Trends in Sea Star Abundance on the US West Coast

Paige Duffin Data Analysis Project

2019-11-27

# Notes to Dr. Handel:

A summary of the current status of my project is as follows.

What I have done (if *italicized*, its a work in progress):

* Cleaned data (processing\_script.Rmd)
  + Wrote about data cleaning in final manuscript
  + Saved all relevant results from data cleaning stage and inserted into manuscript
* Performed exploratory analysis (Exploratory\_Analysis.Rmd)
  + *Wrote about exploratory analysis in final manuscript*
  + *Saved relevant results from exploratory analysis stage and inserted into manuscript*
* Performed continuous outcome modeling analysis (Continuous\_Outcome\_Modeling.Rmd)
  + *Wrote about continuous outcome modeling analysis in final manuscript*
  + *Saved relevant results from continuous outcome modeling analysis stage and inserted into manuscript*
* *Performed categorical outcome modeling analysis (Categorical\_Outcome\_Modeling.Rmd)*
  + *Wrote about continuous outcome modeling analysis in final manuscript*
  + *Saved relevant results from continuous outcome modeling analysis stage and inserted into manuscript*
* Figured out how to save files into their correct location and call them in other scripts. This is something that I couldn’t figure out how to do on the last assignment so I’m proud of myself for working through it.

What I haven’t yet done:

* Write abstract
* Write final introduction
* Write discussion
* Figure out how to use Bibtek

Considering how utterly lost and frustrated I was as I submitted my last portion of this project, I am very pleased with where I am at and now feel as if I can actually succeed in completing this project. Thank you for reducing the workload in other aspects of this course so that I could take (and will continue to take) the time I need to really understand how to create a fully reproducible project like this. I can’t BELIEVE how far I’ve come since the beginning of this semester.

# Summary/Abstract

*Write a summary of your project.*

# Introduction

## General Background Information

Outbreaks of infectious disease are an important, yet poorly understood, driving force in population biology and, while prevalent in both terrestrial and marine systems, the manner by which outbreaks influence biotic relationships, age demographics, community structure and function, and trophic interactions in the aquatic realm consistently lags behind that of terrestrial disease ecology. The growing body of evidence in this field, however, supports the observation that marine disease epidemics are increasing in frequency and severity (Orth et al., 2006; Waycott et al., 2009). Elucidating the relationship between aquatic species and their pathogenic diseases is pertinent in the field of marine ecology because infection outbreaks have the potential to drastically alter ecosystem functionality (Blakesley et al., 2002; Burge et al., 2013). Many of the organisms which have faced chronic and/or severe outbreaks of disease (such as sea urchins, scleractinian corals and seagrasses) are also considered keystone species and/or ecosystem engineers, meaning that they contribute highly valuable services or functions to their surrounding community and ecosystem. Thus, it comes as no surprise that disease-driven mass mortalities of these species often generate waves of ecological permutation, ranging from temporary local disruptions to permanent phase shifts (Burge et al., 2013), such as the transformation from a coral to macroalgal dominated Caribbean reef structure that accompanied the massive *Diadema antillarum* (black sea urchin) die-off of the early 1980s (Lessios et al., 1984; Lessios, 1988). This event, likely caused by a biological pathogen (Schultz et al., 2016), reduced Diadema populations to less than 1% of their original size (Lessios, 1988). Mass mortality events impacting critical foundation species, ecosystem engineers, or keystone species such as the black sea urchin have been coined ‘marine disease emergencies’ due to the detrimental cascade of events that often succeeds them (Miner et al., 2018).

Sea Star Wasting Disease (SSWD) is another epizootic crisis facing modern day coastal ecosystems. Outbreaks of asteroid wasting have been documented periodically since the late 1970s and SSWD describes a suite of symptoms observed across a broad range of sea star species, most of which play integral parts in shaping their community structure. Generally speaking, the wasting disease events which punctuated the past four decades were relatively brief, in localized areas, and largely failed to capture the attention of the scientific community. Beginning in summer 2013, however, mass mortalities of sea stars due to wasting disease have caused unprecedented damage, owing to the geographical and temporal extent of impact. This ongoing epizootic, which has killed millions of asteroids across over 20 taxa, is widely referred to as the largest disease event sweeping through a wildlife marine species in documented history (Hewson et al. 2014, Gudenkauf & Hewson, 2015, Eisenlord et al. 2016).

## Description of data and data source

### The Source

The vast majority of disease documentation and surveys cover the west coast of the United States. While it is almost certainly true that SSWD is most prevalent in this region, there is a strong bias towards the discovery of diseased organisms owing to the emphasis of long-term survey networks in the area, such as the extensive [Multi-Agency Rocky Intertidal Network](https://marine.ucsc.edu/index.html) (MARINe for short). The vast majority of what we know about SSWD dynamics comes from studies along the Pacific Coast of North America using data produced through the combined effort of dedicated scientists, government agencies, and local citizens. The purpose of this investigation is to make use of this data monitoring system to investigate spatial and temoporal changes in sea star abundance along the west coast of the United States.

### The Data

Below I explore the raw data I downloaded from a link I recieved after submitting a request for the data on the MARINe webpage.

library(readr)  
raw\_SS\_count <- read.csv("../../data/raw\_data/seastarkat\_size\_count\_totals\_download.csv")  
head(raw\_SS\_count, 3)

## group\_code site\_code marine\_site\_name marine\_sort\_order latitude  
## 1 UCLA ALEG Alegria 6420 34.46714  
## 2 UCLA ALEG Alegria 6420 34.46714  
## 3 UCLA ALEG Alegria 6420 34.46714  
## longitude marine\_common\_season marine\_season\_code marine\_common\_year  
## 1 -120.2774 85 SP02 2002  
## 2 -120.2774 85 SP02 2002  
## 3 -120.2774 85 SP02 2002  
## season\_sequence season\_name target\_assemblage method\_code species\_code  
## 1 1 Spring sea\_star IP PISOCH  
## 2 1 Spring sea\_star IP PISOCH  
## 3 1 Spring sea\_star IP PISOCH  
## size\_sort\_order size\_bin total mpa\_designation mpa\_region georegion  
## 1 3 20 1 reference South Coast CA South  
## 2 4 30 3 reference South Coast CA South  
## 3 5 40 16 reference South Coast CA South  
## bioregion state\_province island last\_updated  
## 1 Government Point to Mexico California Mainland 2019-04-05 20:07:59  
## 2 Government Point to Mexico California Mainland 2019-04-05 20:07:59  
## 3 Government Point to Mexico California Mainland 2019-04-05 20:07:59

Notably, it contains the following information (some parameters removed from this overview):

* groupcode: The unique code for each monitoring group
* marine\_site\_name: The name of the site where the survey was conducted.
* latitude and longitude: (self explanatory)
* marine\_season\_code: A four-character code to identify the Sampling season. The first two characters indicate the season and the last 2 characters indicate the year.
* method\_code: The method used for sampling the plot.
* size\_bin: The size of the species being counted, binned to the nearest 5 or 10 millimeter.
* target\_assemblage: The specific organism that this transect was created to monitor.
* total: Total number of individuals counted in a given size\_bin
* mpa\_designation: Describes whether the referenced site is located within a Marine Protected Area (MPA) or is a reference site. If this field is blank, the referenced site is not located within an MPA, or is not a reference site for an existing MPA site.
* georegion/ bioregion: geographic/biogeographic region in which site is located
* state\_province: The State or Province and Country where the referenced site is located.
* island: The name of the island where the referenced site is located. Sites not on islands are designated as mainland.

## Questions/Hypotheses to be addressed

*State the research questions you plan to answer with this analysis*

**1. Are sea star populations changing over time?** I anticipate that there will be a decline in sea star populations over time, but that the change will not occur uniformly across geographical regions and species. Species that are reported to have been affected by SSWD will experience the biggest declines.

**2. Are there definable geographical regions throught space based on predictor variables?** I expect to be able to define rough species ranges.

**3. Is there evidence of sampling bias?** Is there a relationship between important survey parameters (mainly, abundance data) and sampling method (ex. group conducting the survey, plot sampling method, the target assemblage). At the very least, I expect there will be bias based on the targeted organism of the survey; if a given survey aims to characterize the presence of one species of sea stars, it is likely that they will seek out that species and record higher counts.

## Methods and Results

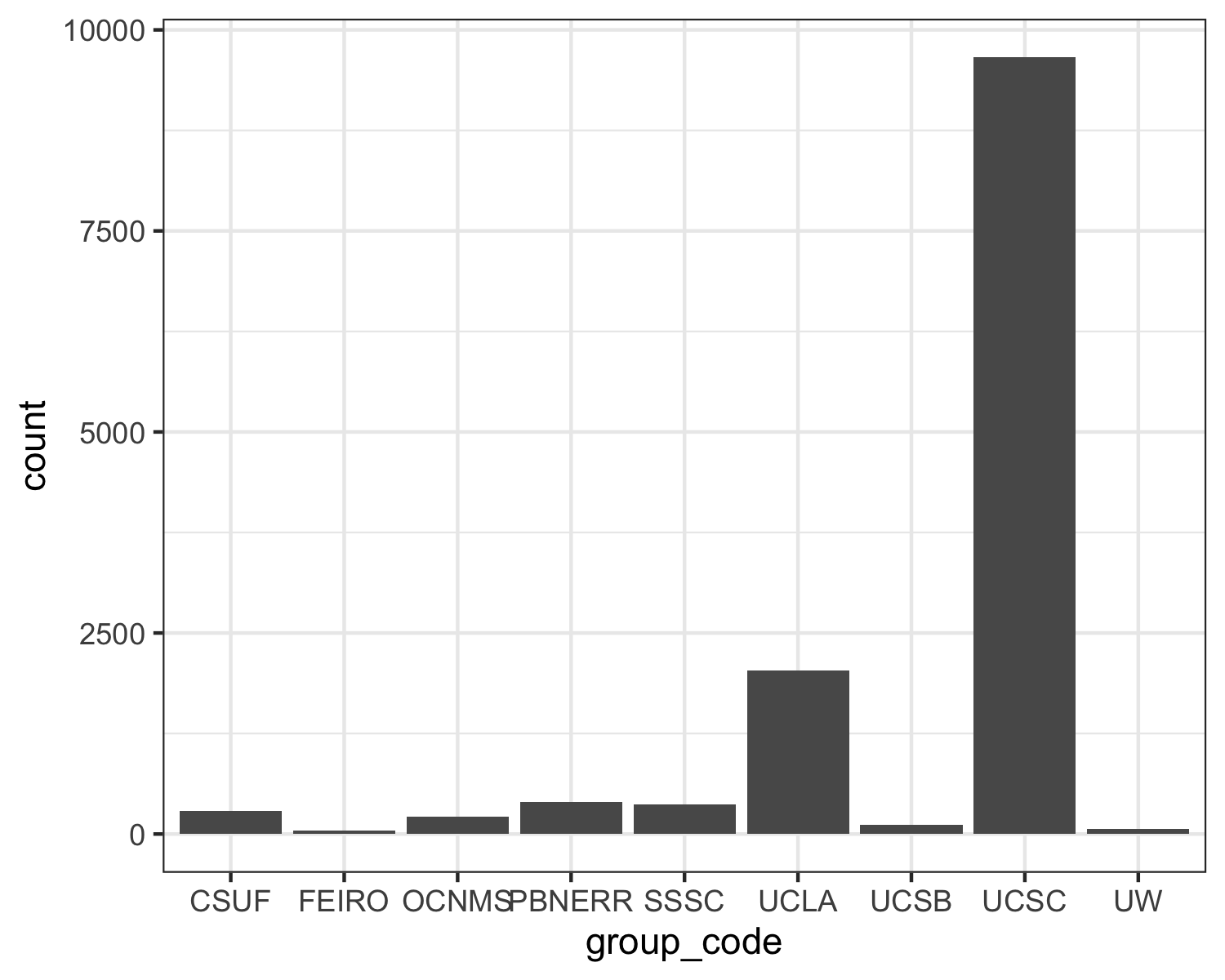
*In most research papers, results and methods are separate. You can combine them here if you find it easier. You are also welcome to structure things such that those are separate sections.*

### Data import and cleaning

*Write code that reads in the file and cleans it so it’s ready for analysis. Since this will be fairly long code for most datasets, it might be a good idea to have it in one or several R scripts. If that is the case, explain here briefly what each file does. The files themselves should be commented well so everyone can follow along.*

The original data was downloaded from a link I recieved after submitting a request for the data on the Multi-Agency Rocky Intertidal Network, or MARINe, [webpage](https://marine.ucsc.edu/index.html). The raw data is found here: ./data/raw\_data/seastarkat\_size\_count\_totals\_download.csv.

