Predicting Community Assemblage in The Rocky Intertidal of Acadia National Park

Data Preprocessing and Feature Engineering Report

DSE6311 – SU2 2025
Nick Lagoni
8/5/2025

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).

2 Methods

2.1 Preprocessing Plan

Fortunately, the data provided by the National Park Service has already been cleaned up and validated with their quality assurance and quality control protocols, but I still anticipated doing some preprocessing work. The primary concern I faced with this dataset is the zero skewed distribution, which is a result of a large quantity of zeroes for different species abundance and percent cover values. To attempt to prepare the data for use with typical regression models that require a normal distribution, my plan was to use progressively more and more intensive data transformations and to evaluate the distribution of my data at each step, starting with a simple square root transformation, and transitioning to a binary presence absence transformation.

For feature engineering, my main plan is to cull any columns that are entirely zero, as these represent no meaningful data. I also plan to utilize Non-metric Multidimensional Scaling (NMDS) and Principal Component Analysis (PCA, with a Hellinger transformation) to attempt to reduce the dimensionality of the data, though these would reduce the interpretability of any modeling.

2.2 Data Preprocessing

My first step was to consolidate the data 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 per plot sample. 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 one sampling effort for 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. Since some of the mean measurement values were nonexistent (for plots with no respective motile invertebrates), zeroes were imputed to replace missing values during the joining process.

My next step was to begin the escalating, iterative data transformation process. I applied a square root transformation to the numeric columns of the final joined dataset and visually evaluated the distribution of the target variables using a histogram plot for both *Ascophyllum* and *Fucus* percent cover. I repeated this process for a log(x+1) transformation and, finally, a presence-absence transformation. I selected square root as a natural, simple, starting off point for the process, used log(x+1) as a more substantial transformation, and used presence-absence as the last transformation as it represents the most dramatic change I can feasibly make to the data. This gave me a snapshot of the spectrum of transformations available and what I could expect in terms of resultant distributions.

2.3 Feature Engineering

The first unsupervised feature engineering process I explored was Non-Metric Multidimensional Scaling (NMDS) with a Bray-Curtis dissimilarity index, as it is generally considered to be a good choice for ecological community ordination (Minchin and Rennie 2010).

I then performed a Hellinger transformation on the joined dataset, which first transforms the data into relative species abundance, then takes the square root of it, diminishing the effects of disproportionately large count values (Quebec Centre for Biodiversity Science 2023). I opted to perform a Hellinger transformation in order to then perform a PCA with the transformed data, as this has been shown to improve the efficacy of PCA when working with ecological community data (Minchin and Rennie 2010).

2.4 Training – Test Split

I elected to use the $\sqrt{p}: 1$ method for determining the optimal training-test split for the dataset, as this is generally viewed as the ideal method for determining this split (Joseph 2022). While my dataset has a large number of features (39), it has an appropriately large sample size (1157). For this reason, I believe the theoretically ideal split method is sufficient, as I do not see a need to tailor make a larger test set to account for a small sample size.

3 Results

3.1 Data Preprocessing

The progressive data transformation strategy that I employed yielded mixed to poor results. The square root and log(x + 1) transformations did not improve the distribution, meaning that models that require a normal distribution are not a good option moving forward. The presence-absence transformation did improve the distribution, though it reduces the value of the analysis substantially. Presence and abundance is a less granular metric than abundance, making it inherently less informative.

