Effect of Environmental Factors on Benthic Invertebrate Populations in the South Bay area

Quincy Kapsner
San Diego State University
Chula Vista, USA
qkapsner8362@sdsu.edu

***Abstract* — This project aims to analyze the effects of environmental factors on benthic invertebrate populations in the South Bay area off the Southern Californian coast based on data collected by the Ocean Monitoring Program (OMP).**

Introduction

Understanding the relationships between water conditions and benthic invertebrate populations is crucial for assessing and managing aquatic ecosystems. This project aims to predict the abundance of benthic invertebrates based on water quality parameters by using machine learning techniques.

Benthic invertebrates are small marine animals that live on or in the sediment at the bottoms of oceans, lakes, rivers, and other aquatic environments. Examples include worms, snails, clams, and small crustaceans like shrimp and crabs. They serve important roles in their ecosystems serving as food and helping to recycle nutrients. For scientists, they are vital for monitoring anthropogenic (human) impacts [1].

The data is provided by the City of San Diego's Ocean Monitoring Program. "The OMP monitors the effects of treated wastewater discharged on the local marine environment. It extends 340 square miles total from northern San Diego to northern Baja California, Mexico, and encompasses the shoreline seaward approximately 10 miles, to depths of over 500 meters" [2].

Benthic invertebrate population data was gathered by sediment samples. Water condition data was collected by a Real-time Oceanographic Mooring System (RTOMS).

RTOMS are anchored buoys that have several instruments at multiple depths, collecting data on various water qualities intended to aid in assessing environmental conditions and impacts of oceanographic and anthropogenic events on the local water. All RTOMS data has been reviewed for quality control [3]. This project uses data from the South Bay (SBOO) RTOMS.

This project employs machine learning algorithms to develop predictive models for understanding future trends in benthic invertebrate populations.

Approach

A. Data Collection

This project used data collected by the OMP. The datasets used were:

- Benthic Invertebrates - Abundance of invertebrates (approximately < 1 cm in size) found in sediment samples collected throughout the region [3].
- RTOMS Ocean Chemistry - Ocean chemistry measurements collected by Real-time Oceanographic Mooring System (RTOMS). Ocean chemistry parameters include dissolved oxygen, fraction dissolved carbon dioxide (xCO2), nitrate + nitrite, and total pH [4].
- RTOMS Salinity - Salinity measurements collected by Real-time Oceanographic Mooring System (RTOMS) [5].

- RTOMS Water Quality - Water quality measurements collected by Real-time Oceanographic Mooring System (RTOMS). Quality parameters include Biological oxygen demand (BOD) equivalent, Chlorophyll fluorescence, Colored dissolved organic matter (CDOM) fluorescence equivalent, and Turbidity [6].
- RTOMS Water Temperature - Water temperature measurements collected by Real-time Oceanographic Mooring System (RTOMS) [7].

B. Data Exploration, Cleaning, and Feature Engineering

i. Benthic Invertebrates

The Benthic Invertebrates' original format had the columns "sample", "station", "date_sample", "taxa_name", "abundance", "project", and "depth_m". It recorded where the sample was taken, what species were in the sample, and the abundance of each species. The data was gathered 2 - 4 times a year since 1991. It had data from both the Point Loma location and the South Bay location.

The first step of cleaning this dataset was to remove all data from the Point Loma location. This included making sure all the entries had the same format (as some locations were formatted “PL” and others “PLOO”) and dropping any rows with the standardized format for Point Loma. Then, I dropped the columns “sample”, “station”, and “project”. “Sample” was a unique sample ID for each sediment sample gathered and was dropped because this project would not be incorporating individual sample specifications. “Station” was a code for the location where each sediment sample was gathered and was dropped for the same reason. “Project” was the identifier of which project the sample was a part of and was dropped as it was now unnecessary.

There were no null or default values to account for.

Samples were gathered 2 - 4 times a year but were gathered a few days apart. Meaning that there were small groups of samples a few days apart, and then the next group would be several months later. I tidied up these groups by making all dates from the same month the first of that month.

While this form of the data was tidy and clean, it was not a good form for modeling because it was semi-3-dimensional since it listed the abundance of every individual species. So, I made a second data frame that was better suited for modeling. The columns of this data frame were “date”, “depth_m”, and “abundance”. The depths were cleaned by rounding them to the nearest 10m. The abundance was the collective total of all samples at that depth.

