Nathan Greenslit Data Analysis Project

Nathan Greenslit

3/18/23

# 1. Summary/Abstract

# 2. Introduction

## 2.1 General Background Information

Each year, plumes of Saharan dust travel across the Atlantic via easterly trade winds. These plumes are then deposited in the surface waters of the mid-Atlantic, Caribbean, and Gulf of Mexico. These dust aerosols harbor a wide range of fungi, bacteria, virus-like-particles, minerals, and nutrients (NO3, Fe, PO4). Addition of these nutrients to otherwise oligotrophic settings can result in large and rapid blooms of potentially harmful microbes, presenting a danger to both marine and human health.

Bacteria belonging to the genus Vibrio are marine opportunistic heterotrophs that are ubiquitous in nature and are among the first to respond to the influx of nutrients. Species of Vibrio are known to cause a variety of diseases in humans ranging from cholera (V. cholerae), to ear and wound infections (V. alginolyticus), and in some extreme cases necrotizing fasciitis and septicemia (V. vulnificus). Vibrio have also been shown to impact marine health, causing coral bleaching (V. coralliiyticus, V. shiloi, V. alginolyticus), and shellfish and fish mortality (V. harveyi, V. parahaemolyticus, V. vulnificus). Vibrio population dynamics are primarily driven by temperature and salinity, with prime conditions between 25-30°C and salinities ranging from 20-35. While there is evidence of these blooms occurring in low-nutrient settings like the Florida Keys, less is known regarding microbial response to dust input in areas with *higher baseline nutrients*.

## 2.2 Questions/Hypotheses to be addressed

How do Saharan dust events influence Vibrio populations in high nutrient coastal waters?

*What I am looking for* I expect to see a more dramatic growth response in the Gulf (low nutrient), whereas the higher nutrient sites may exhibit a dampened growth response since they already have high baseline nutrient levels. If we still see a growth response despite already having supportive background nutrient levels at these sites, this may suggest that there are other constituents in the dust that can elicit a growth response. Nutrients, dust input, salinity, and temperature will be the primary factors of interest.

*How I will analyze it* - Look at Vibrio growth over time series (estimated from qPCR) - Examine relationships between dust input and Vibrio growth - Examine influence of site-specific environmental parameters (Temperature, Salinity, Nutrients) on growth response - Run linear models on factors like: dust x growth and nutrients x growth - NMDS to see which parameters have the strongest influence on growth

# 3. Methods

## 3.1 Data acquisition

Data was collected at 3 sites in Corpus Christi, TX before, during, and after a Saharan dust event (starting on July 7th and ending on July 19th, 2022). Sites represent a gradient in background nutrient levels (Blind Oso and Canals are high, the Gulf is low). Water samples were collected for inorganic nutrients, dissolved and particulate organic matter, and microbial analysis. Vertical profiles of salinity, temperature, pH, and dissolved oxygen were obtained using a YSI ProPlus sonde.

Variables will include Temperature (°C), Salinity, Nitrate, Phosphate, Copies/mL produced using quantitative PCR, and Dust Concentration (AOT) from the naval research laboratory satellite data.

## 3.2 Data import and cleaning

This project works with three primary data sets. (1) Enumeration of Vibrio bacteria as copies/mL from quantitative PCR (copies\_master.csv), (2) Dust concentrations as aerosol optical thickness (AOT) derived from the Naval Research Lab (nrl\_conc.csv), and (3) Temperature and Salinity across the daily time series (ysi\_2.csv). Below is a brief summary of how each data set was cleaned. More details can be found in the supplementary files, containing the code and comments describing what each line does. Please refer to the project\_README.md file located in the NathanGreenslit-MADA-Project folder for details on the repository contents and instructions on reproducibility

### 3.2.1 Cleaning of Total Vibrio Enumerations Data set

Quantitative PCR provides enumeration results as “cycle threshold or Cq” values. This depicts the cycle number in which enough of the DNA target was present to be amplified and thus detected. Lower Cq values correspond to higher target concentrations (as it takes less cycles to amplify), and vice verse for higher Cq values. The cleaning of this data set consisted of converting these Cq values into something that can be used to quantify Vibrio (copies of target per mL of seawater). To do so, I needed to take into account the amount of water sample that was concentrated on a filter, the total amount of DNA that was eluted following a DNA extraction, and the amount of DNA template added to each mix for qPCR (to name a few). Taking these into account, we are able to calculate our way from Cq values to copies of Vibrio per mL of seawater. qPCR was conducted in triplicate, so the last step was to take the average of the three replicates to have a final value per sample. A date column was also added for each sample.

