability of finding a correlation. Using these categorical variables, I decided to use scatter plots to achieve this.

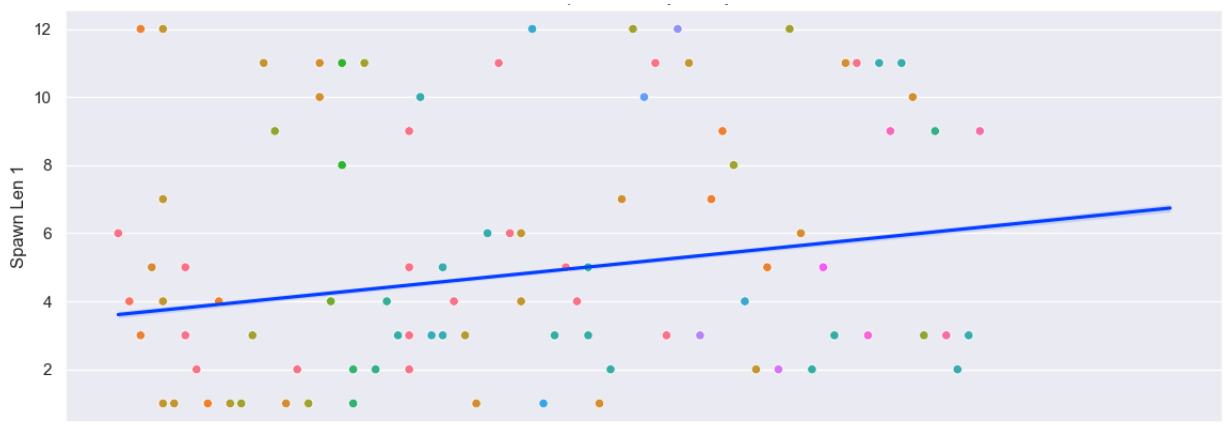

Figure 11. Each Species's spawn length is colored by Family. The blue regression line shows the relationship between the Spawn Length and species.

Looking at relationships between Families, I correlated the species against the spawn length. (Figure 10) This showed the linear regression plotted against the scatter plot with a low correlation between the two. Therefore, zooming out to a broader classifier, I correlated the categorical family feature with a continuous variable of Spawn Length. Using the Cramer's V Coefficient, I got an average of 0.06116 for all the families. Although outliers existed, the average value accurately encapsulates most of the spawn lengths.

Machine learning

Each model utilized the preprocessed data frame to train-test-split and run through the model. The neural network was run with y being the Family feature. Initially, the network went through a sigmoid function, but the ReLU activation function was the only one that was able to

successfully output an answer. With 50 hidden layers and 32 nodes, the model was able to provide values. However, when run against the test set's accuracy score, there was a 6.7130% accuracy. This means that out of the 26 different family options, a random change would amount to 3.8462%. This means the accuracy score is about 2 times a random guess. This model maxed out at 6.7% as slightly manipulating and tweaking the data did not help increase the percentage.

When it came to the Random Forests Model, the data utilized the preprocessed data frame to train-test-split. The data was then run through Sklearn's Random Forest Classifier with 100 trees per forest and a Gini Index to filter out the incorrect predictions. Initially, the Log Loss Function was utilized, but the data made more sense to utilize a Gini Index due to the many of the categorical variables. When it came to setting the y value as Family, an accuracy score was calculated against the test set to produce a score of 31.579%. When it came to setting the y value as Spawn Length, an accuracy score was calculated against the test set to produce a score of 47.294%. With only 12 possible selections, randomly guessing would procure a score of 8.333%. This means that the score for Spawn Length was 5 times better than guessing and, the score for Family was about 5 times more effective with a random guess rate of 3.8462% as mentioned above.

DISCUSSION

The study aimed to create comparisons and correlations between *Bivalivia* species with each other. The first question was whether temperature and other variables were able to impact the spawning of bivalves. Finding direct correlations between the Salinity, Temperature, and Chlorophyll, I was able to find directly correlative evidence between different spawn data to explain the changes happening. The second question honed in on a categorical variable of Family to see relationships within a bivalve's family. The main experiment was to look at how spawning data was affected by family and if there was any noticeable direct correlation. On top of this, I aimed to find other correlations with Family. The third question was to understand if the data presented could create a model that is adequate to predict the Family and Spawn Length of the bivalve presented with limited data. The results of the first experiment were well within my hypothesis on temperature and chlorophyll. However, when it came to the Salinity, there was barely any correlation to how the bivalve's reproduction changed. The second question was answered not according to plan. It did not correlate how Spawn Length was while the only other

data point that had any effect was temperature. This data correlation showed there is not much value to having salinity in this dataset. However, using all this information, all the data was utilized to get the best model through Random Forests.