I then made a secondary version of this data frame that was rotated so that each row had the date and the abundance for each depth.

ii. All Water Data

The data frames with water condition data all had the same format: “project”, “Deployment#”, “unixtime_1000_pst”, “datetime_pst”, “depth_m”, “parameter”, “units”, “values”, and “qualifier_flag”. Due to this, I started by dropping common unnecessary columns from all of them. The columns dropped were “project”, as the water condition data is already split by project so that the entire file is the same project, “Deployment#”, as it is just a label for the RTOMS used, and “unixtime_1000_pst”, as its function was covered by “datetime_pst”.

From the salinity and water temperature data frames, I also dropped the “parameter” and “units” columns as both of those frames only had one type of data.

iii. Ocean Chemistry and Water Quality

Upon first exploring the ocean chemistry dataset, I realized that the depth values were floating point numbers. I changed them to integers because they are integers for every other dataset.

The water condition data is collected every 10 minutes. For this project, I did not need it to be this in-depth, so I decided to cut the data down to once a day. I did this by grouping by date, depth_m, parameter, and the unit, then getting the average of the “value” column and mode of “qualifier_flag”. “Qualifier_flag” has integers that represent if the data is reviewed and good, suspicious, missing, etc. Since the majority of the data for that day was flagged with that qualifier, that flag can be reasonably applied to the averaged data.

Then, I split the ocean chemistry data into separate data frames based on the parameter because each parameter had different nuances of the data, so I felt it would be easier to deal with them separately and then remerge them later. The parameters were 'Dissolved oxygen' (unit: mg/L), 'Fraction dissolved carbon dioxide (xCO2)' (unit: ppm), 'Nitrate + nitrite' (unit: µM), and 'Total pH'.

Next, I utilized the qualifier flag system to clean the data. The flags had the following meanings:

- 1 - pass/good - For data reviewed both automatically and manually
- 2 - provisional/unreviewed - For data that is not reviewed; or data received review but quality could not be determined
- 3 - suspect/questionable - Flagged due to suspicious values for site/season; or as possible instrument drift (such as due to biofouling)
- 4 - bad - Flagged due to out of range for instrument; or manually flagged as clearly bad (such as due to instrument malfunction)
- 5 - value changed/drift-corrected - Used only in post-processing; values have been corrected based on new information, such as water sample results to correct for drift or new calibration factors. For data use purposes, this flag can be treated as a “pass”
- 9 - missing - Placeholder to show missing data; some gaps might be able to be filled in later by downloaded data [4].

Since I am not trained to qualify the data, I determined that data with flags 2, 3, and 5 could be considered as flag 1 for this correlation analysis. Flags 4 and 9 would be treated as missing data. I implemented this by changing any flags 2, 3, and 5 to flag 1 and any flag 4 to be flag 9.

To address missing data, I decided to use time-based interpolation to fill small gaps in the data and to drop any gaps larger than 70 days. I made a function to print the depth, start date, end date, and length of consecutive rows with flag 9. Based on the output of this function, I dropped any large groups and interpolated the rest. I would then drop the “qualifier_flag” column as it was no longer needed. I did this for all four parameter types.

Most of the large gaps were due to a large pause in the year 2021. I do not know the exact reason for the missing 2021 data, but I speculate it is related to the COVID-19 pandemic.

After the data was cleaned, I remerged the data frames, now with each parameter being a different column.

The water quality dataset was formatted similarly: with four parameters of measurement. The parameters were 'Chlorophyll fluorescence' (unit: µg/L), 'Colored dissolved organic matter (CDOM) fluorescence equivalent' (unit: ppb), 'Turbidity' (unit: ntu), and 'Biological oxygen demand (BOD) equivalent' (unit: mg/L). Due to

the similarity, I followed the same procedure: averaging the daily data, splitting it into different data frames based on parameter, keeping data with flags 1, 2, 3, and 5, removing or interpolating data with flags 4 and 9, then finally remerging with each parameter a different column.

vi. Salinity and Water Temperature

The salinity and water temperature datasets had the same format as ocean chemistry and water quality, except they only had one value in the "parameter" and "units" columns. So I again followed the same procedure, excluding splitting the data frames into multiple steps since they both have one parameter.

v. Merge Water Data

For each data frame, the data column names followed the same format: "value_[unit]". Some parameters had the same unit, so before remerging the water data back into a single data frame the columns needed to be renamed. I also decided to rename the columns to preserve the parameter information. I renamed the columns to have the format: "[parameter]_[unit]".