3.2 Feature Engineering

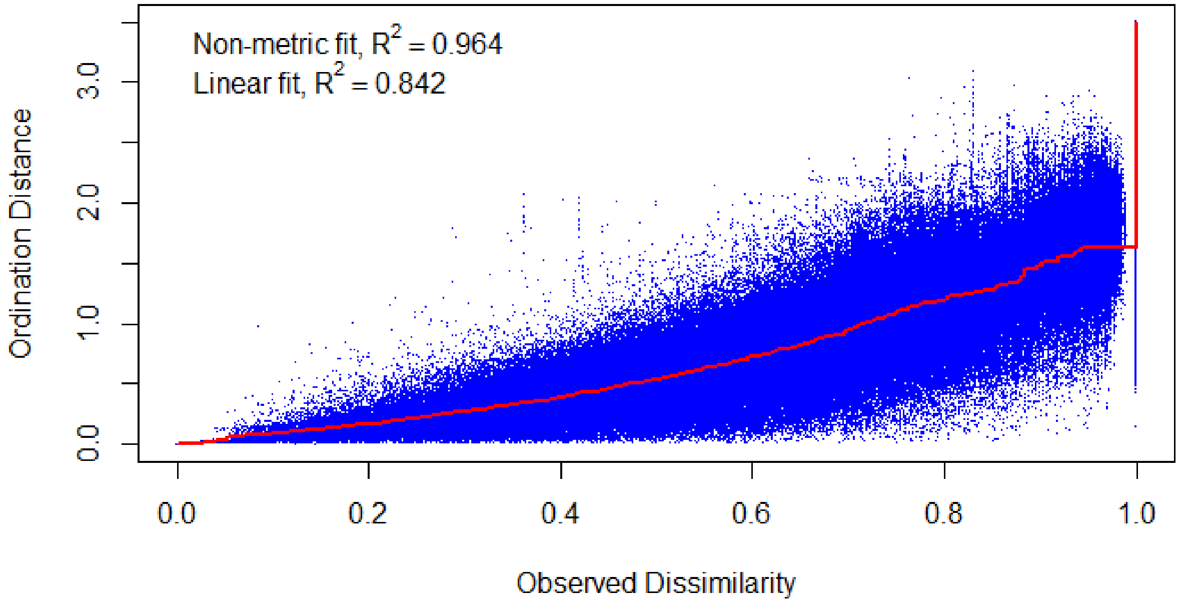

Figure 1 - Non-metric Multidimensional Scaling (NMDS) Plot. The NMDS plot shows little clustering and a high stress (>0.2) despite a relatively strong fit to the data (R^2 = 0.964).

The NMDS plot (Figure 1) revealed no strong clustering, and had an accompanying stress value above 0.2, which is a threshold at which the results are generally considered dubious (Clarke 1993, Dexter et al. 2018). This suggests that the NMDS is not doing an adequate job of capturing the true dimensionality in the data, and that any groupings (which are already not readily apparent from Figure 1) are not trustworthy.