Analysis of environmental correlation data

The correlation of the three features suggests that spawning is affected by changes in the environment. Mainly, the Chlorophyll amount shows the most promising results with a Pearson score of 0.35284. To better understand this scale, Rovetta's paper from the National Library of Medicine, if your score between 0.5 and 0.25 is a fair amount of correlation to make it considerable. For the sake of our study, we will focus on taking the absolute value of the Pearson score, as the negative correlation still shows that there is change, and all we are looking for is a causal relationship. This allows for the temperature to have a poor yet still significant enough correlation to consider for the model. When it came to looking in depth at the temperature data, the Family Mytilldae showed no promising correlation for the temperature data when put against time, spawn length, spawn start/end, and spawn peak. Visually they made me realize that the correlation was barely present. Due to this, the coefficient achieved made lots of sense. Similarly, the rest of the Salinity showed absolutely no correlation. However, Chlorophyll was a different story. Chlorophyll's correlation when it came to being compared against time showed a very distinct pattern of data. The large spikes and sensitivity created extreme standard deviations from the mean. However, the data seemed to properly along with spawn length. This hypothesis stemmed from the idea that bivalves thrive due to the increase in nutrients presented. This gives them the comfort and ability to release more gametes into the ocean for a longer time which explains the positive correlation score. On top of that, the chlorophyll's relation to water temperature suggests that the temperature and chlorophyll have higher Pearson scores for a reason.

Effects of the family

Looking into the definition of a taxonomic family, there is no real denomination on what constitutes a family. However, within the class *Bivalvia*, family denominations rely heavily on physical features (Allmon, 2020). Therefore, these classifications oftentimes do not consider how

different they can be reproductively and internally. For this reason, I believe that the Family class deviated so much from the hypothesis of Question 2. The hypothesis was based on the idea of how plant-like organisms behave. Many of the plants come from similar regions and have characteristic features, but almost all of them have reproductive similarities. Although bivalves do follow this trend, their spawn lengths and start/end period seemed to have more variation and deviance from the plant norm. Much like this information, Cramer's V Correlation suggested that there was no correlation of 0.06116 showing that randomly plotting a point on the graph is almost the same as this distribution.

To mitigate this effect, I looked into the correlation of the environmental variables. All three variables showed no correlation due to the locality of their data. It showed how each family can be local, but there are different climates and areas where these bivalves live within a Family. When it came to each Family's spawn data, the correlation for spawn start and end was essentially non-existent and the same. Even with the created features in spawn, the fact that there weren't a lot of significant correlations brought about the validity of the Family classifications. Classification systems seemed to focus too much on appearance and not as much on the functionality of the ecosystem along with reproduction and feeding. However, creating more quantifiable data points based on the qualities each family possesses is essential to creating a better family classification system.

Family and spawn length modeling

To answer the question regarding the validity of my model, I set a goal to reach an 85% accuracy rating from the test data set. However, I was not able to create a model as effectively as I wanted. Although I do believe that my model for Family performed way better than guessing. It was about 9 times better. However, I do believe that with even 1 less feature available, that would be considerably worse making the model invalid. For that reason, I wanted to create a more simplified version of the model using Spawn Length. The reduction of dimension to a numerical continuous variable allowed for the model to have a lot more regularized options. I believe that the decision trees were able to make a less extreme decision per split. This allowed it to perform to the point where it was getting the correct answer about half the time. This suggests that the combination of the slight correlation was additive and significant.

Limitations of the current models and correlations

With the dataset provided, I believe there were limitations, but I want to focus on the models. For the Neural Networks model, I believe that the use of ReLU as an activation function was apt for some features in the dataset, but features like Country and Hemisphere were limited to being reduced down to a value. Either creating better classification systems or being able to cluster the data properly would have been in terms of accuracy. The model also had a lot of trouble when it came to creating large enough distinctions for the activation function to let correct entries through. Not knowing the sensitivity of the model due to the hidden layers was a major limitation of this network. Moving on to the Random Forest. The random forest seemed to struggle with the amount of data being processed. Utilizing multiple categorical variables with lots of categories to One Hot Encode the data created excessive amounts of data which limited the computing speed and power of the model. Moreover, the overfitting in this model did seem to be a little higher when run multiple times with different train-test splits. Mitigating the overfit of this model seemed to be the biggest challenge as it trained too much on a smaller sample size than wanted.

