Catchment Level Predictive Mapping of 535 Fish Species Distributions

Douglas Patton

January 26, 2022

Abstract

Using an existing dataset of wadeable freshwater stream fish inventories, we apply statistical learning techniques to produce predictive distributions for the Continental United States (CONUS). We use around *400* of the catchment-level metrics in the StreamCat dataset as predictors. We created classification machine learning pipelines with both linear and non-linear statistical learners and we estimated both full-data classification models and probabilistic models. To communicate modeling uncertainty, we also use the cross-validation results to generate maps showing prediction intervals. To quantify the uncertainty of our predictive models

1 Introduction

In order to provide the public with enhanced fish distribution maps and predictive mapping tools, we apply machine learning classification modeling to create predictive maps using an existing dataset of electrofishing inventories spanning much of the Contiguous United States (CONUS).

2 Methods

2.1 Data

We utilize the dataset of electrofishing inventories from citecyterski2000 to create a binary dataset of catchments labeled for each species as present if the species had been found there at least once and absent otherwise. In comparison to citecyterski2000, who defined occurrence as a species appearing in the majority of visits, our approach will create a dataset with a greater share of species occurrences rather than absences for sites visited more than one time. Across the species in our dataset, the distribution and relationship between sample size and the average occurrence rate, $\bar{y}$, can be seen in 2 and 1. The lower left half of 1 is empty due to the lower limit on detecting a species with low occurrence rates for a given sample size, $\bar{y}_{min} = 1/N$.

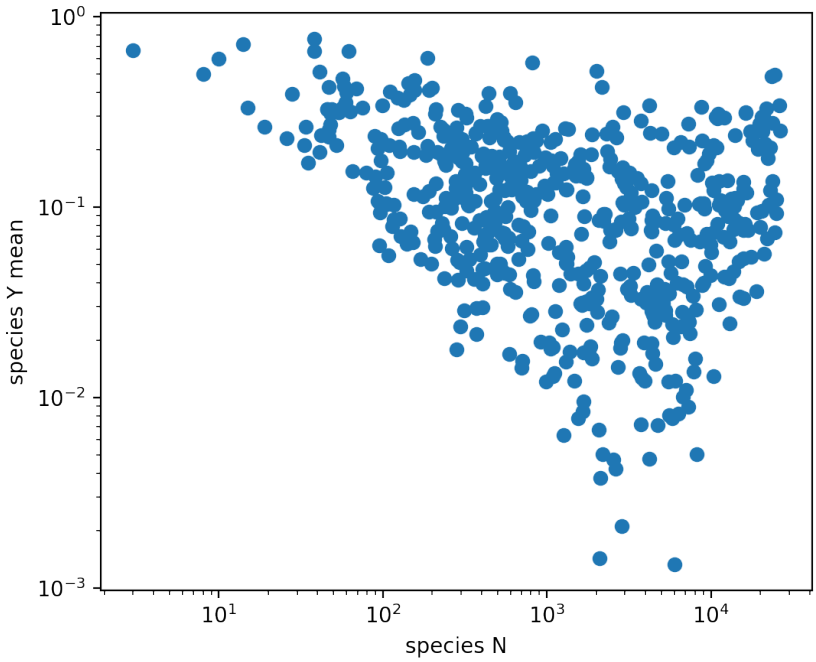

Figure 1: A scatter plot of species sample size on the horizontal axis and species average occurrence rate on the vertical axis

2.2 Modeling

The foundation of our classification modeling is the scikit-learn citepedregosa pipeline.

3 Results

The first set of our results is a basic overview of modeling performance across geographies To provide an overview of modeling performance and aggregate fish distributions, we 3

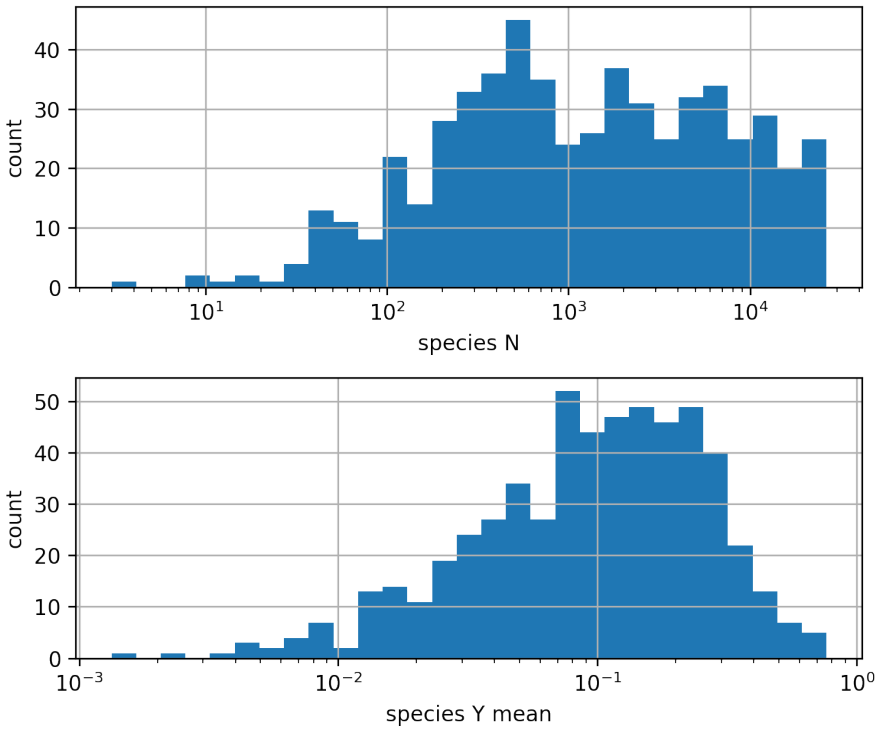

Figure 2: A histogram of each species sample size and each species mean occurrence rate

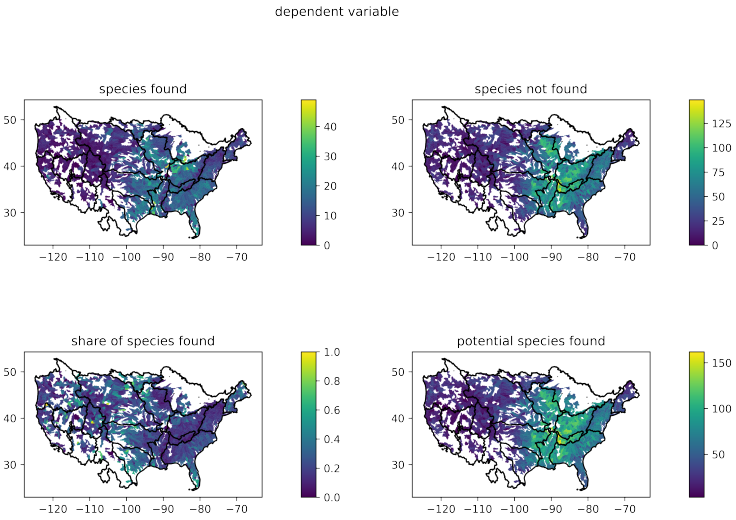

Figure 3: A confusion matrix map with average metrics across comids in each HUC8