### 3.2.2 Cleaning of Dust Concentration (AOT) Data set

Dust concentration (AOT) was collected at time points 0 hour, 6 hour, 12 hour, and 18 hour. In this script, I made different data sets based on time points as well as a data set containing the summed dust concentration (of all time points) per day. For downstream statistical analysis, it will be helpful to have these different data sets to compare to time of sample collection.

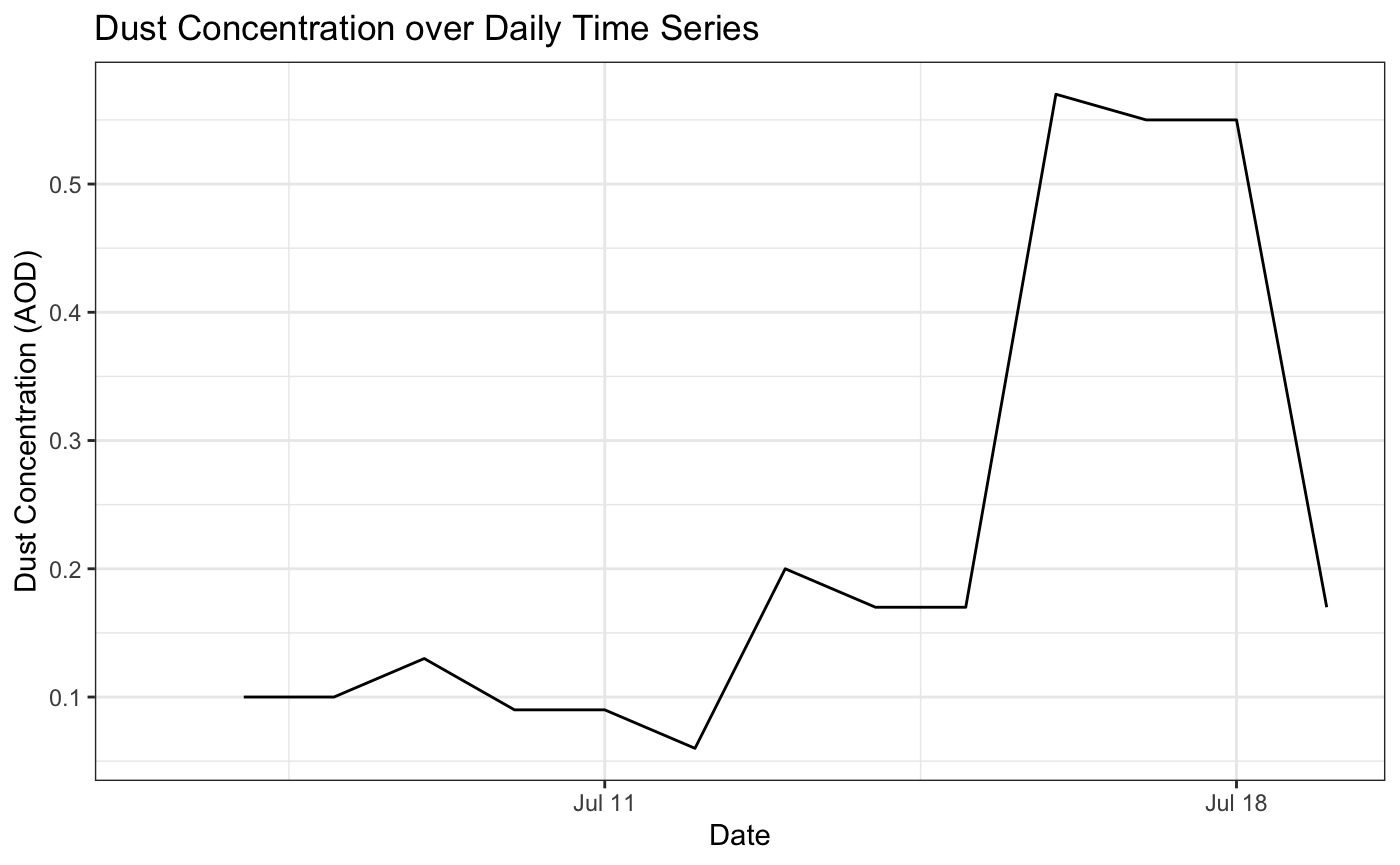
### 3.2.3 Cleaning of Temperature and Salinity Data set

This data set did not require any cleaning.

