Predicting Community Assemblage in The Rocky Intertidal of Acadia National Park

Data Preprocessing, Feature Engineering, and Initial Modeling Report

DSE6311 – SU2 2025

Nick Lagoni

8/12/2025

1 Introduction

1.1 Background

Intertidal communities in the Gulf of Maine are experiencing rapid change, including changes in vertical zonation (Trott 2022) and introduction of invasive and range expanding species (Cohen et al. 1995, Yamada 2001, Epifanio 2013, Johnson 2015, Cheng et al. 2025). These changes may compound over time into significant changes to community structure. Because of this and the importance of these ecosystems ecologically and as a source of ecosystem services, including as interpretive spaces in National Parks, it is important that we understand what the future holds for the rocky intertidal in the Gulf of Maine, and which factors contribute most to that future.

1.2 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 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 Data Preprocessing

The target variables, *F. vesiculosus* and *A. nodosum* percent cover, and the percent cover predictor variables, are percentage data, bounded between 0 and 100. While Random Forest is a non-parametric model that does not require normally distributed or unbounded data, the bounded nature of percentage data can still cause issues for the model. To address these issues and better represent proportional differences across the full range of possible values, a logit transformation was applied to all percentage variables. The logit transformation remaps the bounded interval (a proportion from 0 to 1) to the unbounded real line (negative infinity to positive infinity). An epsilon value of 0.0001 was added to all proportions before transformation to prevent undefined values at exactly 0 or 1.

2.2 Feature Engineering

Previously, I explored unsupervised feature engineering processes with limited success in reducing dimensionality of the data. I performed a non-parametric multidimensional scaling analysis using Bray-Curtis dissimilarity, and a principal component analysis using untransformed and Hellinger-transformed data. NMDS produced no distinct clustering, though it is generally regarded as effective for clustering ecological community data (Clarke 1993). Principal components did not sufficiently capture the variability in the data. Additionally, PCA, even with Hellinger-transformed data, has been shown to be ineffective when working with ecological community data (Minchin and Rennie 2010). Because of this, and to maintain interpretability of the final model, unsupervised feature engineering is not explored any further.

An initial out-of-bag Random Forest model was used to examine the importance of each predictor in predicting *F. vesiculosus* and *A. nodosum* percent cover. Any feature with a percent increase in mean squared error of 0 or below for either target species was removed. Percent increase in mean squared error in this instance is a measure of how much the mean squared error increases when an individual variable is imputed randomly. This means that values of zero represent variables that are not significantly contributing to the model's performance, and, in the case of negative values, are actively adding noise. Additionally, some important variables were removed for the sake of addressing the specific research question posed.

2.3 Training – Test Split

The equation $\sqrt{p}$: 1 is used to calculate the optimal training-test split for the dataset, as this is generally viewed as the ideal method for determining this split (Joseph 2022). The dataset in

use has a sufficiently large sample size to parameter (*n*:*p*) ratio, so the theoretical optimal split is sufficient and there is no need to specifically tailor the split to accommodate the structure of the data.

2.4 Initial Model

The initial model selected for this analysis is a Random Forest model as it is a non-parametric machine learning method that handles mixed data types, the skewed distribution, and non-linear relationships between predictors that has been examined in the NETN monitoring data and is oftentimes found in ecological data. Additionally, Random Forest handles multicollinearity well which has been observed in the NETN data to a small extent. Out-of-bag (OOB) error is used as a first estimate of test error. An OOB model is computationally cheap, fast, and helps mitigate data leakage while still having comparable performance to a fully fleshed out, cross-validated model.

3 Results

3.1 Data Preprocessing

The logit transformation with an epsilon of 0.0001 was fairly successful in reducing the skew of the data, with the untransformed *Ascophyllum* and *Fucus* having skew values of 1.282 and 0.672, respectively, and the transformed *Ascophyllum* and *Fucus* having skew values of 0.767 and 0.067, respectively. This represents a more symmetrical distribution post-transformation which should improve model performance.

