Predicting Community Assemblage in The Rocky Intertidal of Acadia National Park

Exploratory Data Analysis Report

DSE6311 – SU2 2025

Nick Lagoni

7/29/2025

1 Introduction

1.1 Research Question

The primary research question posed by this analysis is: can we accurately predict how the percent cover of ecosystem engineers *Ascophyllum nodosum* and *Fucus vesiculosus* of the rocky intertidal in Acadia National Park will change in the near future, and what variables (including substrate composition and invertebrate abundance) are the strongest predictors of said percent cover? While long-term monitoring has been ongoing through the Northeast Temperate Network (NETN) in the Inventory and Monitoring Division of the National Park Service (NPS) since 2013, no significant analysis has been done on this subject since that protocol began.

I hypothesize that presence and abundance of motile invertebrates will significantly predict the substrate percent cover of *Fucus vesiculosus* (Linnaeus 1753) and *Ascophyllum nodosum* (Linnaeus 1753) in the Acadia rocky intertidal over time.

I predict that an increase in the abundance of motile invertebrates *Littorina obtusata* will significantly be associated with a significant reduction in the algal cover of *F. vesiculosus* and *A. nodosum*, as these periwinkles have been shown to feed on both of these species (Hadlock 1979, Watson and Norton 1987). I predict that *Littorina littorea* abundance will not significantly predict the algal cover of *F. vesiculosus* and *A. nodosum*, as this species has been shown to avoid feeding on these species (Watson and Norton 1985)

1.2 Research Question and Hypothesis

The primary research question posed by this analysis is: can we accurately predict how the community assemblage of the rocky intertidal in Acadia National Park will change in the near future, and what variables (including substrate composition, invertebrate abundance, and temperature) are the strongest predictors of that. While long-term monitoring has been ongoing through the Northeast Temperate Network (NETN) in the Inventory and Monitoring Division of the National Park Service (NPS) since 2013, no significant analysis has been done on this subject since that protocol began.

I hypothesize that temperature and presence and abundance of motile invertebrates will significantly predict the substrate percent cover of *Fucus vesiculosus* (Linnaeus 1753) and *Ascophyllum nodosum* (Linnaeus 1753) in the Acadia rocky intertidal over time.

I predict that an increase in the abundance of motile invertebrates *Littorina obtusata* will significantly be associated with a significant reduction in the algal cover of *F. vesiculosus* and *A. nodosum*, as these periwinkles have been shown to feed on both of these species (Hadlock 1979, Watson and Norton 1987). I predict that *Littorina littorea* abundance will not significantly predict the algal cover of *F. vesiculosus* and *A. nodosum*, as this species has been shown to avoid feeding on these species (Watson and Norton 1985). I predict that increased temperatures will be correlated with a decrease in *F. vesiculosus and A. nodosum* in the higher plots, and an increase in both target species in the lower plots (i.e., the red algae plots); these species have been shown to expand their range as waters warm (Jueterbock et al. 2013, Marbà

et al. 2017), which may extend to their local zonation range as well, expanding into deeper water.

2 Methods

2.1 Data Source

NETN has a dataset spanning 8-years (2013 – 2021) of long-term ecological monitoring data across six sites in Acadia, as well as three sites in the Boston Harbor Islands National Park (Northeast Temperate Network 2021). All nine sites are visited annually for data collection. The dataset is broken down into different sub-protocols, including: motile invertebrate and substrate percent cover photoplots, line intercept vertical zonation transects, tide pool motile invertebrate band transects, barnacle recruitment photoplots, and water temperature logging. As of now, this analysis is focused exclusively on the motile invertebrate and substrate percent cover photoplots. I am working with the NETN Data Manager to procure the other portions of the dataset, which are currently not available for download.

