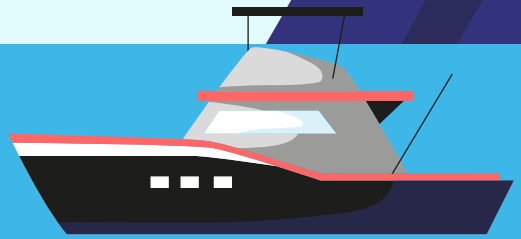
Chemical Compounds' Effect on Anaerobic Oxidation of Methane in the Gulf of Mexico

Kevin Nguyen, Andy Chuang, and Mary-Kate Mahoney



Background and Motivation

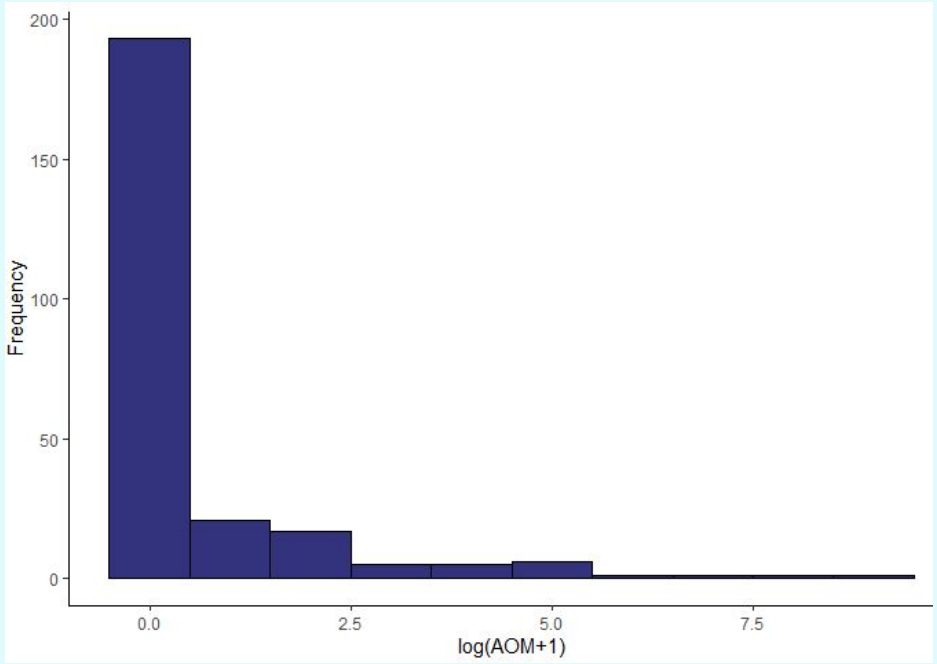
While there are numerous kinds of greenhouse gases responsible for climate change, methane is the most potent of the gases in terms of warming the Earth's climate. Many organisms on the seafloor derive energy from methane and convert it into other compounds through a process called Anaerobic Oxidation of Methane (AOM). Researchers are interested in using soil samples to investigate the precise way micro-organisms perform this methane conversion. Many chemical compounds from soil samples can be easily measured "at-sea," but measuring AOM requires a land-based lab and the process is expensive and time consuming.