3.2 Feature Engineering

Location name, target species, and plot name were manually removed despite having high scores from the OOB Random Forest importance metrics. Location name was first filtered to only include observations from Acadia National Park, then was removed as a column, as this analysis is only focused on Acadia, not the Boston Harbor Islands. Target species, the value referring to the original target species of the plot (i.e., *Ascophyllum* plots, red algae plots, etc.) was removed, as separating the site into areas that are historically associated with a specific species is not a helpful lens for this analysis. Plot name was removed, as they are arbitrary labels for the different plots (i.e., R5 for the fifth red algae plot on any given site) and do not contain any meaningful information. There is no meaningful comparison between an R5 plot on one site and an R5 plot on another, aside from the fact that they are both in the red algae area.

3.3 Training – Test Split

The optimal split is an 85:15 training to testing split. This is in the same realm as the general use 80:20 or 70:30 splits.

3.4 Initial Modeling

The initial OOB Random Forest models with the logit-transformation performed well. When predicting *A. nodosum* percent cover, the model had an R^2 value of 90.35, which suggests a strong fit to the data, and a mean squared error of 2.135. When predicting *F. vesiculosus* percent cover, the model had an R^2 value of 82.2, which again suggests a strong fit to the data, and a mean squared error of 4.505. The error values are more difficult to evaluate as they are

output from the model in logit space, which is not directly comparable to their original raw values. Back-converting from the logit transformation is not one to one either, and varies depending on the original value. Figure 1 examines the change in back-converted error value compared to true percent cover value, which shows a slight increase in the error rate in the middle of the range for both *Fucus* and *Ascophyllum*, with relatively low error at either end of the spectrum, as well as the highest density of data points at either end. This suggests that the model performs best when predicting at the extreme ends of the spectrum, which is also the most common observed case for both target algal species.