2.2 Data Preprocessing

Prior to exploratory data analysis (EDA), the data needed to be transformed and consolidated into one singular data frame. The data was in long form, with duplicate rows for each plot at each site during each year. To resolve this, I pivoted each of the three data files (the photoplot percent cover data, the motile invertebrate count data, and the motile invertebrate measurement data) to wide form, using the species name as the indicator of where to pull values from for the new columns, and the percent cover, count, or mean measurement as the

values for the new columns. The motile invertebrate dataset contained up to 10 measurements per species, meaning it had more duplicates than the other datasets, and therefore had to be averaged down to one mean measurement. The three datasets were then joined together using a left join, with the photoplot dataset as the "host," as the target variables are contained within it. Each row in each dataset at this point had been reduced to a single plot from a single site at a single date, so they were joined using these three fields to keep each row as one observation.

2.3 Exploratory Data Analysis Methodology

My first step for EDA was to construct a correlation matrix, in order to give me an overview picture of the trends in my data, and how different features may influence each other. This proved more challenging than I anticipated due to my large number of features (44) making it difficult to create a legible correlation plot. I tried creating a correlation heatmap, as the color gradient is easier to visually process than numbers or biplots when shrunk down, though this too was illegible with all 44 features. I settled on calculating correlation values between each pair, then filtering for only those with a value of >0.7 (a strong correlation), before making a heatmap with those. This resulted in nearly zero correlated variables, so I lowered the threshold to 0.5 to get a better understanding of the data, which resulted in a legible and valuable correlation heatmap.

My next step was to take a preliminary glance at my research question, by examining the relationship between the different motile invertebrate species and my two target variables, *F. vesiculosus* and *A. nodosum* percent cover. I created five scatterplots for each, wrapped into two sets of panels, one for each target variable, for easy interpretability and comparison. These

scatterplots included the target variable percent cover on the Y-axis, the specific motile invertebrate count on the X-axis, and a trendline, calculated via a simple linear model method.

I created a summary table for my continuous variables showing central tendency (mean, median), spread (SD, min, max), and shape (skewness, kurtosis). This table was designed to act as a quick reference for the structure of each of my target and predictor variables. I considered creating a summary table for the categorical variables as well, but the only categorical variables contained in my dataset are labels for each observation, not actual observational data. These variables (i.e., site, plot target variable, and year) may be used in the final analysis but cannot be summarized, as they are, by design, present in equal amounts through the dataset; there are an equal number of plots at each site and each site is visited once per year.

I also conducted a quick principal component analysis (PCA) to look at broad-strokes relationships in the data, to hopefully get a better sense at what variables were responsible for explaining most of the variance in the data. I had hoped to discern which species had strong effects on the structure of the data using this method. I did a preliminary Scree plot to examine the loadings of each principal component to get a sense of the structure of the data.

Lastly, to get another look and a better understanding of my target variables visually, I created a histogram for each, to visualize their frequency distribution and to start thinking about possible data transformations or other considerations for the model.