This data was cleaned extensively prior to use in this analysis, and the code which performs this cleaning is found here: ./code/processing\_code/processing\_script. Briefly, this involved examining each variable by plotting factors as bar plots and integers/numerics as histograms. This allowed me to observe the structure of each variable and make necessary changes to clean the data. Below, I describe some of these changes:

Some of the factors were very skewed towards one group. For example, group\_code was very biased towards UCSC.  In this example, I made a second variable that describes what group surveyed the data at that sampling point, titled group\_code\_UCSC\_other that lumped all other levels of the group\_code factor into a new level called Other besides the most frequent level (here, UCSC). I repeated this process with the variable method\_code.

There was a discrepency between the variables mpa\_region and mpa\_designation (see processing\_script for more details). As a result, I removed the less informative variable, mpa\_designation.

I renamed factor levels that were excessively verbose.

I removed the last\_updated variable, as all values were equivalent and, therefore, uninformative.

I discovered that the type of species surveyed was heavily skewed towards *Pisaster ochraceus*, and made a mental note that I truly only am interested in *Pisaster ochraceus*, so I should strongly consider performing analyses on a subset of the data that focuses only on one species at a time (or even just *Pisaster ochraceus*)

The cleaned data was then saved as a new file here: ./data/processeddata.rds.

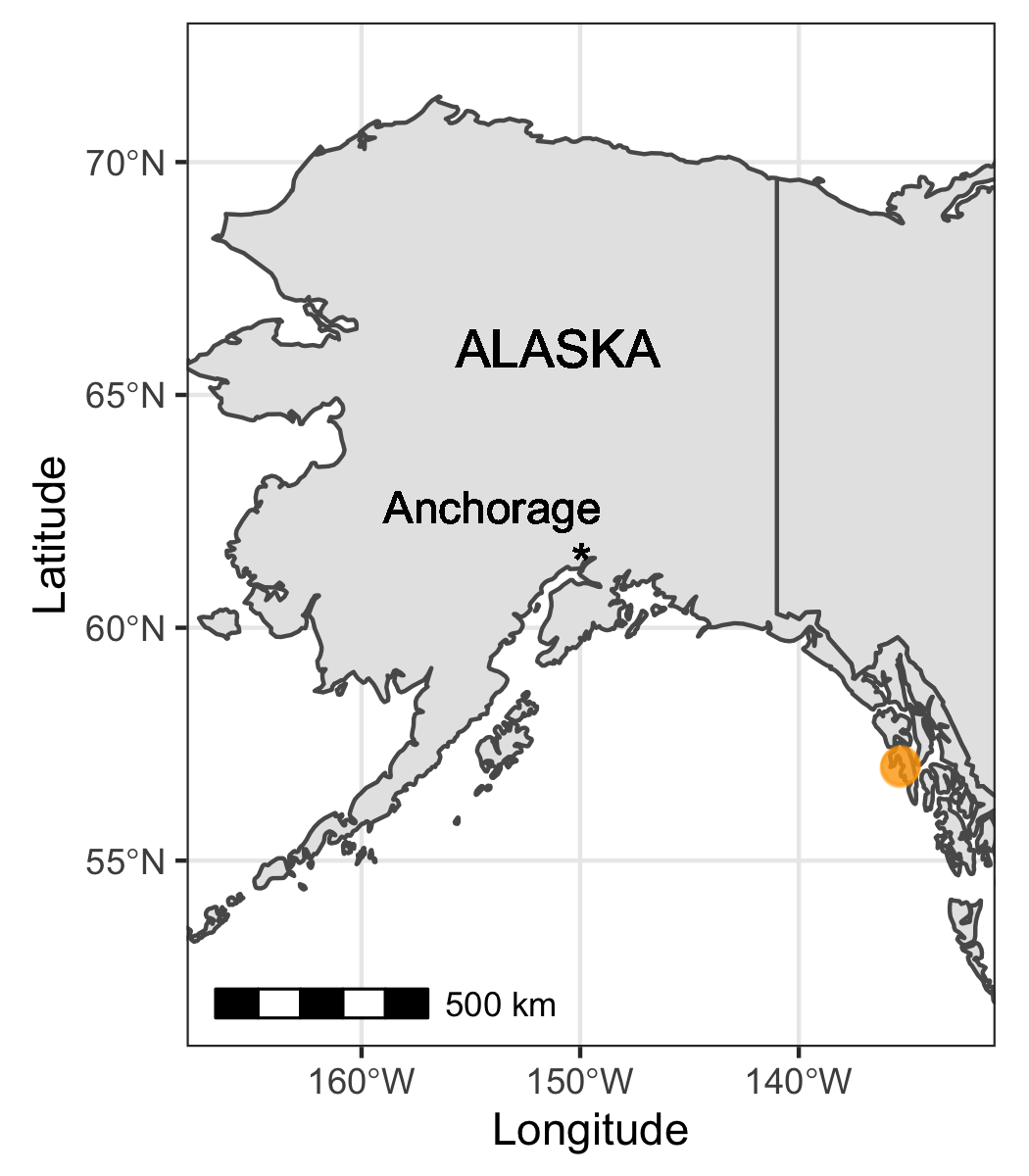
### Univariate analysis

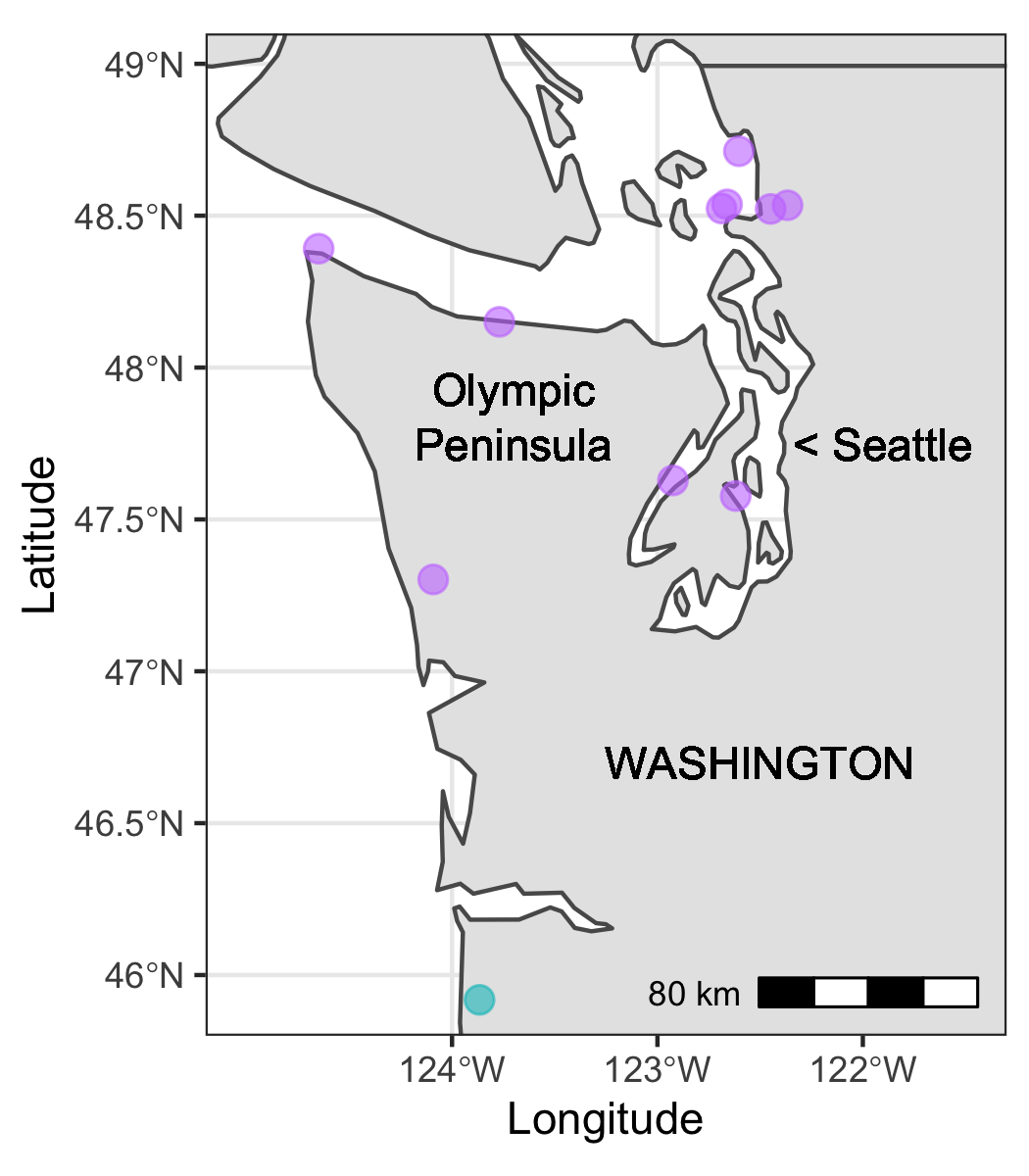
*Use a combination of text/tables/figures to explore and describe your data. You should produce plots or tables or other summary quantities for most of your variables. You definitely need to do it for the important variables, i.e. if you have main exposure or outcome variables, those need to be explored. Depending on the total number of variables in your dataset, explore all or some of the others.*