# 4. Results

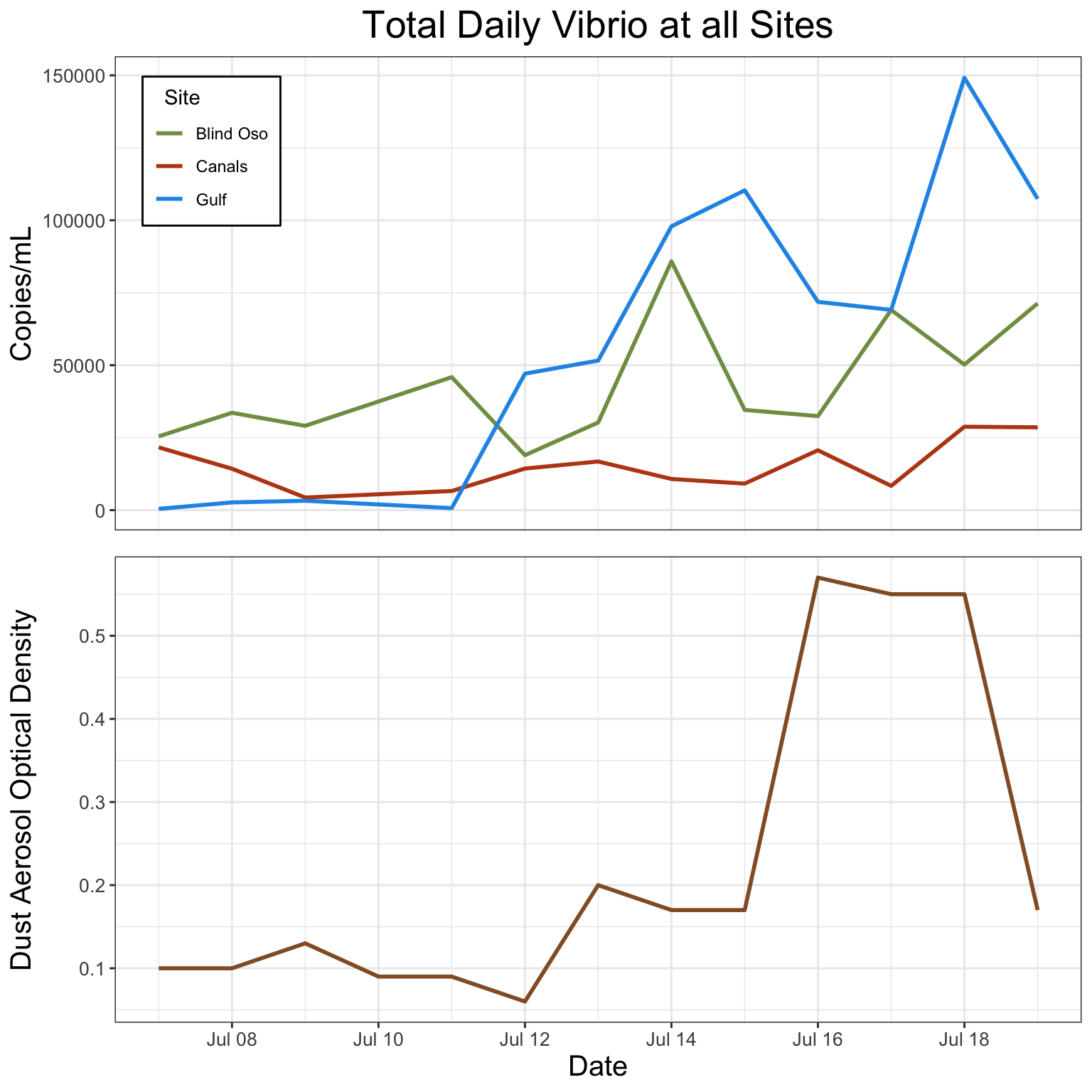
## 4.1 Exploratory/Descriptive analysis

### 4.1.1 Dust



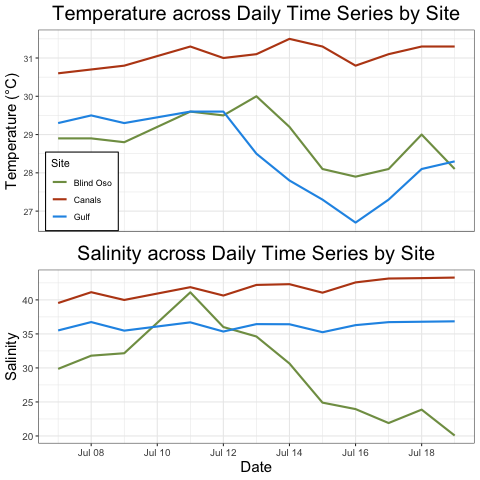
This figure depicts the summed dust concentration 24hr prior to sample collection (time points 13hr–> 7hr the next day) across the daily time series. During the time series, two small periods of higher dust AOT occurred on the 9th and 13th respectively. A much larger spike in AOT was observed on the 16th.

### 4.1.2 Daily Vibrio



The top portion of the figure depicts total enumerated Vibrio as copies per mL, with color by site. There is a noticeable shift in copies in the Gulf site, a less noticeable but still present shift in Blind Oso Bay, and practically no changes in the Canal site. The bottom portion of the figure depicts the dust concentration (AOT) across the daily time series. Initial shift in copies per mL at the Gulf and Blind Oso occur 24-48hr following initial dust introduction on the 13th, and experience another shift around 24-48hr of higher dust input on the 16th.