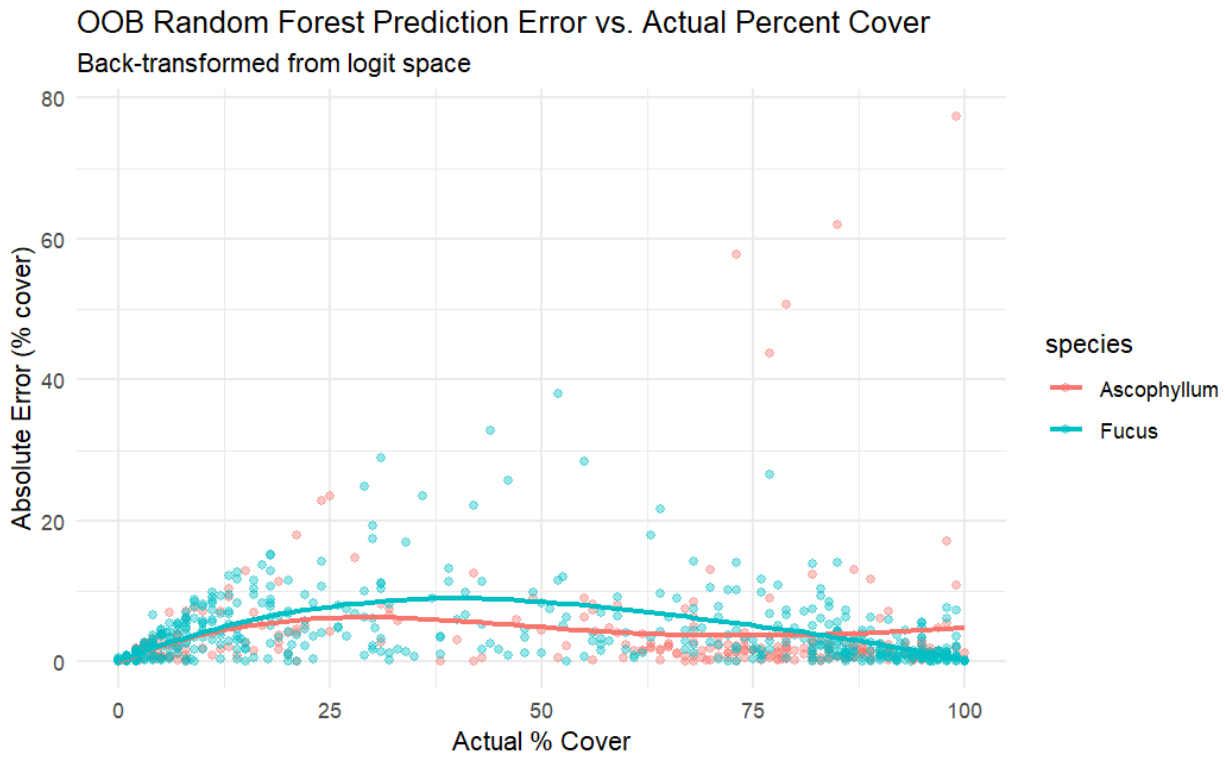

Figure 1 - Out of Bag Random Forest model prediction error compared to actual percent cover value of *Fucus vesiculosus* and *Ascophyllum nodosum* in the rocky intertidal of Acadia National Park. Most data points are clustered around the extremes (0 – 25 percent cover or 75 – 100 percent cover). The mean absolute error rate is consistently low, below 10% cover, with a slight increase in the middle of the range (between 25 – 50 percent cover) in both target species. This suggests that the model is best at predicting values in the extreme ranges of percent cover, which is also where the two target species tend to fall most frequently; the model is best at predicting the most common cases.

4 Discussion

4.1 Conclusions

The initial test run of the model as an out-of-bag Random Forest shows promising results, with a strong fit to the data and low error rates when predicting both *Fucus* and *Ascophyllum* percent cover. While the Random Forest model can handle a high number of predictor variables, the features have been successfully pared down to a more manageable amount, removing unimportant variables manually and by evaluating importance via Random Forest. The model is likely not yet performing optimally but does serve as a proof of concept that the data does support a predictive model for *Fucus* and *Ascophyllum* percent cover.

4.2 Next Steps

The most critical next step is to adapt the model that has been developed so far into one that better addresses the research question. What I have realized late in the initial modeling phase is that, while the model performs well and is able to predict algal percent cover within a year, the output of this particular does not actually predict the future year's percent cover values. If we could collect all the other data for future years and input it into this model, we could reasonably predict the target species percent cover values, but that is not feasible nor valuable for the specific need at hand. Instead, I will explore creating a time-lagged model that uses previous years' values as predictors to predict the current year's target species percent cover. If that model is successful, then that will more sufficiently address the research question posed. I will also look to use cross validation to tweak the Random Forest hyperparameters like the

number of trees and the number of variables in each tree, as right now I have arbitrarily selected 500 trees and 10 variables per tree as a starting point that I saw as sufficiently complex to get a sense for model performance.

5 Appendix A: Data Dictionary

5.1 NETN Rocky Intertidal Long Term Monitoring Protocol Data Dictionary

VARIABLE NAME	DESCRIPTION	UNITS / SCALE	TRANSFORMATION	TYPE
LOC_NAME	Site name within Acadia National Park where survey was conducted (6 total sites)	Categorical (string)	None	Identifier
START_DATE	Date of annual survey for that site/year	Date (MM/DD/YYYY)	None	Identifier
ROCKWEED_FUCUS_SPP	Percent cover of rockweed (*Fucus* spp.)	Percent cover (%)	Logit transformed	Numeric
KNOTTED_WRACK_A_NODOSUM	Percent cover of knotted wrack (*Ascophyllum nodosum*)	Percent cover (%)	Logit transformed	Numeric
BARNACLE_E_G_S_BALANOIDES	Percent cover of barnacles (*e.g.*, *Semibalanus balanoides*)	Percent cover (%)	Logit transformed	Numeric
MUSSEL_E_G_MYTILUS_EDULIS	Percent cover of mussels (*e.g.*, *Mytilus edulis*)	Percent cover (%)	Logit transformed	Numeric
IRISH_MOSS_CHONDRUS_MASTOCARPUS	Percent cover of Irish moss (*Chondrus crispus*, *Mastocarpus stellatus*)	Percent cover (%)	Logit transformed	Numeric
KELP_E_G_LAMINARIA_ALARIA	Percent cover of kelp (*e.g.*, *Laminaria* spp., *Alaria* spp.)	Percent cover (%)	Logit transformed	Numeric
DULSE_PALMARIA_PALMATA	Percent cover of dulse (*Palmaria palmata*)	Percent cover (%)	Logit transformed	Numeric
LAVER_PORPHYRA_SPP	Percent cover of laver (*Porphyra* spp.)	Percent cover (%)	Logit transformed	Numeric
SEA_LETTUCE_ULVA_LACTUCA	Percent cover of sea lettuce (*Ulva lactuca*)	Percent cover (%)	Logit transformed	Numeric
ARTICULATED_CORALLINES	Percent cover of articulated coralline algae	Percent cover (%)	Logit transformed	Numeric
CRUSTOSE_NON_CORALINE	Percent cover of crustose non-coralline algae	Percent cover (%)	Logit transformed	Numeric
OTHER_ALGAE_GREEN	Percent cover of other green algae species	Percent cover (%)	Logit transformed	Numeric
OTHER_ALGAE_RED	Percent cover of other red algae species	Percent cover (%)	Logit transformed	Numeric
FUCUS_EPIBIONT	Percent cover of epibionts on *Fucus* spp.	Percent cover (%)	Logit transformed	Numeric
ASCOPHYLLUM_EPIBONT	Percent cover of epibionts on *Ascophyllum nodosum*	Percent cover (%)	Logit transformed	Numeric
OTHER_INVERTEBRATE	Percent cover of other sessile invertebrate species (non-listed)	Percent cover (%)	Logit transformed	Numeric
ROCK	Percent cover of bare rock substrate	Percent cover (%)	Logit transformed	Numeric
OTHER_SUBSTRATE	Percent cover of substrate types not listed elsewhere	Percent cover (%)	Logit transformed	Numeric
NOT_SAMPLED	Percent cover of quadrat area not surveyed or unable to be sampled	Percent cover (%)	Logit transformed	Numeric