3 Results

3.1 Correlation Heatmap

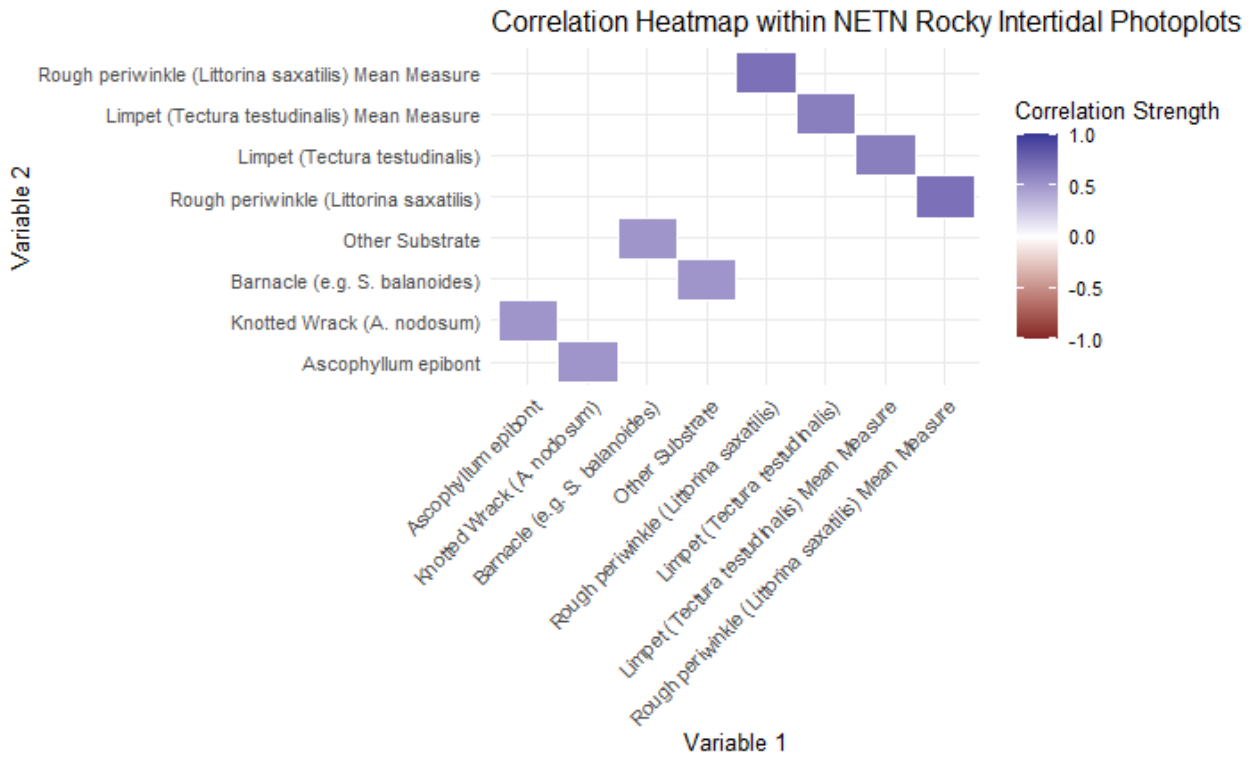

Figure 1 - Correlation Heatmap between moderately correlated variables in the NETN rocky intertidal long term ecological monitoring protocol photoplots. Variables include motile invertebrate count data, motile invertebrate mean length measurement data, macroalgae percent cover data, and substrate percent cover data.

The correlation heatmap showed that there is very little collinearity within this dataset. The strongest collinearity was seen between the motile invertebrate count data and their respective mean length measurement data. This conceptually makes sense as, in plots that are more hospitable to certain species (higher count abundance values) those same species would be more likely to thrive and grow larger. The other correlations were between *A. nodosum* and its epibiont, which requires *A. nodosum* as substrate to grow on, and barnacles and other substrate. Scanning the notes column in the raw data revealed that dead barnacles/leftover shells were frequently labelled as other substrate, which explains that correlation.