### 4.1.3 Temperature and Salinity

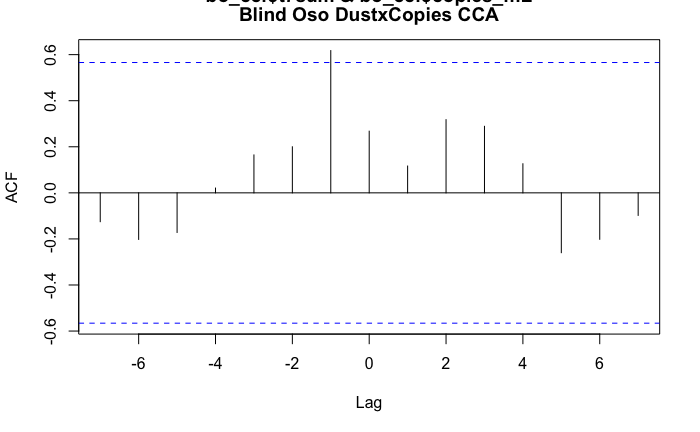


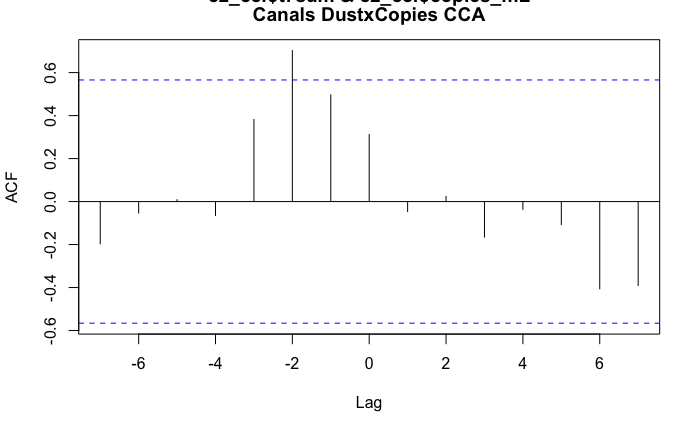
Temperature and salinity at the Canal site is relatively high and consistent as compared to the other sites. Blind Oso Bay experiences some shifts in temperatures and drastic fluctuations in salinity. The Gulf site had a drop in temperature (possibly due to upwelling) and relatively consistent ocean salinities

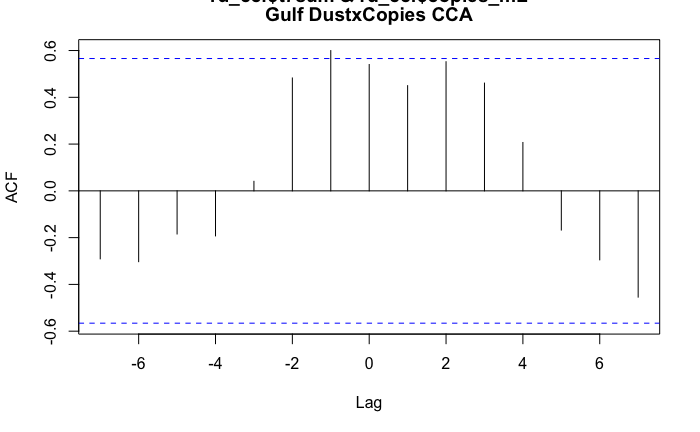
## 4.2 Basic statistical analysis

### 4.2.1 Cross-correlation Analysis

Previous literature has shown that Vibrio respond 12-24 hours following dust deposition events. Based on this, we need to look at lags between the two times series for dust data and growth data. To do this, we can set up a cross-correlation analysis, which allows us to examine the lag/lead relationship between the two variables.