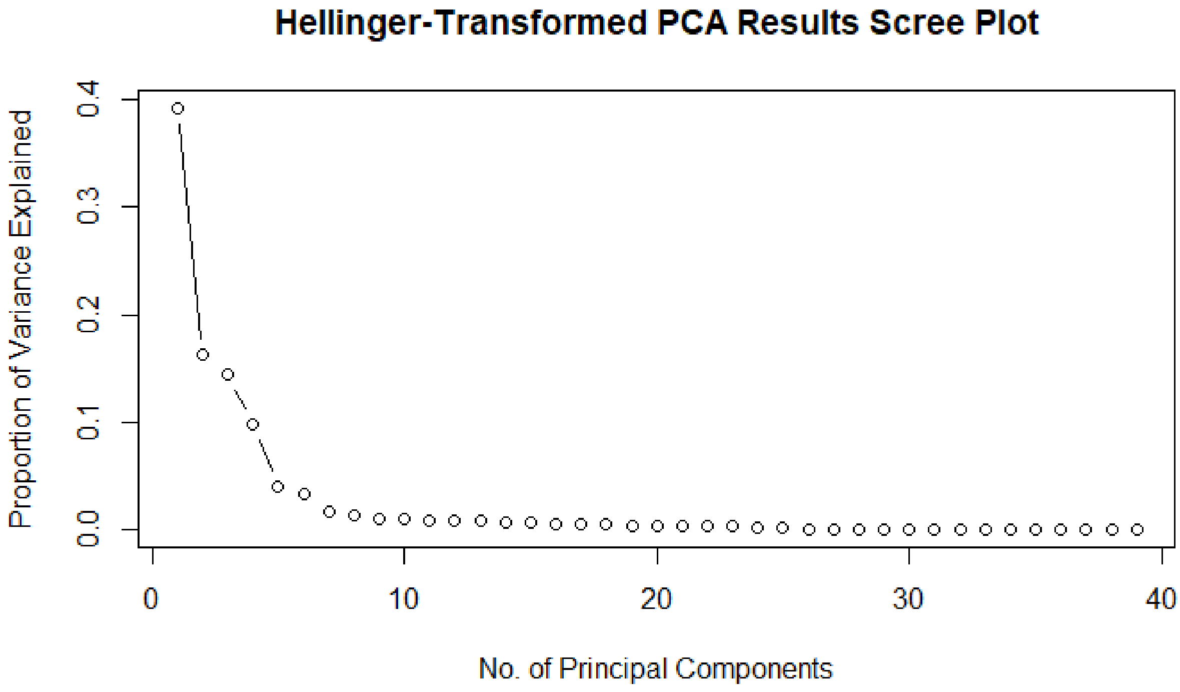

Figure 2 - Scree plot examining the proportion of variance captured by the Principal Component Analysis on Hellinger-transformed data. PC1 captures 39.5% of the data, followed by PC2 and PC3 with a much lower 16.25% and 14.4%, respectively. After PC3, each component contributes a negligible amount to explaining the variation in the data.

The Hellinger-transformed PCA, shown in the Scree plot in Figure 2, had slightly more promising results than the NMDS, and was an improvement over the exploratory PCA I conducted during the Exploratory Data Analysis portion of this project. The Hellinger-transformed PCA did a moderate job at capturing the variability in the data, with the first principal component accounting for 39.25%, the second component accounting for 16.25%, and the third accounting for 14.4%. After the first component the accounted for variability saw a precipitous drop, which continued after the third principal component as well. This is an improvement in performance to the earlier PCA, but still leaves much to be desired in explaining the variability in the data.

4 Discussion

4.1 Conclusions

The overarching conclusion from this preprocessing and feature engineering analysis is that the zero-inflated nature of the data is still dominant. It may not be feasible to resolve this issue with preprocessing, data transformation, or feature engineering. Instead, models specifically tailored to a non-normally distributed dataset, or a zero-inflated dataset, may be a more suitable option for this analysis. Dimensionality reducing techniques like NMDS and PCA showed some limited success, especially in tandem with data preprocessing tools like the Hellinger transformation, but these may alter the interpretability of the data too far to be worthwhile, even if they were highly successful in capturing the variation and dissimilarity in the data. The removal of features will all zeroes in their counts or percent cover data is worthwhile, though the removal of *Carcinas maenas*, the European Green Crab, is still an error caused by incorrect data, in my opinion. I have not yet heard back from the Data Manager regarding the repaired dataset, so for now it will stay removed. I hope to see it return as it is an ecologically significant invasive species, which has drawn a lot of attention from both academics and the public in recent years.

4.2 Next Steps

I took a preliminary stab at the next step of this analysis, which is to explore non-parametric models and other suitable techniques for modeling with a zero-inflated dataset like this. I ran two out-of-box random forest models to get a sense of how that non-parametric model would do, without jeopardizing the testing set by inviting data leakage. This model was fit well to the training set, explaining ~95% of the variance in both *Fucus* and *Ascophyllum* percent cover, but it failed to predict accurately, with a mean error of squared residuals value of 46.6 and 70.7, respectively (on a scale out of 100). I experimented with a logit-transformation on only the percent cover values, as they are bounded (0 – 100), with an epsilon value of 0.0001. This resulted in improved performance compared to the unmodified out of bag random forest model, so I intend to explore this avenue further. I also intend to branch out to other models that may handle the data well, such as gradient boosting. I have also been looking into how I might use a feedforward neural network for this problem, but I am not confident that I have the sample size I need to do so. I am thinking if I reduce the number of features down, either by combining algal species into groups (i.e., red algae, brown algae, green algae), or by trimming down the less informative groups (i.e., other substrate, unidentified substrate, rock, etc.), or even limiting the scope to *just* investigating the impact of motile invertebrates on *Ascophyllum* and *Fucus*, rather than the other substrates and algae as well.

References

Clarke, K. R. 1993. Non-parametric multivariate analyses of changes in community structure. Australian Journal of Ecology 18:117–143.

Dexter, E., G. Rollwagen-Bollens, and S. M. Bollens. 2018. The trouble with stress: A flexible method for the evaluation of nonmetric multidimensional scaling. Limnology and Oceanography: Methods 16:434–443.

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.

Joseph, V. R. 2022. Optimal ratio for data splitting. Statistical Analysis and Data Mining: The ASA Data Science Journal 15:531–538.

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.

Minchin, P., and L. Rennie. 2010. Does the Hellinger transformation make PCA a viable method for community ordination? 95th Annual ESA Meeting Contributed Oral Papers.

Quebec Centre for Biodiversity Science. 2023. Chapter 10 Transformations | Workshop 9: Multivariate Analyses in R.

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.