3.2 F. Vesiculosus *and* A. nodosum *relationship with Motile Invertebrates*

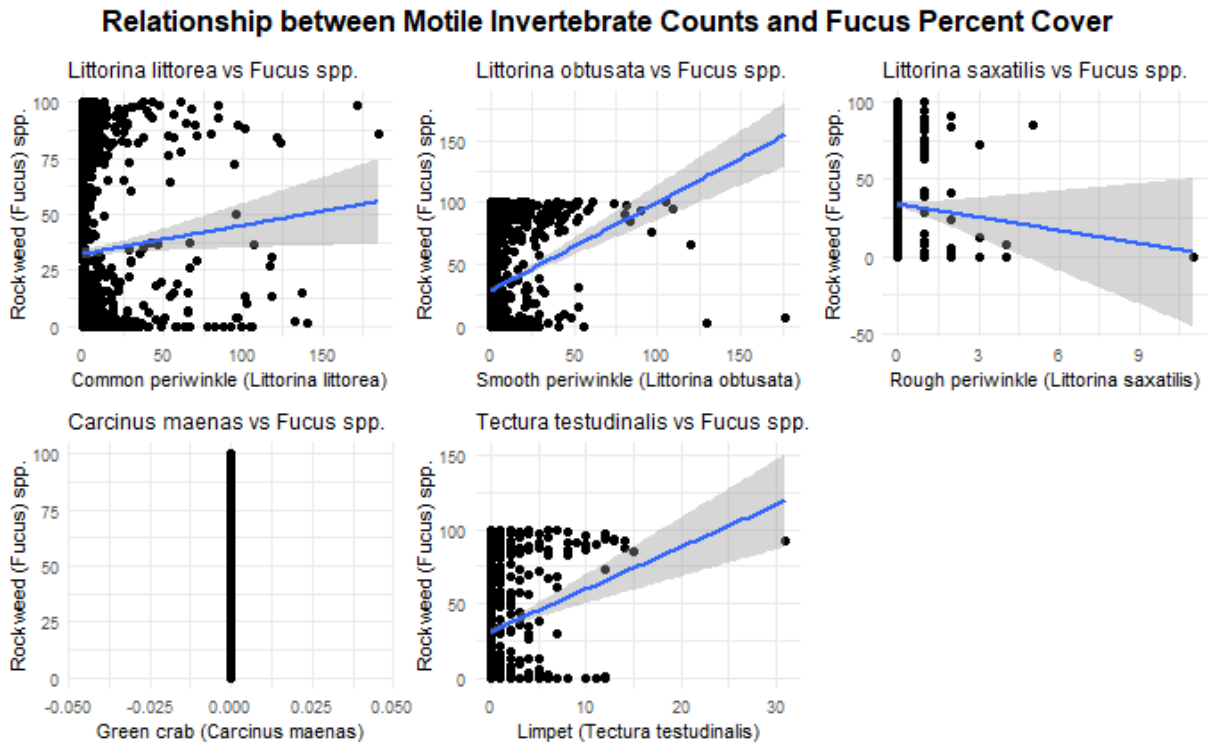

Figure 2 - Relationship between Fucus Vesiculosus percent cover and motile invertebrate abundances in NETN rocky intertidal photoplots.

Examining *F. vesiculosus* percent cover as a function of motile invertebrate abundances suggests that common periwinkles, smooth periwinkles, and limpets have a positive correlation, while rough periwinkles have a negative correlation. This may suggest that either the commons, smooths, and limpets promote *Fucus* in some way, or that they prefer that as habitat.