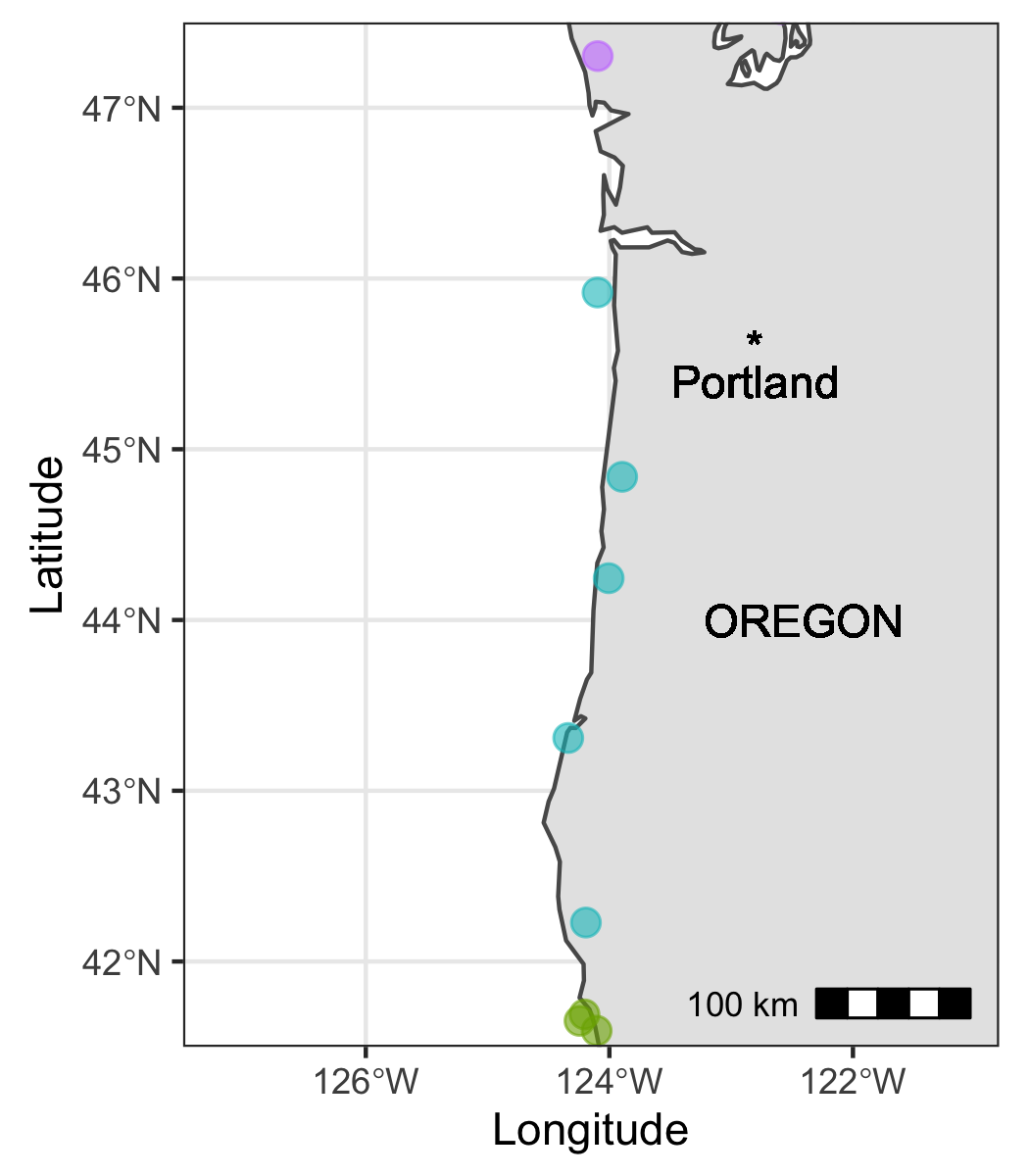
The code which performed my univariate analyses can be found here: ./code/analysis\_code/Exploratory\_Analysis.Rmd

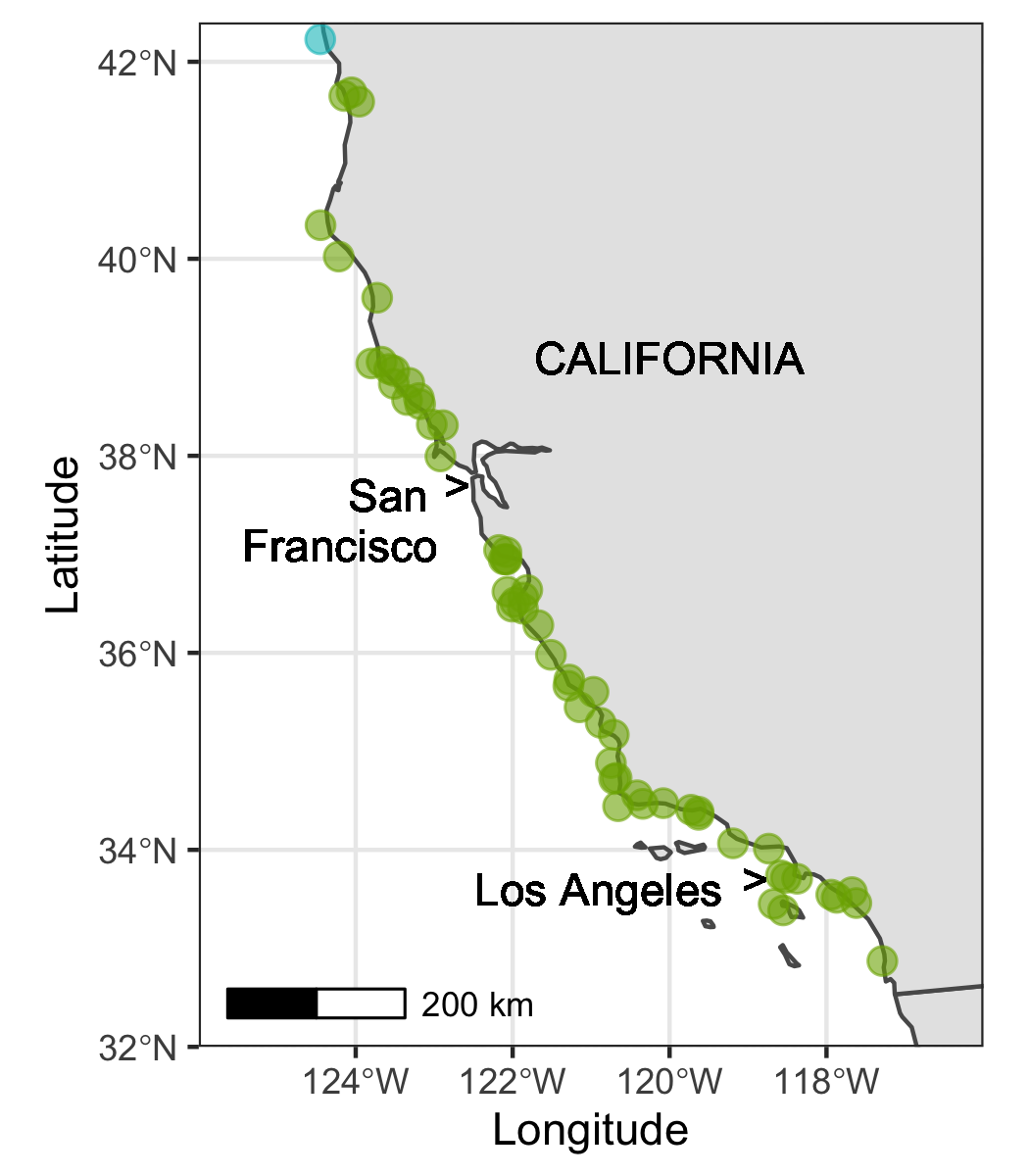
#### Mapping Survey Sites

First, I created geographical maps to visualize the sample sites. Sites were sampled in four states:

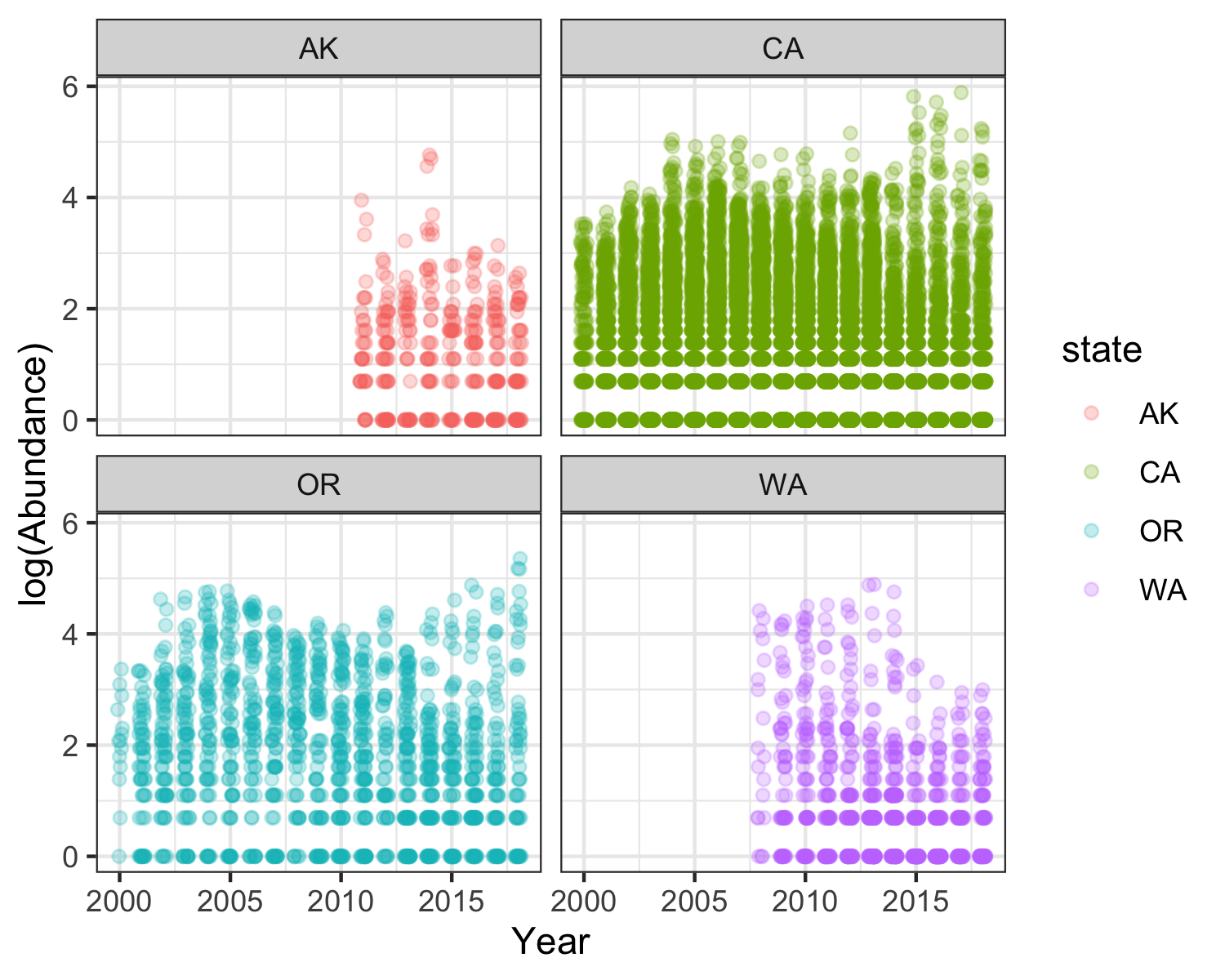
Alaska: 