Diving deeper into the data, there were multiple limitations to the dataset. First and foremost, the data was limited by time and the time range. Since the data ranges from the 1950s to the 2010s, there is a sparse amount of data over a large period. Looking past the barriers of the technology evolving for these measuring tools, not enough data for such a large time period became the biggest issue. The data within periods weren't enough to train the model for consistent results. Much like not having enough data, the features or columns of this dataset were very limited to location and spawn. Spawning didn't have a precise measurement component making it difficult to relate the data with other features. The inclusion of spawn rates would have given a few new dimensions to improve accuracy in modeling. For the location, having a few bivalves per certain localities forced the data to depend solely on these few species. The more diversity included in the data would have created a stronger model.

Future directions in classification

Despite all these limitations, this model is a rudimentary proof of concept for implementing machine learning concepts into research and natural sciences. Utilizing the current dataset, I believe that there are a multitude of papers to collect more data on bivalves and their reproduction. In fact, I want to focus on adapting this model to create correlations between family and classification systems. Creating classifications on the function and form of the bivalve is something that I believe needs to be focused on and defined more clearly. However, this is not limited to the rows of data. Increasing the type of information that is extracted from the studies is even more essential. Collecting data on spawning rates and how many gametes are released could drastically influence the validity of the model. Although it is hard to collect data on the gametes being released, such data will add higher dimensionality to the data while paving new paths for new research. Much like this, finding more environmental data would help create more distinct classifications. I want to utilize different models that don't have the numeric limitations of Neural Networks and Decision Trees. Looking into natural language processors and clustering algorithms is essential for being able to categorize the locality and time.

From here, the final form of this model is to employ the model's classification system to have a limited amount of data input. This input then runs through the model to give a predictive percentage on what type of bivalve it is. For instance, only having the spawn rate, spawn period, locality, and local temperature, a few selections of species and families should show. Each family will contain a percentage likelihood while the species within each family will be presented. Through continued usage and time, the model collects data with these inputs to refine its results.

The furthest yet most promising future for this model is based on utilizing autoregressive integrated moving averages (ARIMA) to look at the chronological trend of the environmental factors. ARIMA has specialized in new correlations for a time due to its lagged usage of averages. The model could then be adapted to procure data on future trends of a specific bivalve's survivability with predictions on environmental factor trends. Such a model could largely benefit fishermen and biologists by simulating extreme conditions with a quickly changing climate. For once, many indicator bivalves can then be protected without having to go through these drastic conditions.

In our study, we found that environmental factors like temperature, salinity, and chlorophyll levels exerted varying influences on the reproductive patterns of the *Bivalvia* family. While temperature and chlorophyll demonstrated significant statistical differences in each

bivalve's reproductive behavior, salinity did not exhibit the same level of significance. Additionally, bivalves within the same taxonomic family displayed divergent reproductive lengths and cycles, indicating minimal correlation. Notably, variations in temperature and chlorophyll content, particularly in cooler climates, significantly impacted bivalve reproductive patterns, leading to fluctuations in spawn lengths and peaks over time. Utilizing this data, we developed a predictive model that integrated environmental variables and taxonomic classifications, achieving approximately 50% accuracy in forecasting spawn lengths. These findings underscore the need for a critical reassessment of taxonomic criteria and the substantial variability observed among closely related species, particularly concerning fundamental traits such as spawning behavior. Despite these challenges, our results offer promise for the development of more refined models capable of classifying species with limited data, thereby assisting scientists in predicting species survivability and reproductive success more effectively.