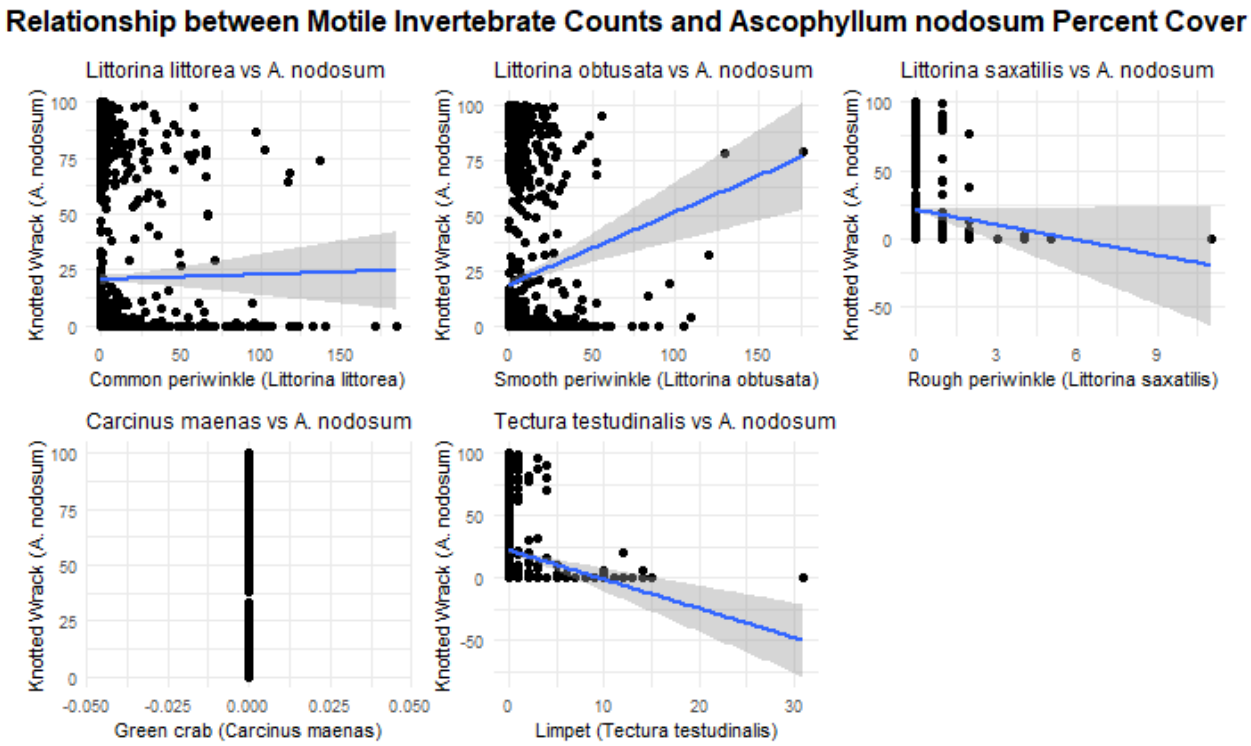

Figure 3 - Relationship between Ascophyllum nodosum percent cover and motile invertebrate abundances in NETN rocky intertidal photoplots.

The plots examining *A. nodosum* percent cover as a function of the various motile invertebrate abundances suggests that common periwinkle abundance has no meaningful relationship with *Ascophyllum* percent cover, while smooth periwinkle abundance is associated with improved percent cover, and both rough periwinkle and limpet abundances are associated with reduced percent cover. The common periwinkle trend is in line with the predictions I made in the proposal, which predicted that there was no relationship between the abundance of the two species as common periwinkles do not preferentially feed on *A. nodosum*.

Both sets of plots suggest that there are some relationships between macroalgal percent cover and motile invertebrate abundance, which is promising for the model. These plots also reveal quite starkly that *Carcinus maenas* only contains values of 0 for abundance, across all

observations. This is an error in the dataset as I personally have collected and recorded *C. maenas* specimens in these plots while assisting NETN, so I will follow up with the Data Manager to see what has occurred here.