After renaming the columns, I merged the data frames into one based on matching date and depth.

There was a lot of missing data in the total water data frame because not all parameters were measured at all depths. The depths data was measured at were 1m, 10m, 18m, and 26m.

Only salinity and temperature were recorded at 10m, so I decided to drop the data at that depth because there would only be two features for that model.

I then split the data into separate data frames based on depth and dropped the now unnecessary date column from each.

Nitrate, fraction dissolved carbon dioxide (xCO2), biological oxygen demand (BOD) equivalent, and pH were not measured at 18m, so I dropped those columns from that data frame. Similarly, I dropped the xCO2 column from the depth 26m data frame and the BOD column from the depth 1m data frame.

I then had to address the remnant 2021 gaps in the data. Though the gaps were previously dropped or interpolated while the data were in separate frames, when they were put together, the gaps reappeared as the exact dates for the start and ends of the large gaps did not exactly line up. I suspect this is because the OMP was not paused for the pandemic all at once. Rather, I believe that instruments were kept on for as long as possible before being forced to shut off.

Regardless of the reasoning, there was once again a large gap in each of the water condition data frames. The gaps span several months, so it is not reasonable to interpolate or fill them with the mean or other values. The only option was to drop the entire gap. I did this by printing the null values of each data frame. This gave me the first and last dates that had null values, effectively providing the span of the 2021 gap. I would then drop all rows within those two dates.

With that, the water conditions data frames were finished.

Water data at depth 1m had the columns: "date", "o2_mg/L", "nn_uM", "xco2_ppm", "ph", "salinity_psu", "turb_ntu", "cdom_ppb", "chlf_ug/L", and "temp_C".

Water data at depth 18m had the columns: "date", "o2_mg/L", "salinity_psu", "turb_ntu", "cdom_ppb", "chlf_ug/L", and "temp_C".

Water data at depth 26m had the columns: "date", "o2_mg/L", "nn_uM", "ph", "salinity_psu", "turb_ntu", "cdom_ppb", "bod_mg/L, "chlf_ug/L", and "temp_C".

C. Analysis

i. Visual

To visually analyze the data for any trends, I plotted the data frames on line graphs.

I noticed that benthic invertebrate populations have largely been consistent except for a large spike in 07-2022 at 20m and an upward trend for 60m and 40m. Invertebrates at a depth of 20m have regularly fluctuating populations based on the time of year. Finally, I saw that populations are lowest at 50m.

It is also clear that several measurements at a depth of 1m spiked in 05-2020. Additionally, at a depth of 18m, several measurements started fluctuating around 03-2022. The same 05-2022 spike can be seen in nitrate levels at 26m, but the rest of the measurements are consistent.

ii. Correlation Matrices

Next, I plotted heatmaps of the correlation matrices to see how each measurement correlated, and therefore may affect, the abundance of benthic invertebrates at each depth.

To do this, I first had to merge the population data with the water condition data, keeping the water depths as separate data frames and population depths as different columns within each frame.

For the water data at depth 1m, I found these notable correlations:

Exact correlations:

- Abundance at 20m with temperature, CDOM, XCO2, and dissolved oxygen levels
- Abundance at 30m with temperature and dissolved oxygen levels
- Abundance at 40m with abundance at 50m
- Abundance at 60m with turbidity

Biggest correlations (over +-0.75):

- Abundance at 20m with abundance at 30m, 40m, 50m, chlorophyll fluorescence, and salinity
- Abundance at 30m with abundance at 20m, 40m, 50m, chlorophyll fluorescence, CDOM, XCO2, and salinity
- Abundance at 40m and 50m with abundance at 30m, 20m, temperature, CDOM, pH, dissolved oxygen levels, and the date
- Abundance at 60m with chlorophyll fluorescence, salinity, and nitrate

For the water data at depth 18m, I found these notable correlations:

Exact correlations:

- Abundance at 20m with temperature, and chlorophyll fluorescence
- Abundance at 40m with abundance at 50m

Biggest correlations (over +-0.75):