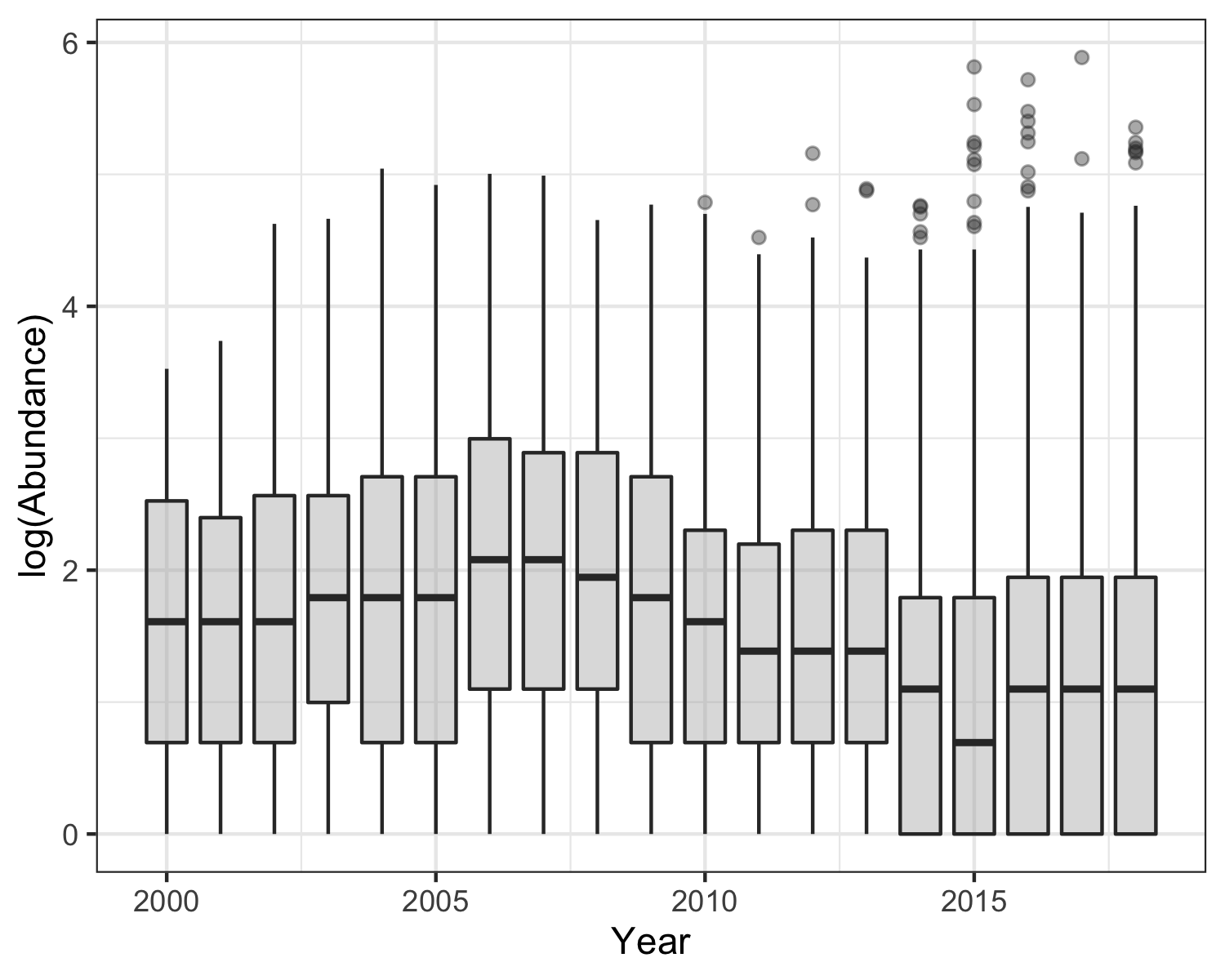
Washington: 

Oregon: 

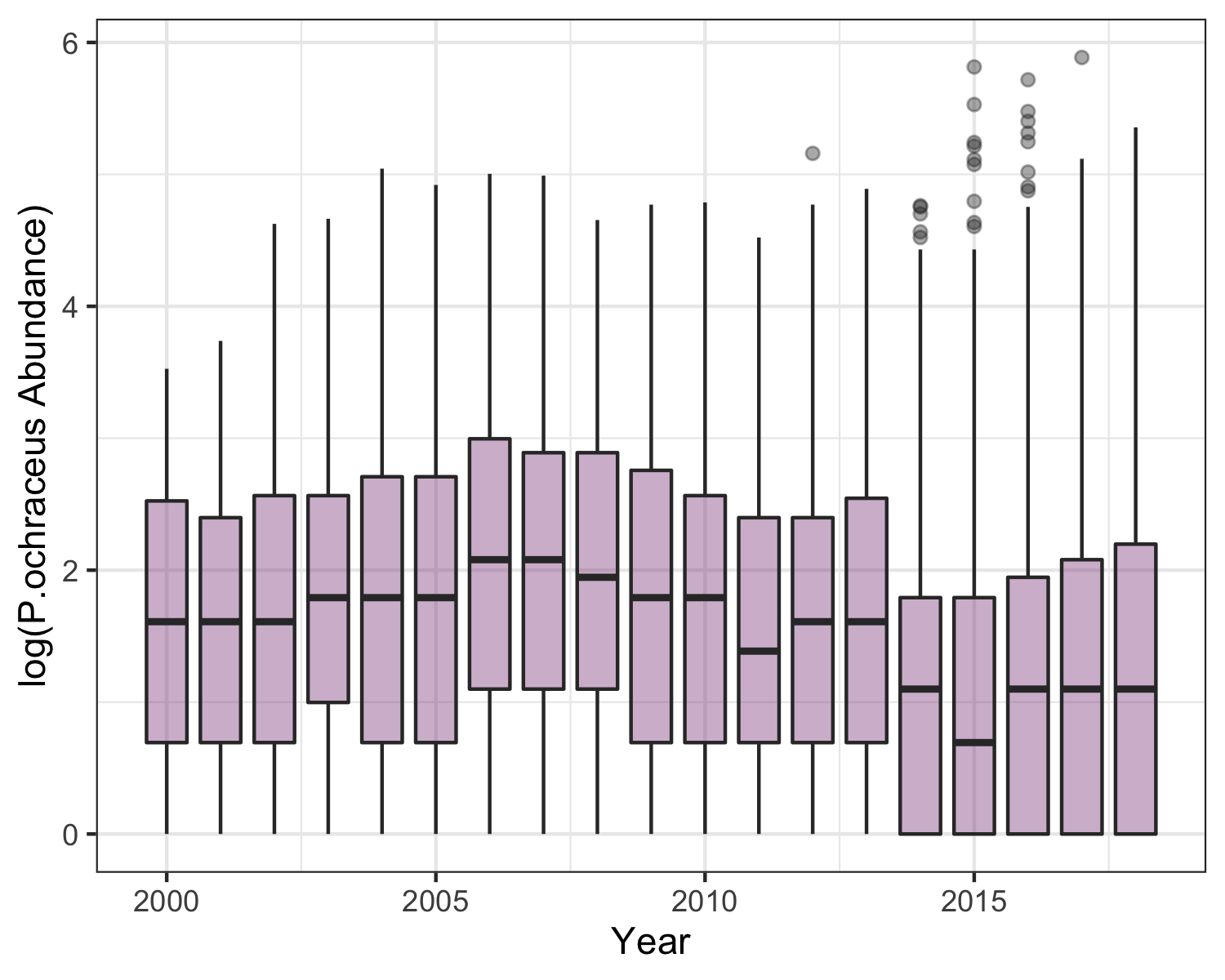
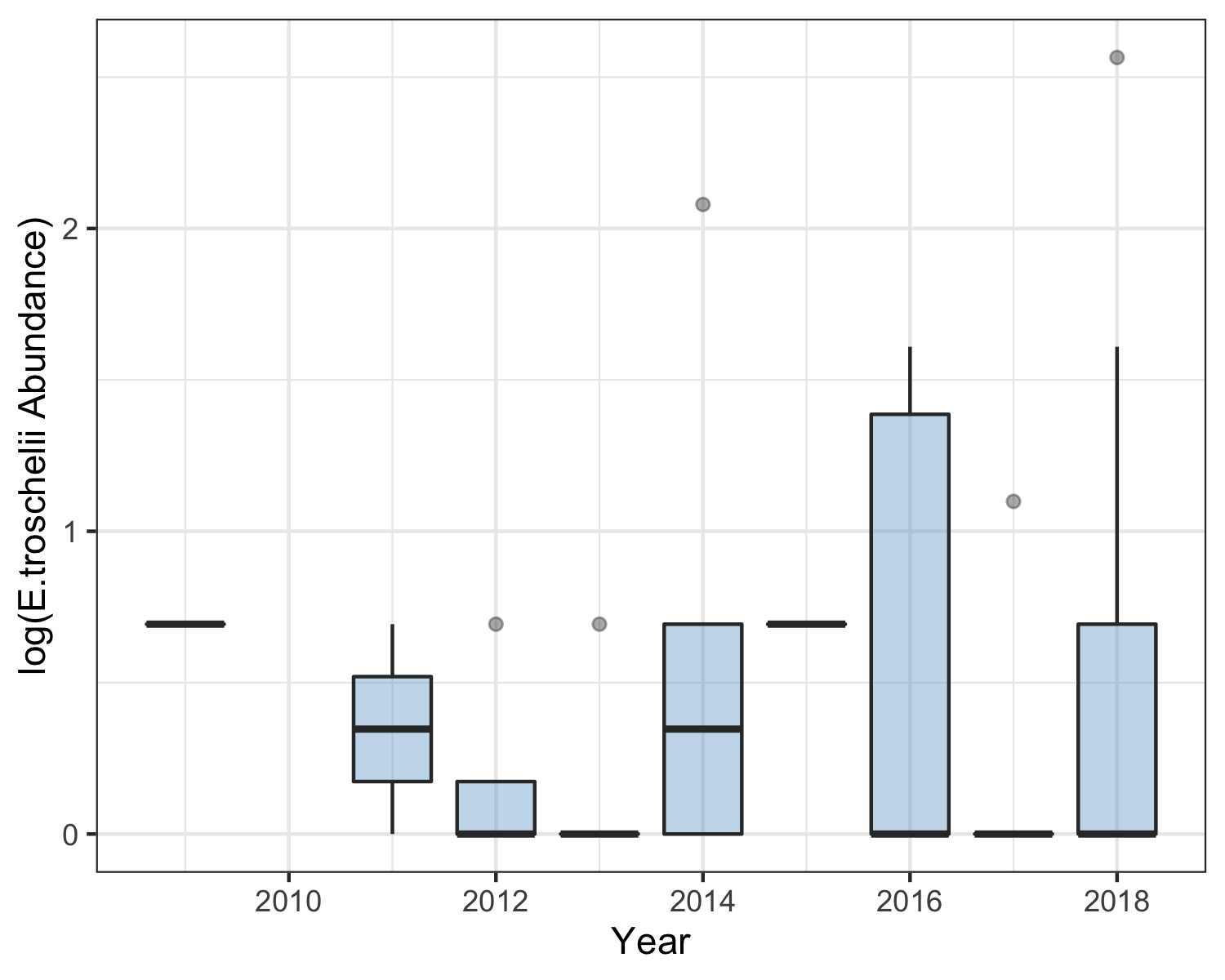
and, finally, California: 