ACKNOWLEDGEMENTS

The largest thanks to Dr. Daniel Killam for mentoring me through this whole process by consistently proposing new avenues to explore while providing consistent and constructive support. I was so thankful to be able to connect with Dr. Killam and carry out the modeling project through his dataset procured throughout the majority of 2019 to 2020. Without his help, I would not have been able to find a more relevant and impactful project. Not only that, I carry the utmost respect and thanks to Dr. Patina Mendez and PhD candidate Melissa von Mayrhauser for supporting me through the thesis process journey. Patina was always patient with me and pushed me to find and connect with people outside of UC Berkeley to procure a thesis topic in my respective field. Thanks to her guidance from topic formation to thesis completion, I was able to follow the necessary steps to finish this thesis. Last but certainly not least, Melissa von Maryhauser gave excellent suggestions and revisions when it was most needed. She gave me to constructive yet encouraging support I needed in revising my mistakes to enhance my paper.

BIBLIOGRAPHY

Allmon, W. D. 2020. Class Bivalvia - Digital Atlas of Ancient Life. https://www.digitalatlasofancientlife.org/learn/mollusca/bivalvia/.

Cárdenas, E. B., and D. A. Aranda. 2000, January 1. A review of reproductive patterns of bivalve mollusks from Mexico: Ingenta Connect.

Fisheries, N. 2021, April 5. How Much Is A Clam Worth To A Coastal Community? | NOAA Fisheries. https://www.fisheries.noaa.gov/feature-story/how-much-clam-worth-coastal-community

Garcia-Soto, C., L. Cheng, L. Caesar, S. Schmidtko, E. B. Jewett, A. Cheripka, I. Rigor, A. Caballero, S. Chiba, J. C. Báez, T. Zielinski, and J. P. Abraham. 2021. An Overview of Ocean Climate Change Indicators: Sea Surface Temperature, Ocean Heat Content, Ocean pH, Dissolved Oxygen Concentration, Arctic Sea Ice Extent, Thickness and Volume, Sea Level and Strength of the AMOC (Atlantic Meridional Overturning Circulation). Frontiers in Marine Science 8.

Guinotte, J. M., and V. J. Fabry. 2008. Ocean Acidification and Its Potential Effects on Marine Ecosystems. Annals of the New York Academy of Sciences 1134:320–342.

IBM. (n.d.). What Is Random Forest? | IBM. https://www.ibm.com/topics/random-forest.

Lewandowska, A. M., D. G. Boyce, M. Hofmann, B. Matthiessen, U. Sommer, and B. Worm. 2014. Effects of sea surface warming on marine plankton. Ecology Letters 17:614–623.

Li, Q., S. Sun, F. Zhang, M. Wang, and M. Li. 2019. Effects of hypoxia on survival, behavior, metabolism and cellular damage of Manila clam (Ruditapes philippinarum). PLoS ONE 14:e0215158.

Rovetta, A. (n.d.). Raiders of the Lost Correlation: A Guide on Using Pearson and Spearman Coefficients to Detect Hidden Correlations in Medical Sciences. Cureus 12:e11794.

Soula, D. W. 2024, January 6. Cramer's V.

Turney, S. 2022, May 13. Pearson Correlation Coefficient (r) | Guide & Examples. https://www.scribbr.com/statistics/pearson-correlation-coefficient/.

UC Berkeley. 2001. The Bivalvia. https://ucmp.berkeley.edu/taxa/inverts/mollusca/bivalvia.php.

US DoC, N. O. and A. A. 2023, January 20. What is a bivalve mollusk? https://oceanservice.noaa.gov/facts/bivalve.html.

US EPA, O. 2017, February 8. Inventory of U.S. Greenhouse Gas Emissions and Sinks. Reports and Assessments. https://www.epa.gov/ghgemissions/inventory-us-greenhouse-gas-emissions-and-sinks.

USNPS. (n.d.). Monitoring Razor Clams as an Indicator of Nearshore Ecosystem Health (U.S. National Park Service). https://www.nps.gov/articles/razorclams.htm.

Vaughn, C. C., and T. J. Hoellein. 2018. Bivalve Impacts in Freshwater and Marine Ecosystems. Annual Review of Ecology, Evolution, and Systematics 49:183–208.

What is a Neural Network? - Artificial Neural Network Explained - AWS. https://aws.amazon.com/what-is/neural-network/.

Zhao, X., Y. Han, B. Chen, B. Xia, K. Qu, and G. Liu. 2020. CO2-driven ocean acidification weakens mussel shell defense capacity and induces global molecular compensatory responses. Chemosphere 243:125415.

Zhong, W., and J. D. Haigh. 2013. The greenhouse effect and carbon dioxide. Weather 68:100–105.