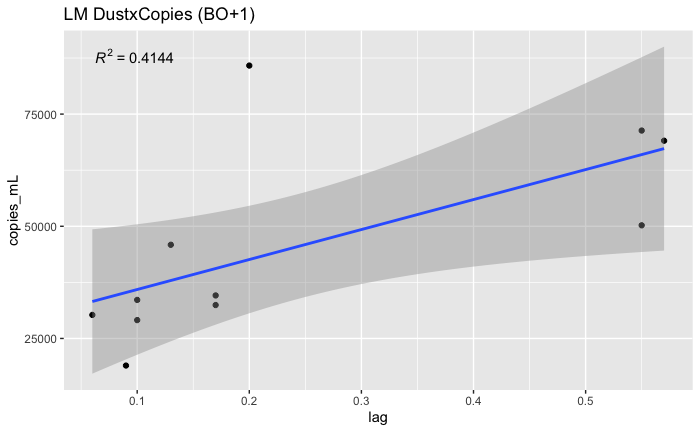


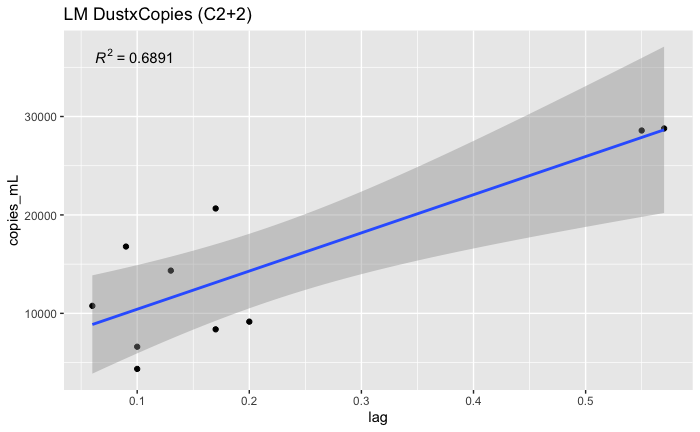


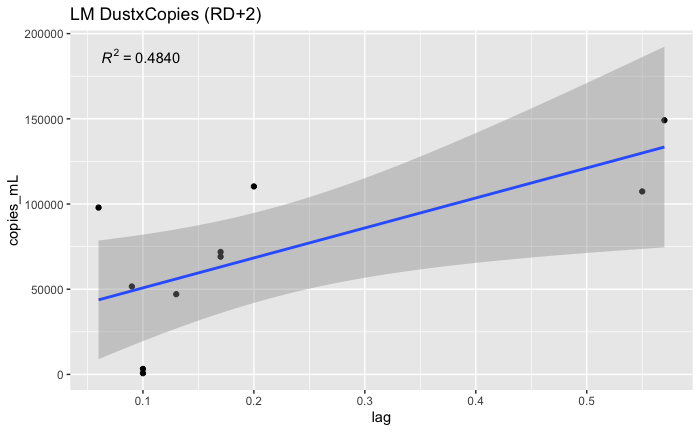
According to the ccf plots, BO has a significant lag relationship a value of -1. In other words, dust occurs, and one day later, we see a response in copies\_mL. C2 has a lag at -2, and RD at -1.

**One important thing to note**: Our data set contains dust data for all of 2022. Our copies\_mL data only contains values from 7/7/22 –> 7/19/22. Therefore, these CCA were made with only this time series. This limits our analysis, as we have dust data prior that can explain copies\_mL on the earlier dates 7/7 or 7/8. But we cannot incorporate this into the CCA as we have NA for copies\_mL for those early days. Below we will run linear models that DO contain these prior dust days. And since we have identified a lag, we can shift the dust data to line up with the copies\_mL that it corresponds with directly - and we no longer have NAs. But this ultiamtely means that the lag identified above. may not be the best fit for our model below, so some additional work is needed that tests other lags in the code.