3.3 F. Vesiculosus *and* A. nodosum *Frequency Distribution*

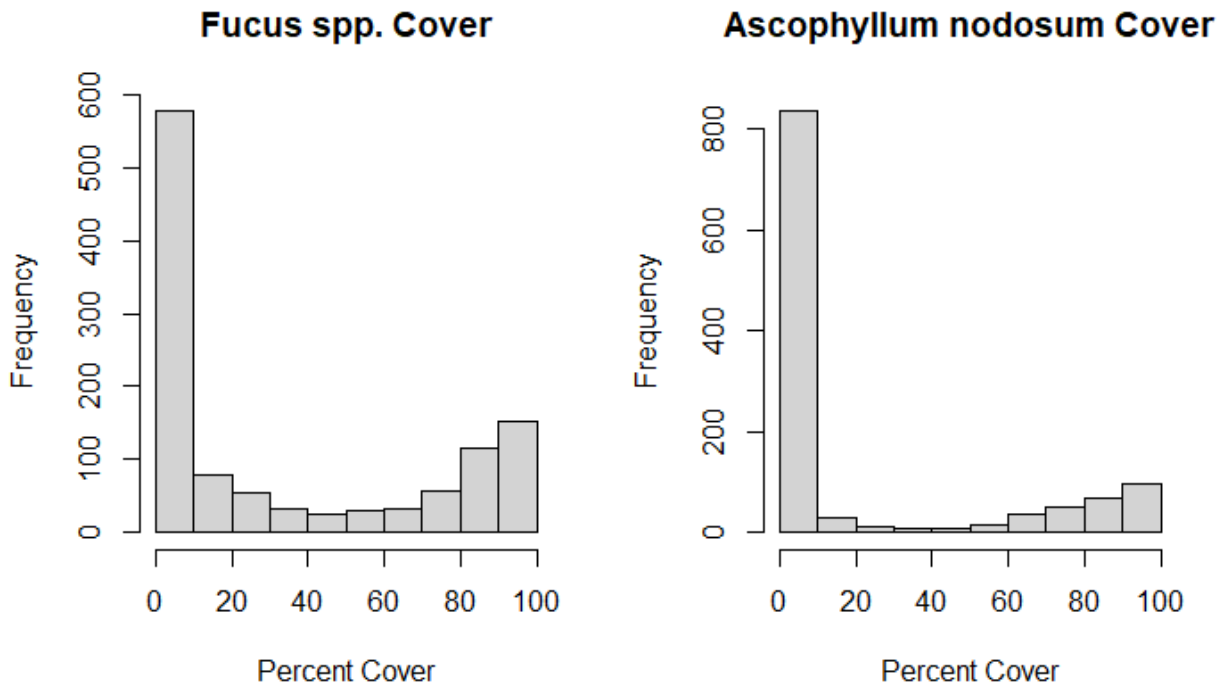

Figure 4 - Frequency distribution histograms for* Fucus *and* Ascophyllum *percent cover values in NETN rocky intertidal photoplots.

Both *Fucus* and *Ascophyllum* have distributions weighted very heavily towards 0. This suggests strong spatial constraints on these species, as they are completely absent from most plots. Their next most common value is 100% cover, which suggests that, when they are present, they are able to do quite well in their niche. This zero-inflated distribution poses difficulties for modeling, as this will violate many common assumptions such as normality.

Table 1 - Summary statistics for numeric variables including percent cover data, invertebrate abundance data, and invertebrate mean measurement data. Summary statistics include mean, median, standard deviation, minimum, maximum, skew, and kurtosis.

Variable	Mean	Median	SD	Min	Max	Skew	Kurtosis
Other Algae - Brown	0.00	0.00	0.11	0	3.00	24.26	616.23
Other Algae - Green	0.55	0.00	2.95	0	47.00	8.62	93.74
Other Algae - Red	1.36	0.00	5.68	0	75.00	6.68	55.30
Articulated Corallines	1.07	0.00	5.27	0	54.00	6.72	49.50
Ascophyllum epibont	2.54	0.00	6.58	0	56.00	3.55	14.34
Knotted Wrack (A. nodosum)	20.66	0.00	35.12	0	100.00	1.31	-0.09
Barnacle (e.g. S. balanoides)	15.81	0.00	29.05	0	100.00	1.75	1.58
Irish Moss (Chondrus / Mastocarpus)	12.01	0.00	25.16	0	100.00	2.01	2.68
Crustose coraline	0.07	0.00	1.21	0	39.00	29.61	939.51
Fucus epibiont	1.09	0.00	4.66	0	96.00	13.81	253.43
Rockweed (Fucus) spp.	33.20	10.00	38.32	0	100.00	0.64	-1.33
Kelp (e.g.Laminaria/Alaria)	0.03	0.00	0.56	0	17.00	25.43	725.61
Mussel (e.g. Mytilus edulis)	1.49	0.00	6.84	0	60.00	5.92	37.64
Crustose non-coraline	1.46	0.00	5.14	0	55.00	5.56	36.61
Not Sampled	0.57	0.00	2.59	0	54.00	14.41	255.32
Other Invertebrate	0.17	0.00	0.59	0	6.00	4.54	25.29
Other Plant	0.00	0.00	0.00	0	0.00	NaN	NaN
Other Substrate	0.53	0.00	1.49	0	17.00	4.28	24.73
Dulse (Palmaria palmata)	1.33	0.00	5.32	0	51.00	5.56	35.40
Laver (Porphyra spp.)	0.61	0.00	3.11	0	46.33	8.96	99.18
Rock	3.44	1.00	7.50	0	58.00	3.82	17.44
Sand	0.01	0.00	0.26	0	9.00	33.93	1150.01
Tar	0.00	0.00	0.00	0	0.00	NaN	NaN
Grass kelp (Ulva intestinalis)	0.01	0.00	0.12	0	3.00	17.71	359.18
Sea Lettuce (Ulva lactuca)	1.90	0.00	5.57	0	56.00	4.54	25.46
Unidentified	0.06	0.00	0.34	0	5.00	7.66	76.79
Green crab (Carcinus maenas)	0.00	0.00	0.00	0	0.00	NaN	NaN
Asian shore crab (H. sanguineus)	0.00	0.00	0.00	0	0.00	NaN	NaN
Common periwinkle (Littorina littorea)	8.26	1.00	20.62	0	185.00	4.08	19.62
Smooth periwinkle (Littorina obtusata)	6.34	1.00	14.13	0	177.00	4.92	35.90
Rough periwinkle (Littorina saxatilis)	0.09	0.00	0.50	0	11.00	11.68	205.17
Dogwhelk (Nucella lapillus)	3.05	0.00	8.50	0	117.00	6.11	52.88
Limpet (Tectura testudinalis)	0.71	0.00	2.12	0	31.00	5.54	47.74
Smooth periwinkle (Littorina obtusata) Mean Measure	5.23	7.75	4.90	0	21.00	-0.03	-1.74
Common periwinkle (Littorina littorea) Mean Measure	8.76	9.50	8.74	0	29.00	0.19	-1.61
Dogwhelk (Nucella lapillus) Mean Measure	8.45	0.00	10.72	0	42.00	0.70	-1.11
Limpet (Tectura testudinalis) Mean Measure	2.16	0.00	4.27	0	17.00	1.58	0.82
Rough periwinkle (Littorina saxatilis) Mean Measure	0.45	0.00	1.93	0	16.00	4.36	18.83
Green crab (Carcinus maenas) Mean Measure	0.00	0.00	0.00	0	0.00	NaN	NaN