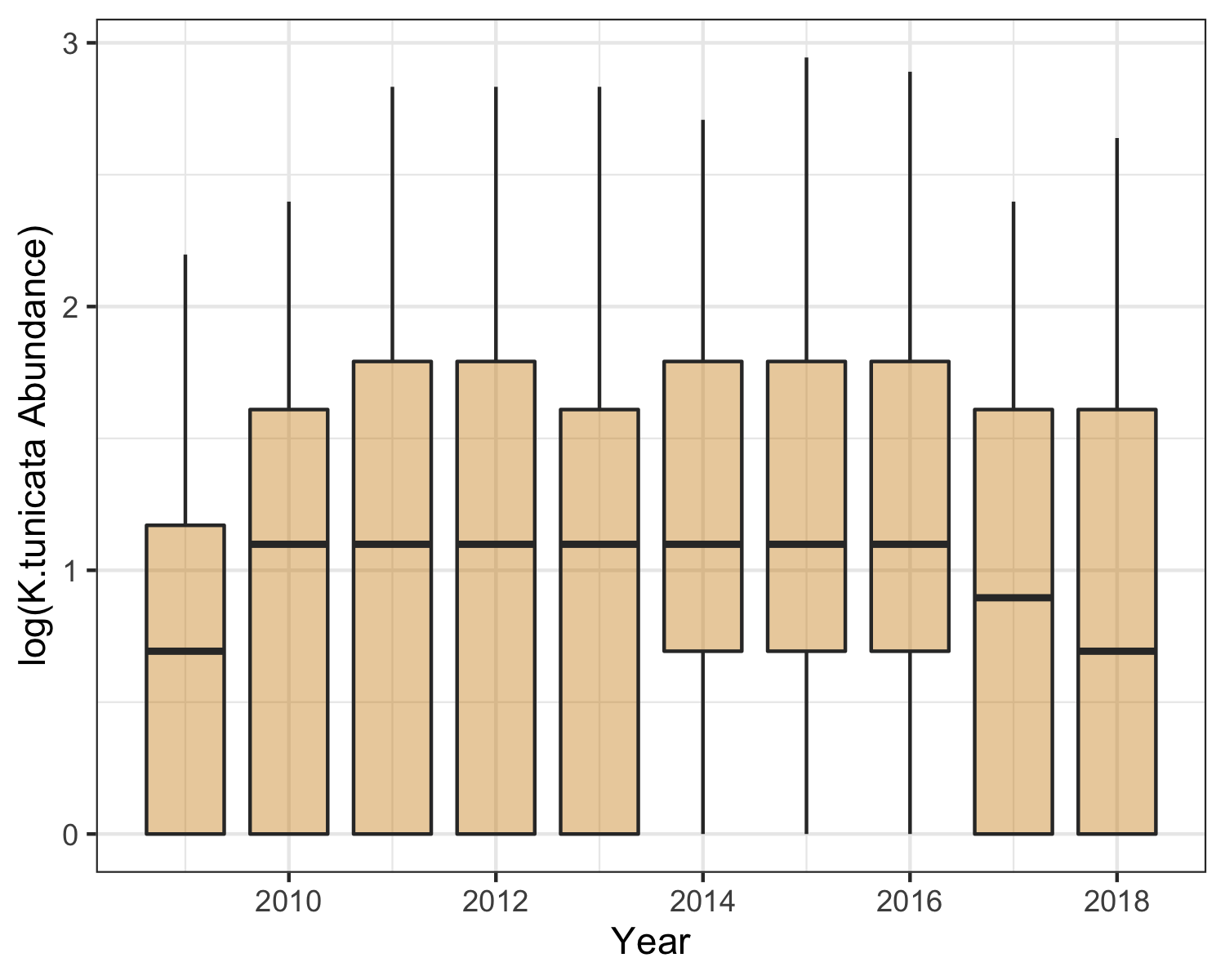
#### Exploring Count Data by State and Year Surveyed

I was interested in exploring the structure of the abundance (count) data across space and time. I opted to take the log of the abundance counts to make the plots more readable.  This graph shows… [FINISH]

Here’s a box-plot showing the same information across all sites/states: 

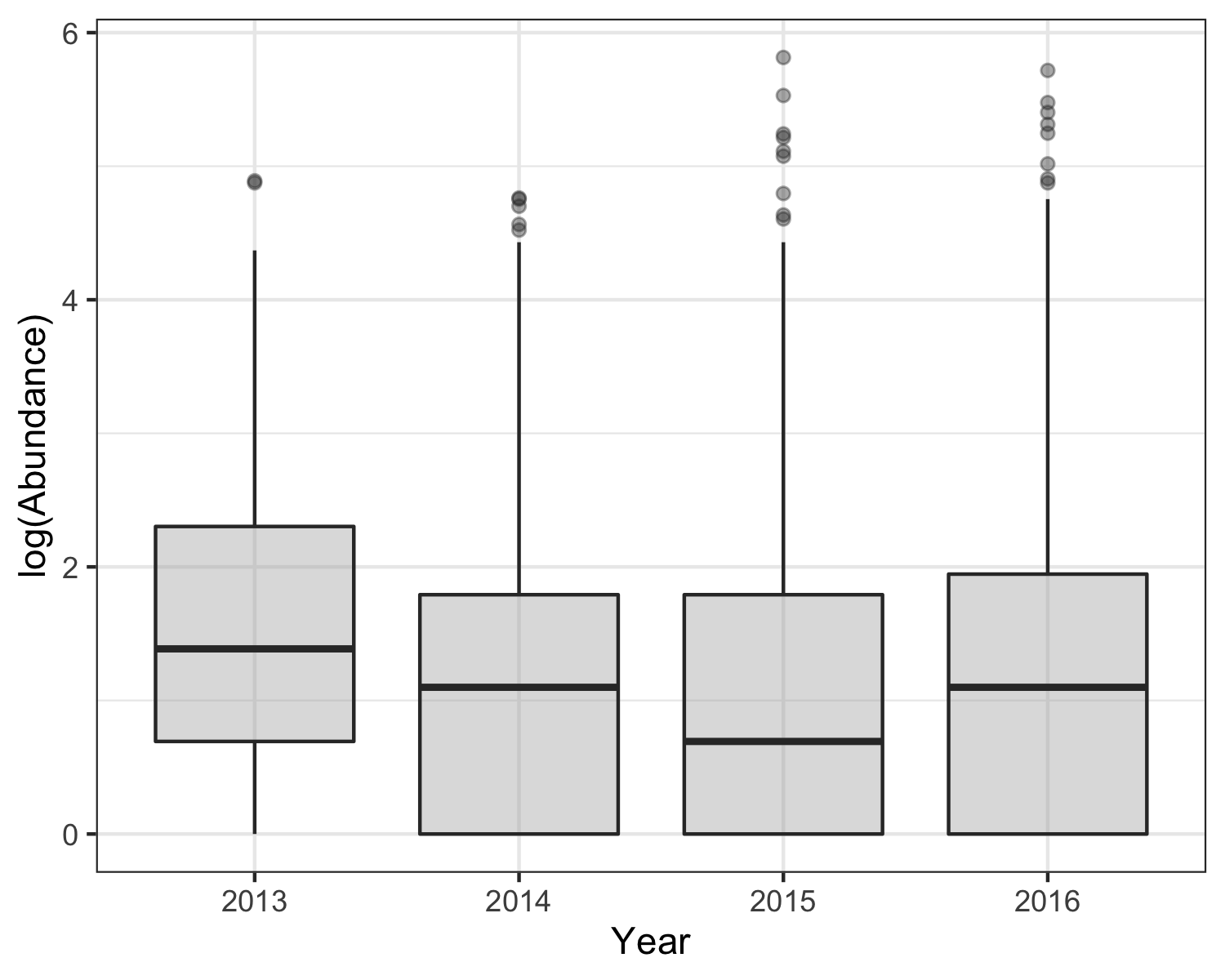
Then, I wanted to know how this relationship looked if I broke the analysis up by species.