### 4.2.2 Linear Regression Models







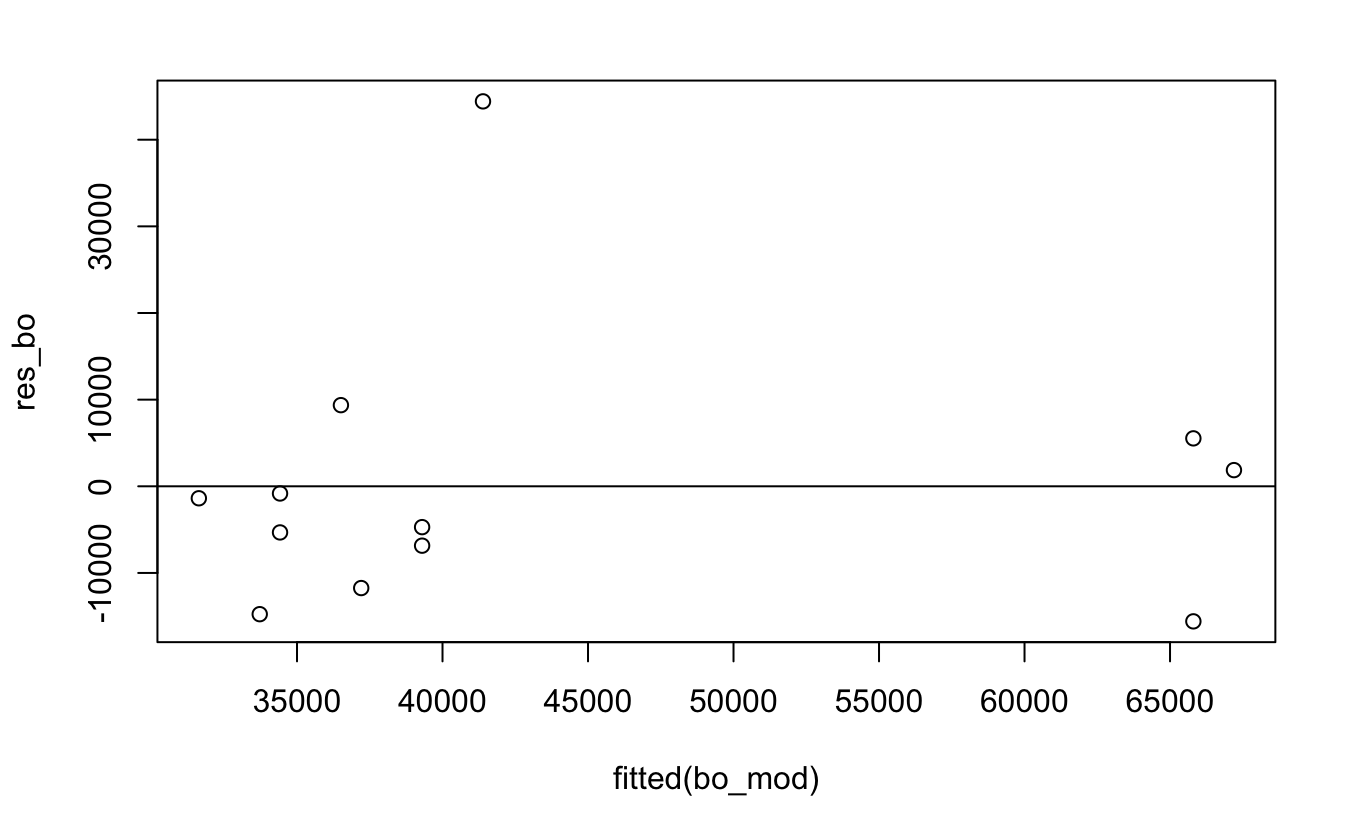
Above are linear regressions that examine the relationship between dust concentration and copies\_mL. As mentioned above, the lags identified may not correspond to the most significant relationship between the two variables. While a lag of -1 for Blind Oso and -2 for the Canals explains the data well, a lag of -2 for the Gulf site was more fitting. This makes sense, as taking a look at the copies\_mL plot above, we do see a ~2 day lag in response, where we have a spike in dust, and 2 days later, a spike in copies\_mL.