Some immediately noteworthy trends in the summary statistics table are the high frequency of zeroes (many variables have a median value of zero), which suggests a very zero heavy dataset. Some variables have no data at all, which is either an error (i.e., in the case of *C. maenas* as described above), or suggests total absence from the plots. The target variables of *F. vesiculosus* (or *spp.*) and *A. nodosum* have moderately high average cover, but have high standard deviation, suggesting high variability in percent cover. They are both relatively present but also have spikes of high concentration in certain plots.

3.5 Principal Component Analysis Scree Plot

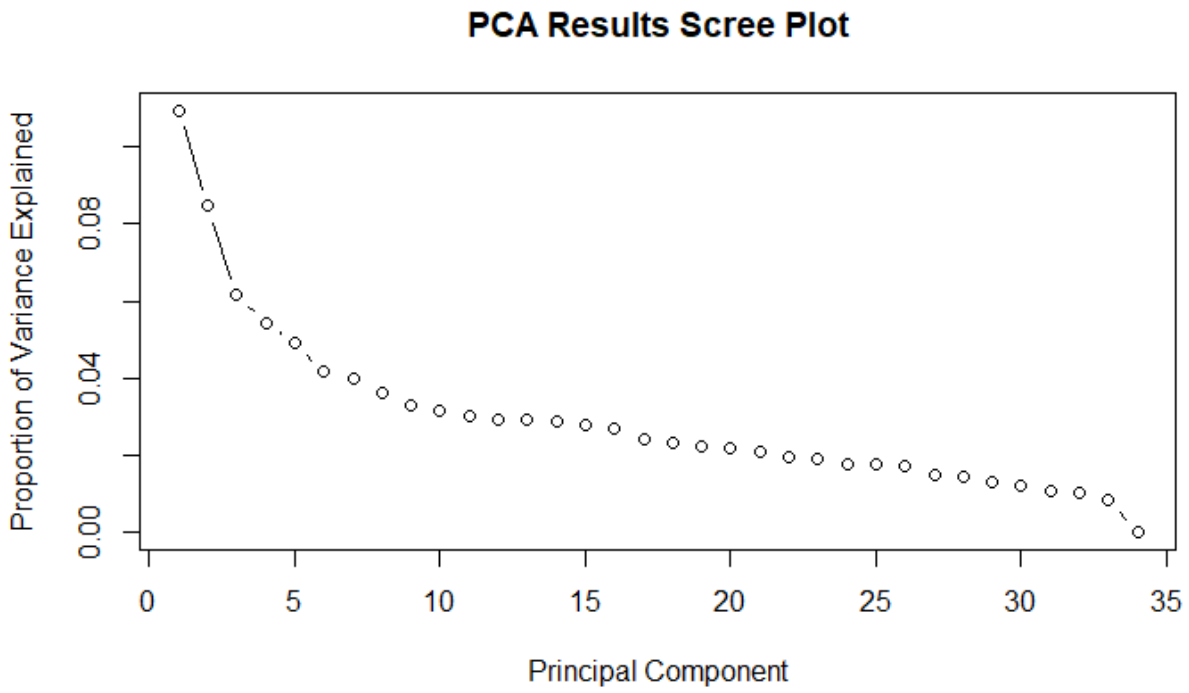

Figure 5 - Principal Component Analysis proportion of variance explained by each principal component. Variance explained per PC is low, drops off precipitously between the first three PCs, and has a consistent but slow decline afterwards.