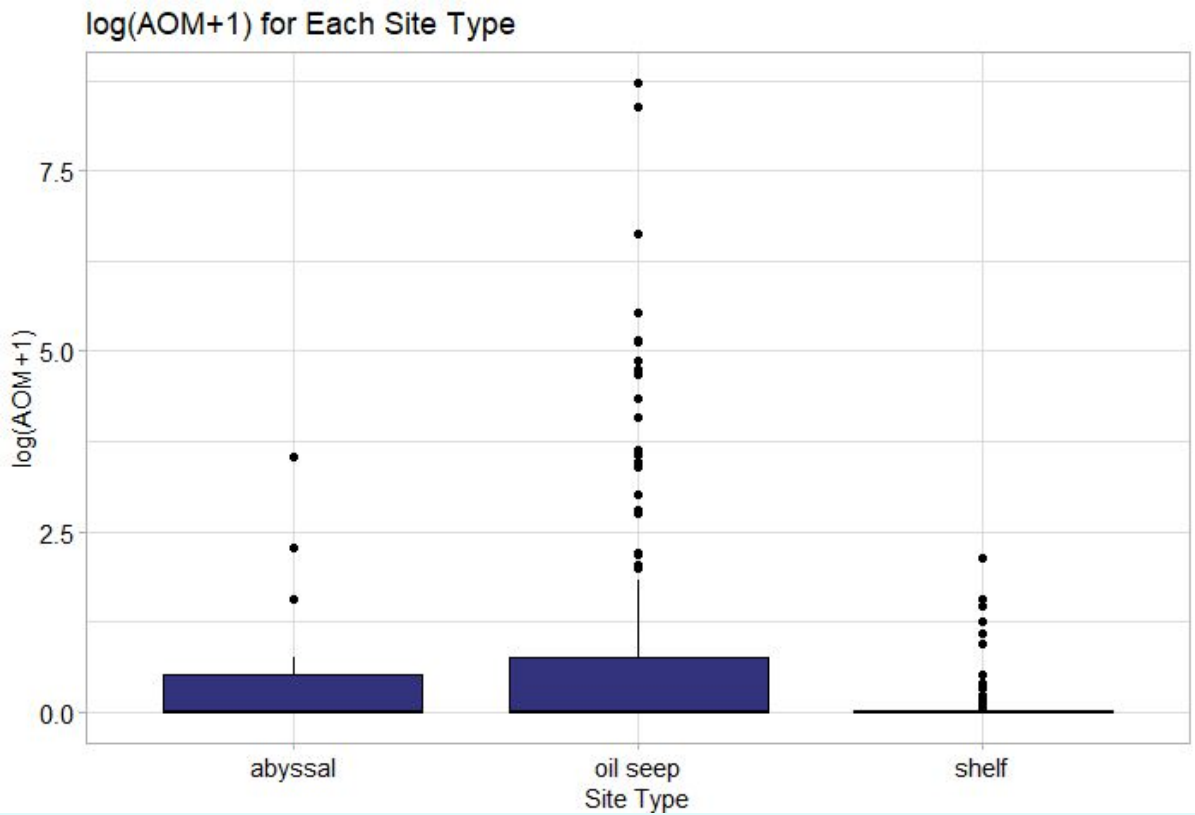
The primary goal of this consultation is to find a method that allows us to use "at-sea" measurements to predict the presence of AOM in soil samples. We also want to identify what attributes of a soil sample are associated with AOM.