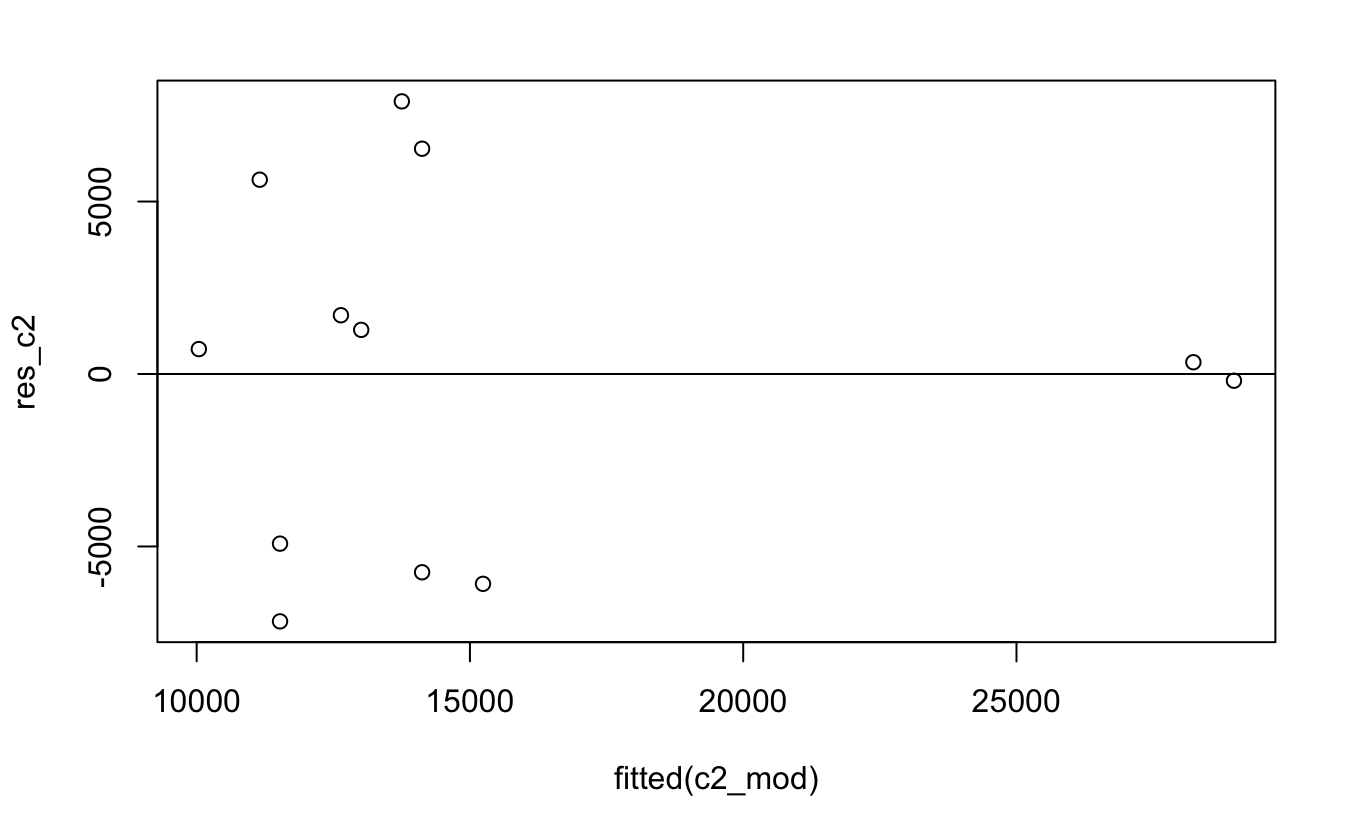
A few things to note:

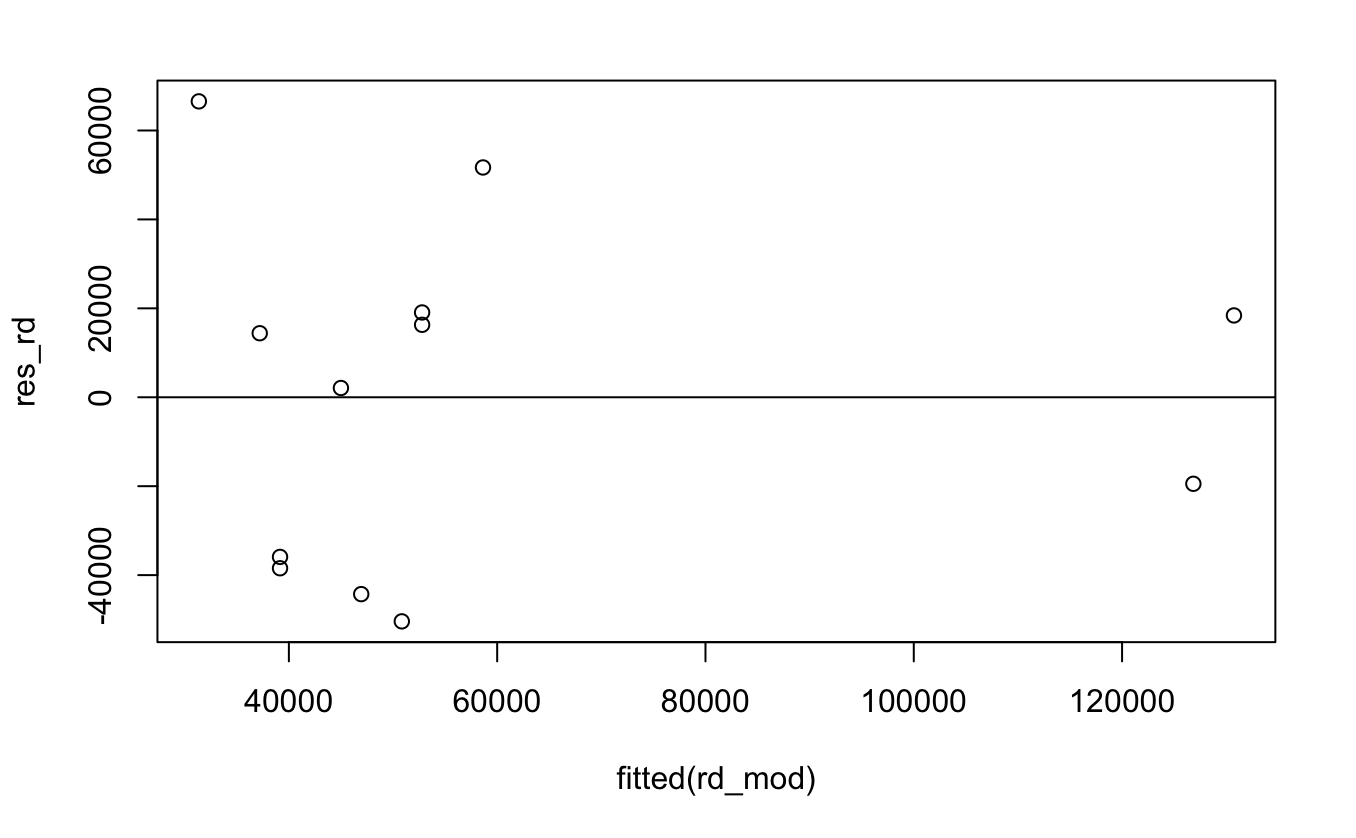
* My stats code contains some chunks that create predicted values based off of the current data set. Since my data set is small, I am not able to portion of ~25% to use to test my model as it could impact the strength of my model. Therefore, plotting the predicted vs. actual will give an R2 of 1.00. I am continuing to find a good validation method.
  + After reviewing module 9, it looks like Information Criteria may be a proper alternative (REF: MOD9/Model Performance and Overfitting).
    - Will need to interpret findings as exploratory and hypothesis generating, and will need to be careful about trying to draw generalizable conclusions.
    - AIC:
      * Blind Oso = 271.25
      * Canals = 243.97
      * Gulf = 292.13
* I am also planning on running a multi-variate model to look at multiple predictors of interests (temp, salinity, nutrients), but am currently running into an issue with my model output, where rather than getting values, I receive NaN. This code can be found in my dust\_copies\_stats.qmd code.

### 4.2.3 Residuals:

Instead of (or in addition to) plotting observed versus predicted for continuous outcomes, you can plot the difference between the two. These differences are called the residuals. What you are looking for is a cloud of points with no discernible pattern. If there is a pattern (e.g., an overall skew, or more points above the 0 y-axes than below), it again suggests that there is still some pattern/signal in the data that the model didn’t capture.







## 4.3 Full analysis

# 5. Discussion

## 5.1 Summary and Interpretation

## 5.2 Strengths and Limitations

## 5.3 Conclusions

# 6. References