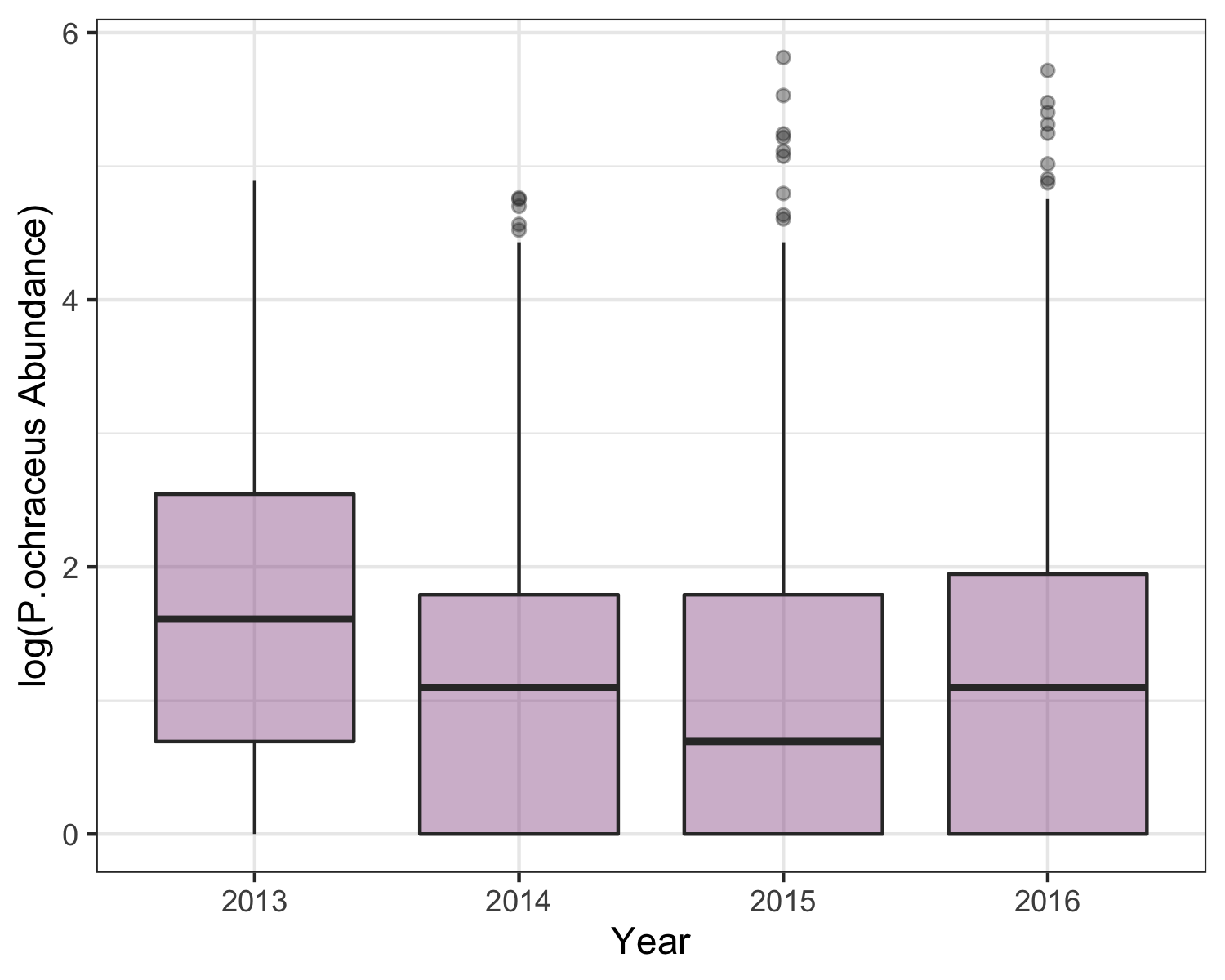
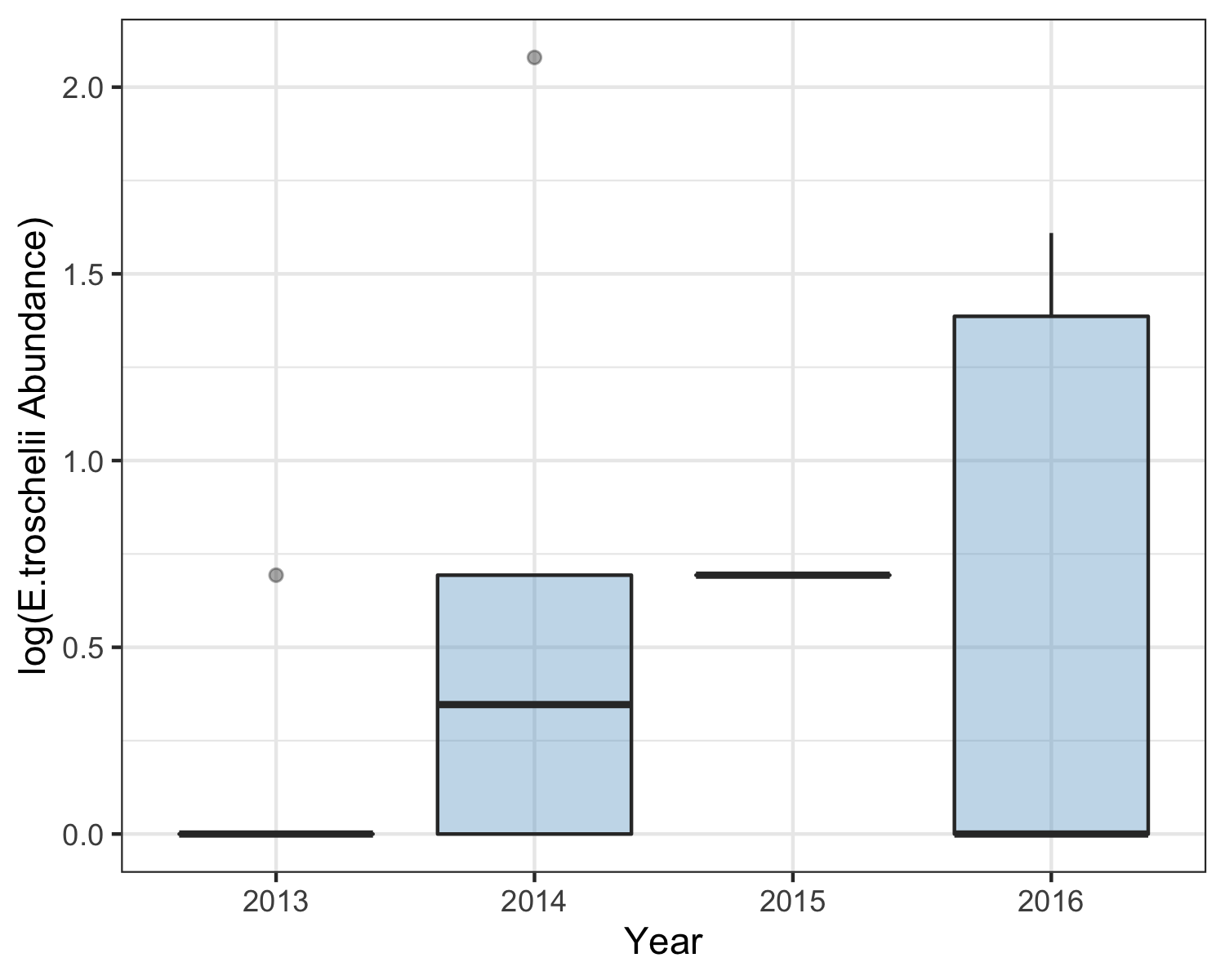
P.ochraceus:  E.troschelii: 

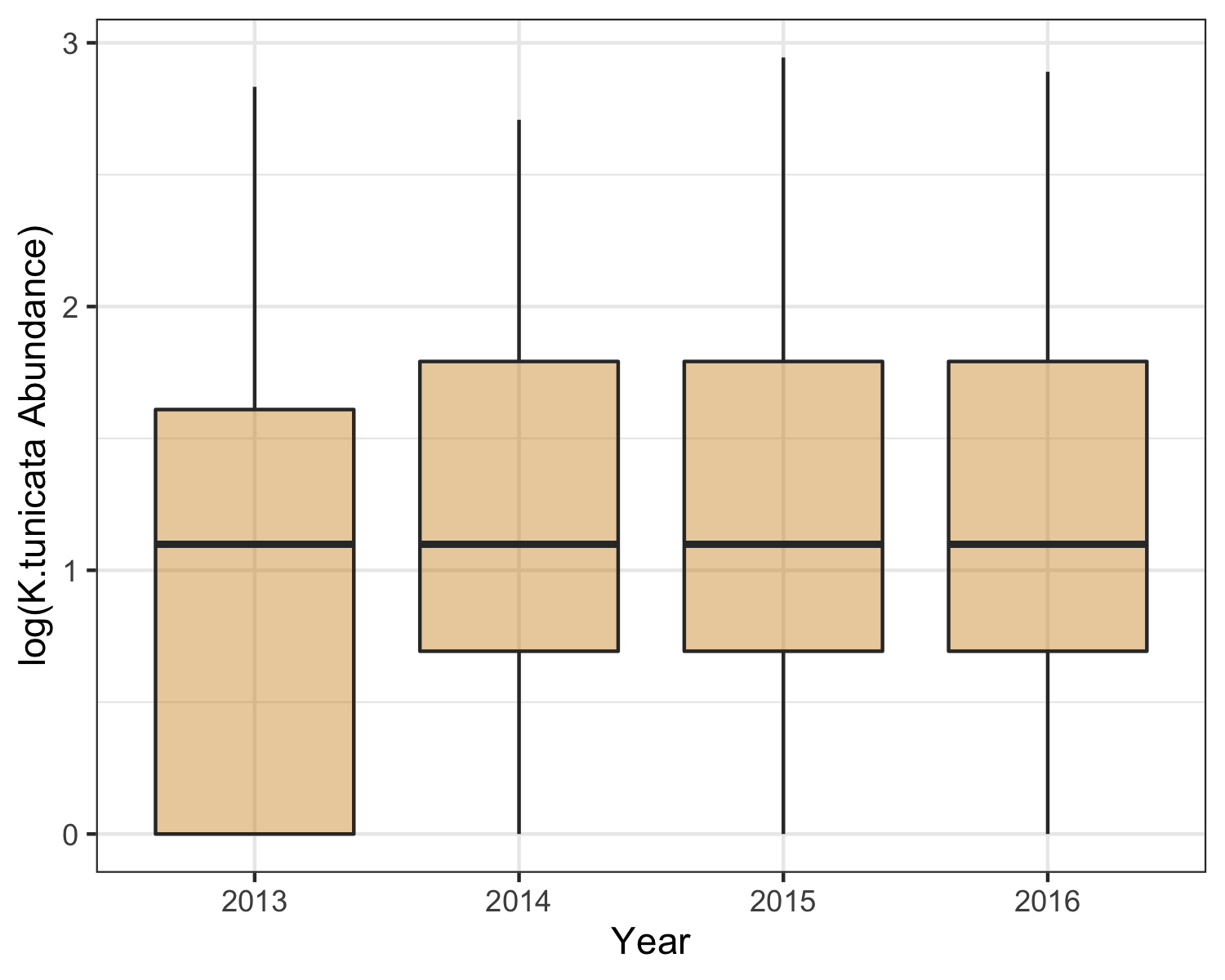
K.tunicata: 

Specifically, I’m interested in this data set through the lens of a very important piece of background knowledge: there was a massive die-off of sea stars on the west coast starting around 2014. So, I narrowed the time scale of my analysis a bit: 2013-2016.