Data Cleaning

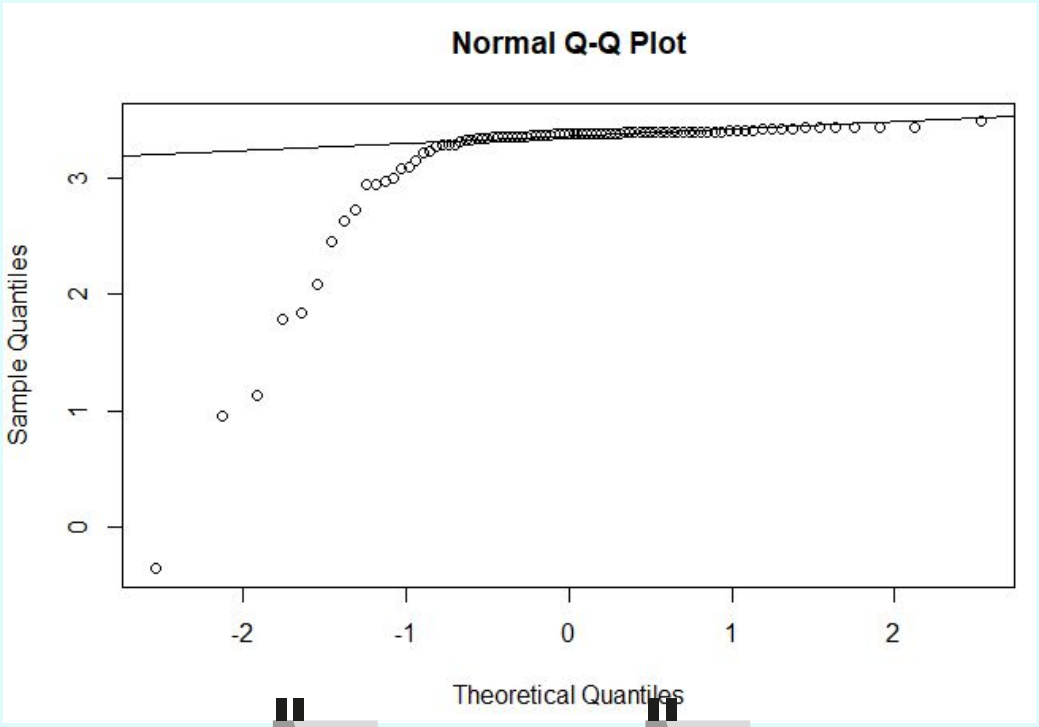
- Remove unnecessary columns
  - Sample ID
  - Ship ID
  - Site
  - POC
  - DOC
- Convert Site Type to factor (“chr”)
- Used “midpoints” for Sed Depth, using Oily Layer as “0”
  - Convert Sed Depth to numeric (“chr”)
- Create new column that converts AOM to binary variable of 1 if >.01, 0 otherwise (a “binary” AOM column)
- Removed AOM NA values for logistic regression

Exploratory Data Analysis




log(AOM+1) for Each Site Type
7.5
5.0
2.5
0.0
log(AOM+1)
abyssal
oil seep
shelf
Site Type

QQ Plot for AOM



Robust Logistic Regression

- Purpose: Can we predict the presence of AOM with any sort of accuracy using “at-sea” measurements?
- Logistic regression: the process of modeling (through classification) the probability of a discrete outcome given an input variable. Models a binary outcome.
  - 0 = no presence of AOM
  - 1 = presence of AOM (AOM is at least .01)
- Predictors are skewed
  - Sed Depth values are being inputted (midpoint value of 1.5 cm vs 0-3 cm) => there will be errors with this particular predictor
  - Use a robust method: glmrob() function in robustbase package

Robust Logistic Regression (cont).

- Split data into training and test (80% training, 20% test)

```
model <- glmrob(AOM~., family = "binomial", data = train)
```

- Specify family = “binomial”, otherwise a linear regression is fit instead
- Create a confusion matrix with testing data to predict accuracy of model
- Example:

	0	1
0	30	12
1	8	56

- Accuracy = 100* ((30+56) / (30+12+8+56)) = 81.13 (rounded to two decimals)

Breakout Questions

1. How should we account for the 24 NA values for AOM?
2. How do we account for the NA values of the other chemical compounds?
3. Should we use the midpoint for Site Depth or keep them as categorical variables?
4. ~98.5% of AOM values fall within the range of 00 to 171. The other values fall far outside this range (largest is >6,000). Should we include these values in the analysis? Why or why not?
5. Suppose we use the rglm to predict the NA values in the binary AOM column
    a. Do we use a robust linear regression model to create proxy values for the missing AOM values, then include them in the final model?
    b. Or should we just run linear regression without the proxy AOM values (i.e. removing NA values)?

Summary

- Clean data
  - Remove unneeded columns
  - Create new column that converts AOM to binary
- Investigate ways to deal with missing data/outliers/zeros
- Run rglm to predict the presence of AOM in soil samples
- Looking Ahead: run rlm to identify what attributes of a soil sample are important

References

1. Methane and climate change [Internet]. Stanford Earth. [cited 2022Sep14]. Available from: https://earth.stanford.edu/news/methane-and-climate-change#gs.bue487.