Application of machine learning models to predict fish species distributions in Alaskan estuaries

Katharine Miller, AFSC Auke Bay Laboratories
Katharine.miller@noaa.gov

Introduction

- Understanding how habitat influences the distribution and abundance of marine organisms is fundamental to successful fisheries management and conservation; however, for most marine species in Alaska this information is severely lacking, especially for early life stages.

- Species-habitat relationships are complex making them difficult to evaluate with standard statistical methods. Estuarine species, in particular, tend to have wide environmental tolerances and their response to environmental factors can be nonlinear or discontinuous. Furthermore, most environmental factors do not act on organisms independently: the relationship between a response variable and an explanatory variable often is affected by one or more other explanatory variables.

- The objective of this research is to develop spatially distributed predictive models of species occurrence and abundance for estuarine associated species in Southeast Alaska.

Sampled (red) and predicted (yellow) estuaries

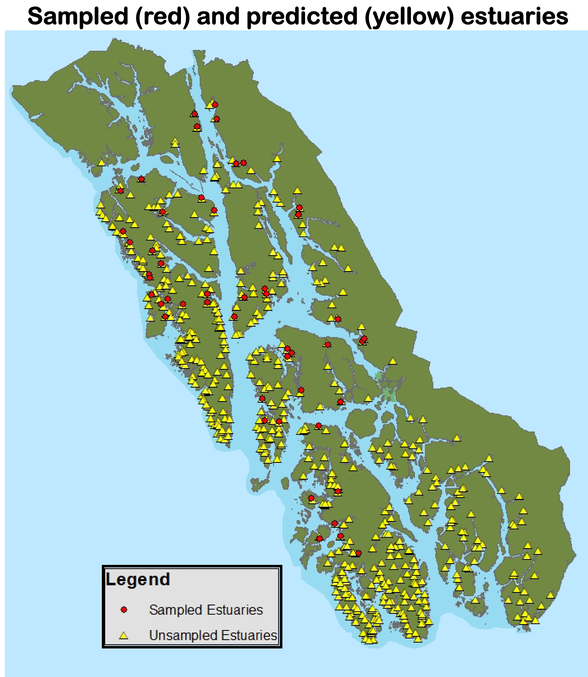

Figure 1: I developed a GIS of 541 estuaries (yellow) in Southeast Alaska and compiled spatial data for each estuary including: bathymetry, size amount of intertidal area, amount of open water, width at mouth, distance to major channel or open ocean, average precipitation (annual and monthly), degree of watershed fragmentation, percent cover of upland vegetation, and number an length of streams. Fish and invertebrate data was taken from a dataset of seine and trawl samples from 54 estuaries (red).

Methods

Occurrence and count data were analyzed using the Random Forest algorithm implemented in R (Breiman, 2001, Liaw and Weiner, 2001).

Random Forest:

- Data are sampled with replacement to construct a large number of regression trees.
- Approximately 70% of the data are used to build each model, the other 30% (termed out-of-bag sample) is used to test the model predictions
- At each node of the tree, a random subsample of predictor variables is used and the best split is chosen and the predictions of all of the individual trees are averaged .
- At each bootstrap iteration, the model is used to predict the out-of-bag sample and results are aggregated to obtain an error estimation for the model.
- Variable importance is calculated as the decrease in mean squared error of the model when the variable is permuted.

For species occurrence data, model predictive power was determined using the receiver operator characteristic (ROC) which plots the true positive rate against the false positive rate of the model's predictions. For count data, reduction in root mean squared error (RMSE) was used to select the best model . Model results were plotted in the GIS as either probability of occurrence or predicted abundances.

Predicted species occurrence

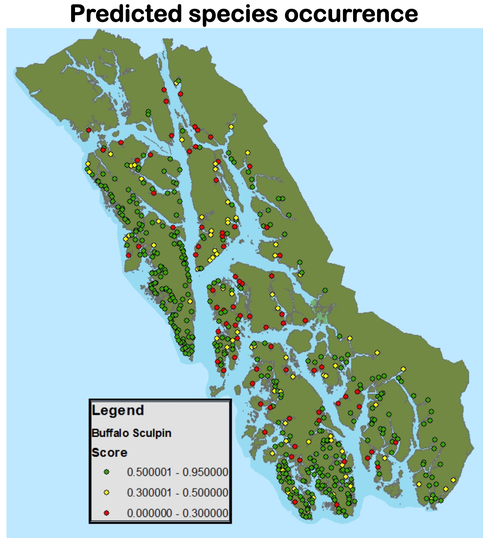

Figure 3: : Occurrence probabilities for Buffalo sculpin are plotted in the GIS. Sites with green dots have occurrence probabilities higher than 0.5. The GIS map of prediction can assist in evaluating regional patterns in species distributions and identify estuaries with similar environmental characteristics.

Results

- Preliminary model results show the predictive accuracy of models for species presence/absence ranged from 58% to 80% for fish and 57% to 90% for invertebrates based on the receiver operator characteristic (ROC). Accuracies for fish were calculated using independent data.

- Count models explained between 3% and 60% of the variation in species abundance.

- Variables important for predicting species occurrence often differed from the variables important in predicting abundance.

- The models were able to capture nonlinearities in the relationship between species occurrence/abundance and environmental variables.

Fish species and ROC scores

Species	Common Name	ROC
Enophrys bison	buffalo sculpin	0.8
Oncorhynchus kisutch	coho salmon	0.79
Oncorhynchus keta	chum salmon	0.77
Pallasina barbata	tubenose poacher	0.77
Blepsias cirrhosus	silverspotted sculpin	0.76
Gadus macrocephalus	Pacific cod	0.75
Syngnathus leptorhynchus	bay pipefish	0.75
Lepidopsetta spp	rock sole	0.75
Limanda aspera	yellowfin sole	0.75
Oncorhynchus gorbuscha	pink salmon	0.73
Gasterosteus aculeatus	threespine stickleback	0.72
Leptocottus armatus	staghorn sculpin	0.71
Aulorhynchus flavidus	tubesnout	0.71
Salvelinus malma	Dolly varden	0.71
Hexagrammos decagrammus	kelp greenling	0.7
Lumpenus sagitta	snake prickleback	0.67
Oncorhynchus nerka	sockeye salmon	0.67
Platichthys stellatus	Starry flounder	0.67
Clupea pallasi	Pacific herring	0.66
Theragra chalcogramma	walleye pollock	0.65
Podothecus accipenserinus	sturgeon poacher	0.65
Cymatogaster aggregata	shiner perch	0.64
Ammodytes hexapterus	Pacific sand lance	0.61
Hexagrammos stelleri	whitespotted greenling	0.6
Pholis laeta	crescent gunnel	0.58

Table 1: Receiver operator characteristic values for fish species presence/absence (***preliminary results***). Values above .60 generally regarded as fair-to-good predictors.

Conclusion

The Random Forest machine learning algorithm has demonstrated good predictive ability for estuarine fish occurrence and abundance. These models have important implications for evaluating spatial patterns in species distributions for conservation and management.

References: Breiman, L., 2001. Random forests. Mach. Learn. 45, 5-23, Liaw, A., Weiner, M., 2002. randomForest: Classification and Regression with Random Forest. R. Package

The recommendations and general content presented in this poster do not necessarily represent the views or official position of the Department of Commerce, the National Oceanic and Atmospheric Administration, or the National Marine Fisheries Service.