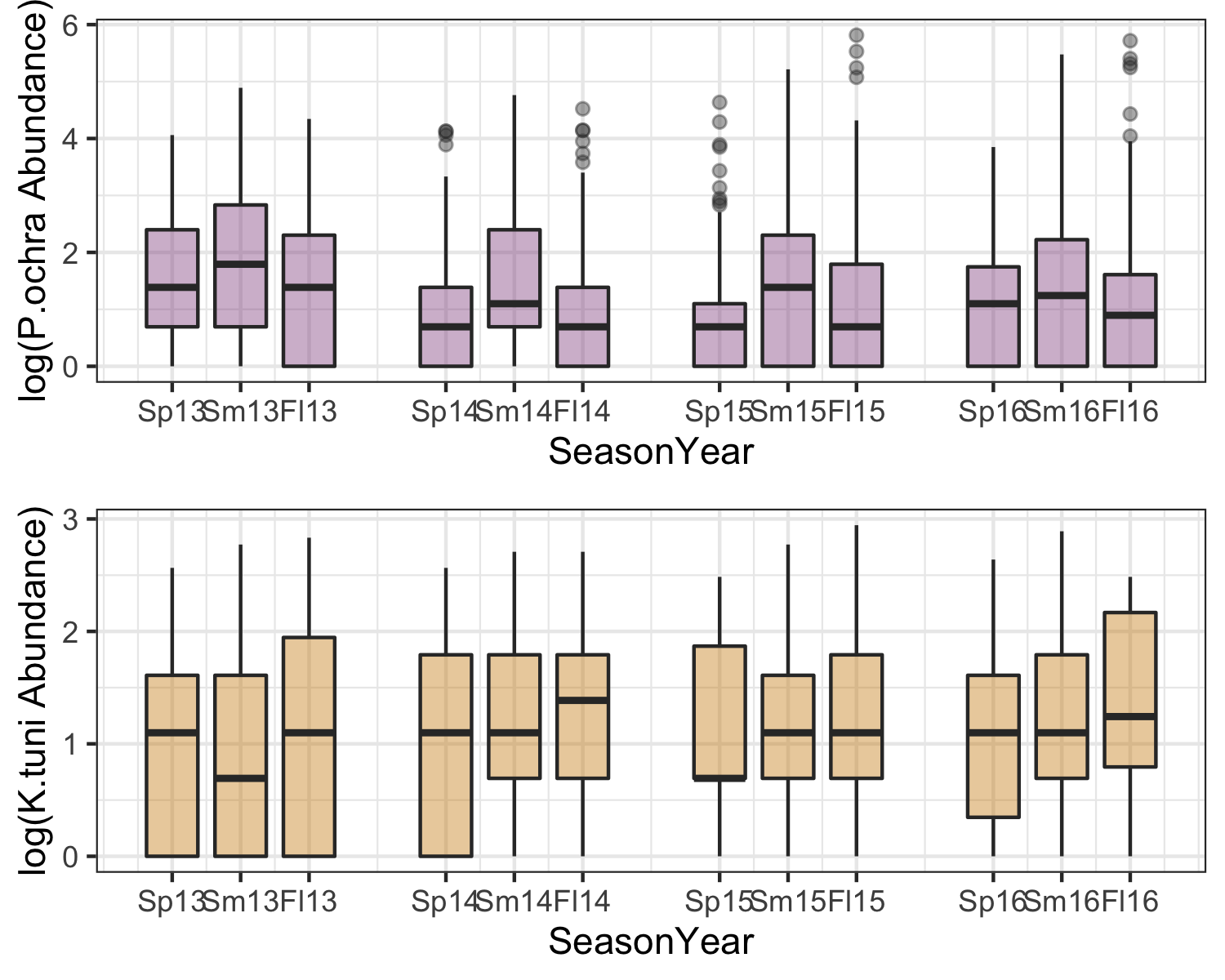
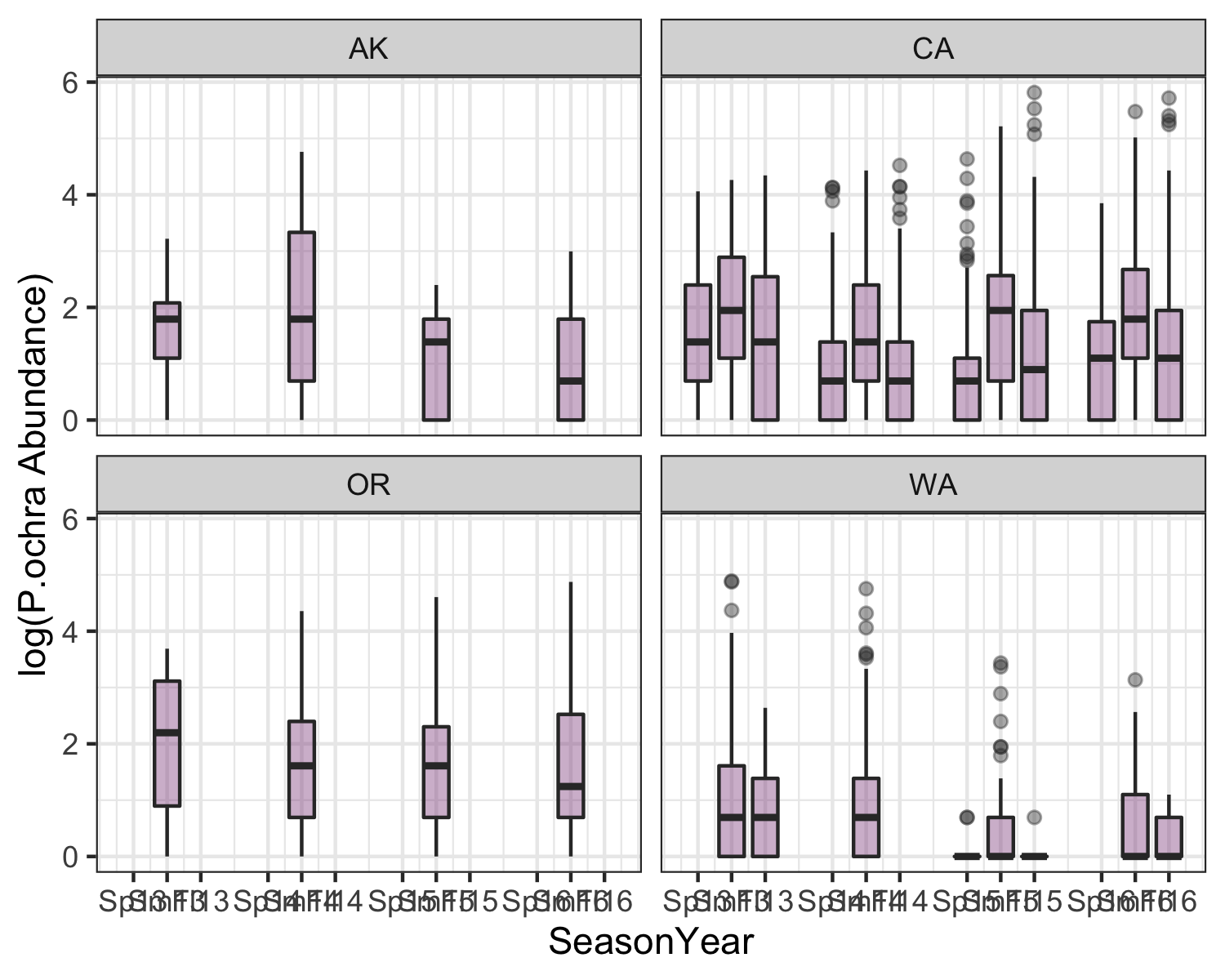
How did sea star abundance look before the epidemic, how did it change during the epidemic, and have the populations rebounded?

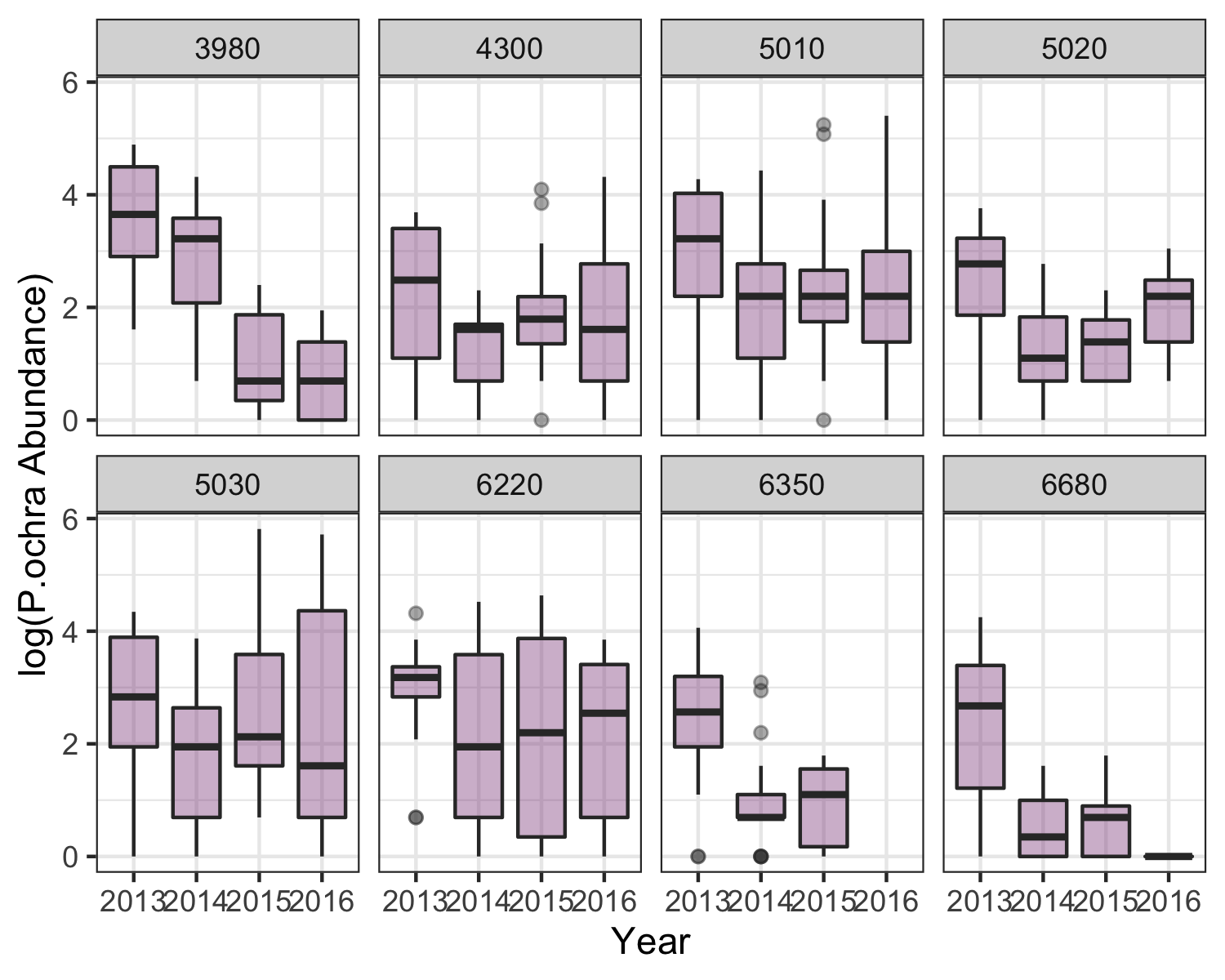
Here’s what abundance looked like across all species: 

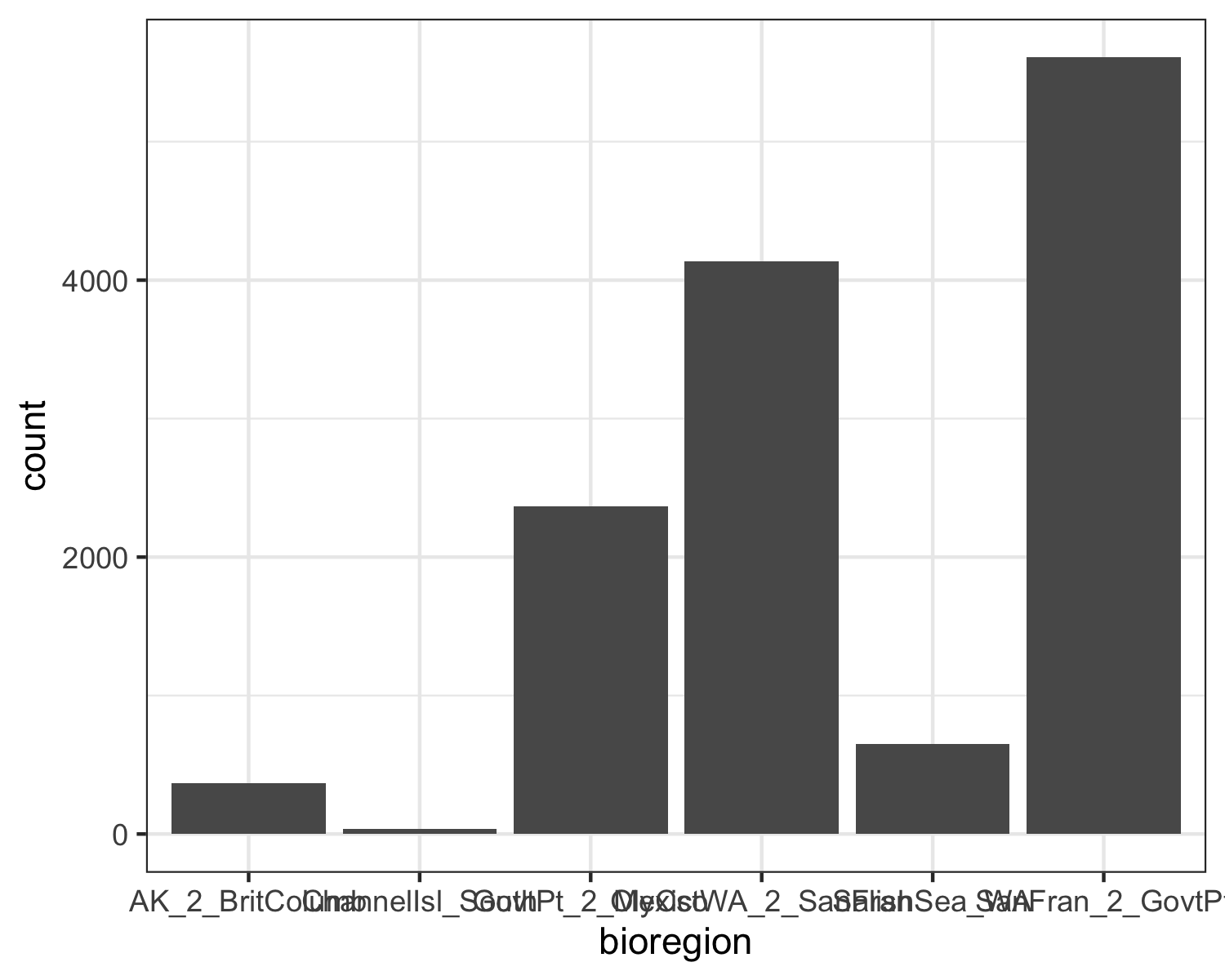
And by species: P.ochraceus:  E.troschelii: 

K.tunicata: 

Then, I focused in on my species of interest, *P. ochraceus*, and expanded the previous graph by breaking down each year by sampling season.