The scree plot shows that each individual principal component from the PCA is capturing very little of the variability in the data. PC1 accounts for barely 10% of the variability, and it takes a

large number of PCs to explain the majority of the variability. This suggests that the data is weakly structured and high dimensional, the latter of which we know to be true. This also may be a result of the large number of zeroes throughout every variable, which may make the variance difficult to encapsulate here.

4 Discussion

4.1 Conclusions

Table 1, Figure 4, and Figure 5 strongly suggest that the copious number of zeroes in both the target and predictor variables are skewing my ability to interpret the dataset and understand what drives the variability. These exploratory plots and table suggest that, in its current state, the data will be difficult to model with and may violate many key assumptions. The distributions of *F. vesiculosus* and *A. nodosum* described in Figure 4 and Table 1 are congruent with what I have anecdotally experienced in the field—both macroalgae species are present throughout the intertidal sporadically, with additional specific areas that they thrive and overwhelmingly dominate. Figures 2 and 3 suggest that there are some relationships between motile invertebrate abundance and macroalgae percent cover of the target variables, which is promising for the research question, though the actual scatterplots look far from conclusive, despite the trendlines suggesting those relationships. Figure 1 suggests that there is very little multicollinearity in the data, which is both welcome and unexpected, as I would have expected many of the percent cover and abundance variables to be correlated.

4.2 Next Steps

The primary issue facing this analysis, as suggested by Figure 4, Table 1, and Figure 5, is the overwhelming skew towards zero across many of the features of this dataset. Some of these may be errors and will be resolved by further communicating with the NETN Data Manager, but it is likely that I will need to transform the data in some way or select a model that can handle zero inflation well. I believe a log transformation will help with the zero inflation issue, and I believe that Random Forest or xgBoost should perform well even with this nonlinear distribution, though the model may still lean too heavily towards predicting zero. If I find that the model is struggling, I may transform my data altogether to a presence vs. absence question, rather than a continuous percent cover variable. This will diminish the value of the conclusion of the analysis in my mind, but may be a necessary pivot if all else fails.

References

Hadlock, R. H. 1979. The distribution of Littorina obtusata (L.) in the rocky intertidal: effects of competition with Littorina littorea (L.). Master's thesis, Department of Zoology, University of Rhode Island, Kingston, RI.

Jueterbock, A., L. Tyberghein, H. Verbruggen, J. A. Coyer, J. L. Olsen, and G. Hoarau. 2013. Climate change impact on seaweed meadow distribution in the North Atlantic rocky intertidal. Ecology and evolution 3:1356–1373.

Linnaeus, C. 1753. Species plantarum, exhibentes plantas rite cognitas ad genera relatas cum differentiis specificis, nominibus trivialibus, synonymis selectis, locis natalibus, secundum systema sexuale digestas. Vol 1.

Marbà, N., D. Krause-Jensen, B. Olesen, P. B. Christensen, A. Merzouk, J. Rodrigues, S. Wegeberg, and R. T. Wilce. 2017. Climate change stimulates the growth of the intertidal macroalgae Ascophyllum nodosum near the northern distribution limit. Ambio 46:119–131.

Northeast Temperate Network. 2021, January 1. Long-term Rocky Intertidal - Database. https://irma.nps.gov/DataStore/Reference/Profile/2289832.

Watson, D. C., and T. A. Norton. 1985. Dietary preferences of the common periwinkle, Littorinalittorea (L.). Journal of Experimental Marine Biology and Ecology 88:193–211.

Watson, D. C., and T. A. Norton. 1987. The habitat and feeding preferences of Littorina obtusata (L.) and L. mariae sacchi et rastelli. Journal of Experimental Marine Biology and Ecology 112:61–72.