- Abundance at 20m with abundance at 30m, 40m, 50m, turbidity, salinity, and dissolved oxygen levels
- Abundance at 30m with abundance at 20m, 40m, 50m, temperature, chlorophyll fluorescence, turbidity, salinity, and dissolved oxygen levels
- Abundance at 40m and 50m with abundance at 30m, 20m, temperature, chlorophyll fluorescence, and the date
- Abundance at 60m with CDOM, salinity, and dissolved oxygen levels

For the water data at depth 26m, I found these notable correlations:

Biggest correlations (over +-0.75):

- Abundance at 20m with abundance at 30m, temperature, chlorophyll fluorescence, BOD, nitrate, and dissolved oxygen levels
- Abundance at 30m with abundance at 20m, 40m, temperature, chlorophyll fluorescence, BOD, and turbidity

- Abundance at 40m with abundance at 30m, 50m, BOD, turbidity, and the date
- Abundance at 50m with abundance at 40m, turbidity, and the date

D. Modeling

The final phase of this project was creating the models. There were many null values in the combined data frames because the population data was gathered only a few times a year, so I forward-filled them. I also made a "unixtime" column to have a numeric value for the dates that can be modeled.

Due to the nature of my data, I needed a model that could have multiple input and output variables. I explored a few different types of models that could do this, but finally opted for Gradient Boosting Regression considering the complexity that so many features would bring to the relationship. I used an 80/20 split and trained a model for each population depth.

Evaluation

I used Root Mean Squared Error (RMSE) and R-squared values to assess the model's accuracy and goodness of fit. The RMSE scores were all very close to 0 and the R-squared scores were all very close to 1, which implied overfitting.

To confirm this, I cross-validated my models. This resulted in very high RMSE values (except for the models at a depth of 50m, which all actually performed well). This meant my models were not good at predicting unseen data.

I tried other kinds of models and hyperparameter tuning, but nothing brought the values down.

Conclusion

I sadly must conclude that this project would not be fruitful in its current state. Based on the consistently poor cross-validation scores, I believe the problem lies in the amount of data. The infrequent sampling of benthic invertebrate populations and the relatively recent start of the RTOMS data in 2020 restricts the amount of available data for training our models.

However, while the project may not have yielded the desired results, it should not be considered a failure. The models for populations at a depth of 50m displayed relatively low RMSE scores across all water depths. This suggests that predictive models could be developed. With continued data collection over the coming years or an increase in the frequency of sample collection, these predictive models have the potential to improve significantly.

In conclusion, while the current limitations resulted in poor models today, they show potential for future improvement and results. These models could help enhance our understanding of the complex relationships between water conditions and benthic invertebrate populations. Understanding and protecting our local marine ecosystems is paramount, not only for the beauty of our coasts but also for its crucial role in sustaining life.

References

[1] "SCDNR - ACE Basin Characterization," www.dnr.sc.gov. https://www.dnr.sc.gov/marine/mrri/acechar/biological/benthicinvertebrates.html#:~:text=Benthic%20invertebrates%20are%20the%20small

[2] "Ocean Monitoring | Public Utilities | City of San Diego Official Website," www.sandiego.gov. https://www.sandiego.gov/public-utilities/sustainability/ocean-monitoring

[3] "Benthic Invertebrates - Ocean Monitoring Program," City of San Diego Open Data Portal. https://data.sandiego.gov/datasets/monitoring-ocean-benthic-invertebrates/ (accessed Apr. 25, 2024).

[4] "RTOMS Ocean Chemistry - Ocean Monitoring Program," City of San

Diego Open Data Portal. https://data.sandiego.gov/datasets/monitoring-ocean-rtoms-ocean-chemistry/ (accessed Apr. 25, 2024).

[5] "RTOMS Salinity - Ocean Monitoring Program," City of San Diego Open Data Portal. https://data.sandiego.gov/datasets/monitoring-ocean-rtoms-salinity/ (accessed Apr. 25, 2024).

[6] "RTOMS Water Quality - Ocean Monitoring Program," City of San Diego Open Data Portal. https://data.sandiego.gov/datasets/monitoring-ocean-rtoms-water-quality/ (accessed Apr. 25, 2024).

[7] "RTOMS Water Temperature - Ocean Monitoring Program," City of San Diego Open Data Portal. https://data.sandiego.gov/datasets/monitoring-ocean-rtoms-water-temperature/ (accessed Apr. 25, 2024).