In the figure, below, I used *K.tunicata* as somewhat of a negative control, as it is not a sea star and therefore not affected by sea star wasting disease.  Then I observed P.ochraceus during this time by state: 

Finally, I wanted to look at this same information, but at the site-level scale. There were way too many sites to view them all, so I resolved to look only at sites that originally had over 40 *P.ochraceus* individuals before SSWD “hit”, or sometime in 2013. 

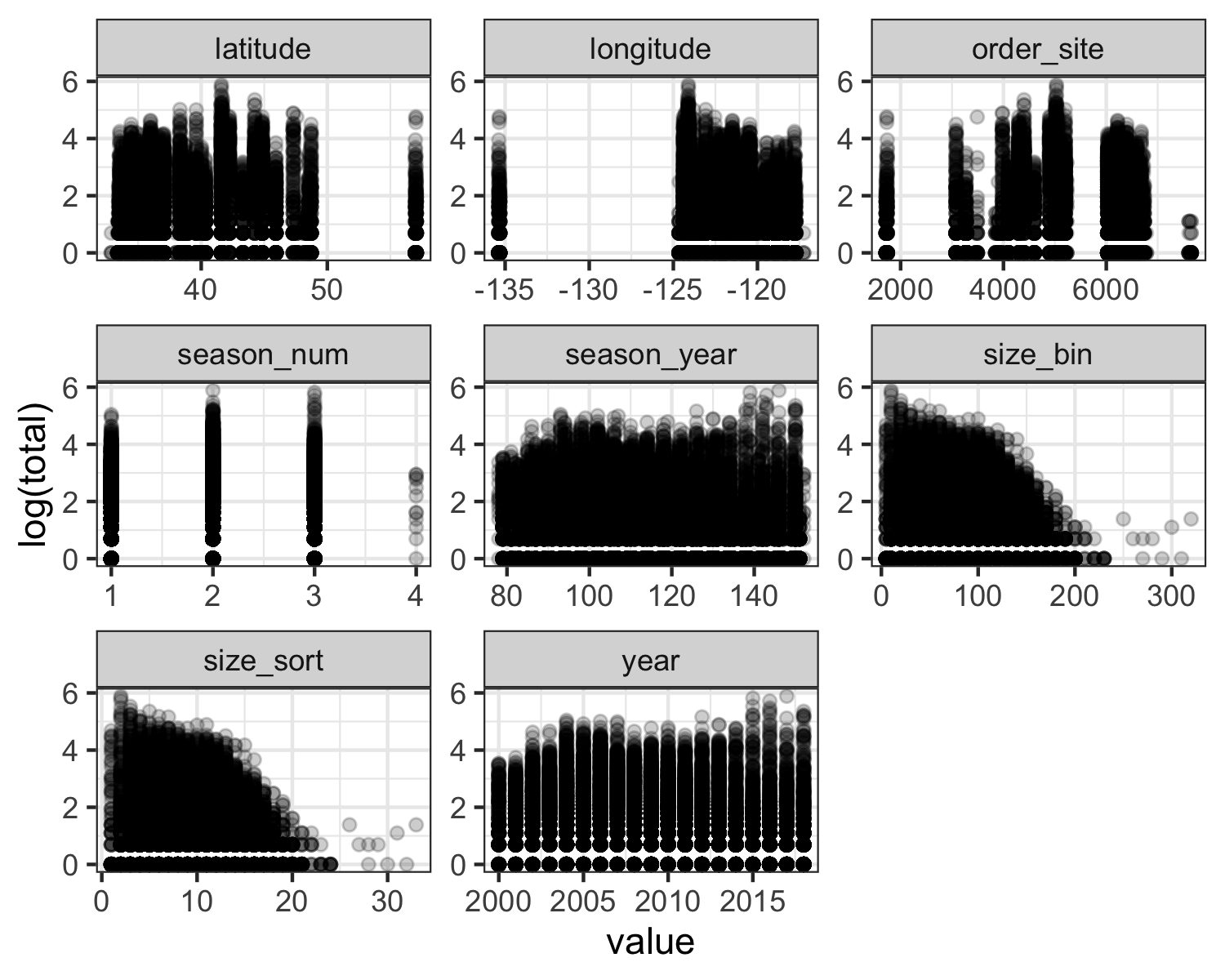


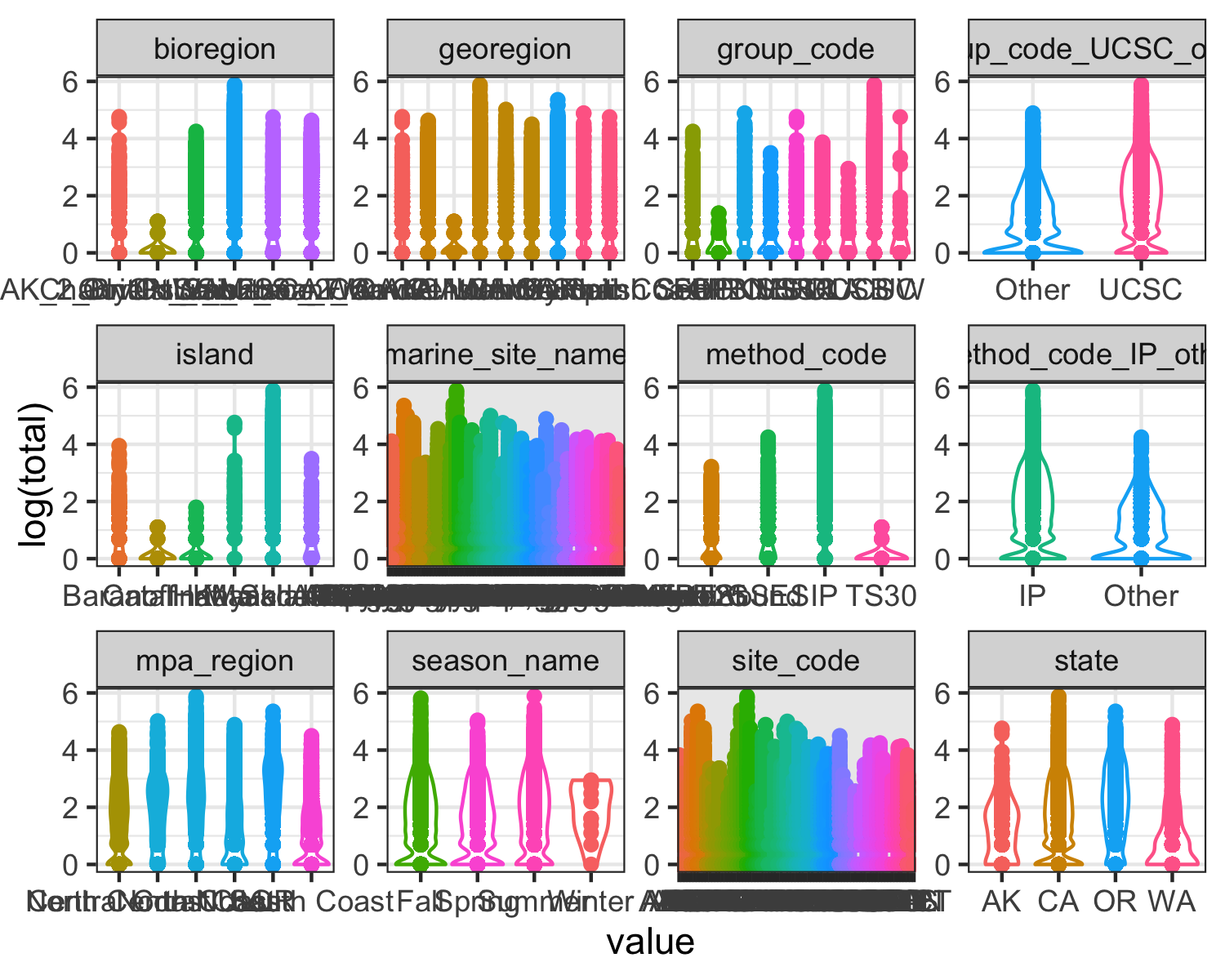
### Bivariate analysis

*Create plots or tables and compute simple statistics (e.g. t-tests, simple regression model with 1 predictor, etc.) to look for associations between your outcome(s) and each individual predictor variable*

The code which performed my bivariate analyses were split between two .Rmd files: ./code/analysis\_code/Exploratory\_Analysis.Rmd and ./code/analysis\_code/Continuous\_Outcome\_Modeling.Rmd.

[FINISH WRITING THIS SECTION]

(partial list of) Figures for this section: 



### Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You can then load the results produced by this code*

I conducted two different full analyses on my data.

The first used abundance counts (or the variable total) as the outcome predictor on a reduced dataset that only retained the species of interest, *Pisaster ochraceus*. This code is found here: ./code/analysis\_code/Continuous\_Outcome\_Modeling.Rmd.

The second focused on the categorical variable bioregion and used a machine learning tree fitting method. This code is found here: ./code/analysis\_code/Continuous\_Outcome\_Modeling.Rmd.

[FINISH WRITING THIS SECTION]

## Discussion

### Summary and Interpretation

*Summarize what you did, what you found and what it means.*

### Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

### Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

reference (Leek & Peng, 2015) here reference (Bull & Lauring, 2014) here

### References

Bull, J. J., & Lauring, A. S. (2014). Theory and empiricism in virulence evolution. *PLoS Pathogens*, *10*(10), e1004387.

Leek, J. T., & Peng, R. D. (2015). Statistics. What is the question? *Science (New York, N.Y.)*, *347*(6228), 1314–1315. <https://doi.org/10.1126/science.aaa6146>