COMMON_PERIWINKLE_LITTORINA_LITTOREA	Abundance of common periwinkle (*Littorina littorea*) per plot	Count	None	Numeric
SMOOTH_PERIWINKLE_LITTORINA_OBTUSATA	Abundance of smooth periwinkle (*Littorina obtusata*) per plot	Count	None	Numeric
ROUGH_PERIWINKLE_LITTORINA_SAXATILIS	Abundance of rough periwinkle (*Littorina saxatilis*) per plot	Count	None	Numeric
DOGWHELK_NUCELLA_LAPILLUS	Abundance of dogwhelk (*Nucella lapillus*) per plot	Count	None	Numeric
LIMPET_TECTURA_TESTUDINALIS	Abundance of limpet (*Tectura testudinalis*) per plot	Count	None	Numeric
COMMON_PERIWINKLE_LITTORINA_LITTOREA_MEAN_MEASURE	Mean shell length of common periwinkle (average of up to 10 individuals if abundance > 10)	Millimeters (mm)	None	Numeric
SMOOTH_PERIWINKLE_LITTORINA_OBTUSATA_MEAN_MEASURE	Mean shell length of smooth periwinkle	Millimeters (mm)	None	Numeric
ROUGH_PERIWINKLE_LITTORINA_SAXATILIS_MEAN_MEASURE	Mean shell length of rough periwinkle	Millimeters (mm)	None	Numeric
DOGWHELK_NUCELLA_LAPILLUS_MEAN_MEASURE	Mean shell length of dogwhelk	Millimeters (mm)	None	Numeric
LIMPET_TECTURA_TESTUDINALIS_MEAN_MEASURE	Mean shell length of limpet	Millimeters (mm)	None	Numeric

References

Cheng, H., M. D. McMahan, S. B. Scyphers, L. McClenachan, and J. H. Grabowski. 2025. Observations, perceptions and concerns of the American lobster industry regarding the range-expansion of Black Sea Bass. Marine Policy 173:106517.

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

Cohen, A. N., J. T. Carlton, and M. C. Fountain. 1995. Introduction, dispersal and potential impacts of the green crab Carcinus maenas in San Francisco Bay, California. Marine Biology 122:225–237.

Epifanio, C. E. 2013. Invasion biology of the Asian shore crab Hemigrapsus sanguineus: A review. Journal of Experimental Marine Biology and Ecology 441:33–49.

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.

Johnson, D. S. 2015. The Savory Swimmer Swims North: A Northern Range Extension of the Blue Crab Callinectes Sapidus? Journal of Crustacean Biology 35:105–110.

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.

Trott, T. J. 2022. Mesoscale Spatial Patterns of Gulf of Maine Rocky Intertidal Communities. Diversity 14:557.

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.

Yamada, S. B. 2001. Global invader: the European green crab.