Data Analysis Workflow: Habitat Modeling

Executive Summary

This document outlines the logic and processing steps involved in modeling habitat suitability for aquatic indicator species to estimate the potential ecological uplift of management options on aquatic biodiversity.

Scope

This process is applied to each indicator (fish species), with 5-6 species assessed for each ecoregion in NC. The unit of analysis is the NHD+ catchment, and the focus is on catchments in which each species has been observed to occur; we are using occurrence data compiled by Mark Endries for this purpose (Endries 2014). The scope of habitat suitability analysis is intended to be the ecoregion (i.e., Mountain, Piedmont, or Coastal Plain) but for purposes of this analysis the scope is defined by the geographic range of each species. In this, we estimate the range as those HUC8 watersheds containing the NHD+ catchments in which each species has been observed. This ranges serves as a useful approximation of where the species *might* have been observed. In practice, this range is assembled from HUC8 watersheds that might extend beyond the focal ecoregion. For convenience, data processing is done at the scale of the HUC6 basin; larger ranges can be collated from more than one HUC6 while smaller ranges can be clipped as needed to the appropriate HUC8’s.

Predictions of habitat suitability and potential changes in suitability due to management are made at the scale of the Cataloging Unit, or HUC8. This means that the habitat suitability modeling is done just one time for each of the indicators—at the scale of its potential range—while estimates of uplift can be repeated for many different Cataloging Units using the same model.

Processing Logic

Processing a habitat suitability model proceeds in several stages: (1) assembling the raw data sets needed for each species; (2) exploratory data analysis and editing of these data; (3) estimating the habitat suitability model(s); and (4) applying this suitability model to an analysis area (e.g., HUC8).

Two data sets are required for this assessment:

(1) A set of catchments, and the habitat or environmental descriptors associated with these catchments, for all catchments in which the species has been observed to occur. These observations constitute the “presence” samples.

(2) A set of catchments for all HUC8’s in which the presences occurred, with the environmental variables associated with these samples. These constitute the samples in which the species might plausibly been observed, referred to here as “available” habitats but also called “background” samples or “pseudo-absences in other modeling literature.

The aim of the habitat suitability modeling is to statistically contrast the “presences” from the “available” habitat samples.

1. Data Assembly

Two master data sets are used in the processing. The first is a file that records, for each species, the number of observations of that species in each NHD+ catchment in NC. In this, the samples (rows) are catchments and the columns are species. This data set was collated from the data provided by Mark Endries. Note that this file does not include all catchments in the state; it includes only those catchments in which focal species have been observed. This is file spp\_freqs.csv.

The second file includes all habitat and environmental predictors for each NHD+ catchment, collated at the scale of a 6-digit HUC. For example, the Santee HUC6 includes the Catawba HUC8, one of our pilot watersheds. In the example documented here, this is HUC 030501, as huc030501.csv.

The habitat data were pre-processed in ArcGIS (processing documented separately) and exported in CSV format. All further processing and habitat modeling is conducted in the R environment (R Development Core Team 2013).

Data processing entails using these two master files to create a “presences” and an “available” data set for each species.

### Species “Presences”

This process begins by extracting a single focal species from the species frequencies file, saved as a data frame with the NHD+ catchment identifier (featureID), the HUC8 to which the catchment belongs (HUC8), and the species (spp). This file is then match-merged with the habitat data (huc.data). Because these data sets match on *only* the featureID, the merging creates a new data set that includes the featureID and the habitat variables from huc.data: habitat data for the “presences”.

|  |  |
| --- | --- |
| Field name | Description |
| featureID | NHD+ catchment identifier |
| HUC8 | 8-digit hydrologic unit code in which the catchment occurs |
| spp | Species name |

### “Available” Habitat

The second data set consists of habitat data for all catchments for all HUC8’s in which the “presences” occur. The list of these HUC8’s is extracted by finding the unique names in this field of the “presences” data set. A utility function huc8catchments (huc8catchments.R) is then used to collate the habitat samples for all of these catchments. Note that if the species range extends beyond a single HUC6, this step must be repeated for each HUC6 and the resulting data sets concatenated into a single “available” data set.

These two data sets are then concatenated into a single file for subsequent analysis. In this, an identifier column is attached to the data frame, named spp and coded 1 for species presences and 0 for the available habitat samples.

2. Exploratory Data Analysis

The data set assembled in the first step includes a large number of variables, only some of which will be used to estimate a habitat suitability model. The data set is edited in several steps.

### Missing Values

Missing values are tallied using utility function tallyNAs (tallyNAs.R), and variables with missing values are discarded.

### Correlations with Species Occurrence

The potential habitat variable are correlated with species occurrence (variable spp) using utility function SHcor.R. This function returns a file with the names of each habitat variable (as rows) and the estimate and P-value of its correlation (as columns) with the occurrence the focal species. This is essentially a *t*-test contrasting the 0’s and 1’s coded for spp. Variables with nonsignificant correlations are discarded before habitat suitability modeling.

### Redundant Habitat Variables

The remaining habitat variables (those with correlations significant at *P* < 0.05) are then tested for redundancy using utility function screen.cor (screen\_cor.R). Variables with correlations much stronger than 0.70 are discarded as redundant (cite).

### Variable Selection (Ordering) for GLM

In the case of the generalized linear model (GLM), the order that predictor variables are entered into the regression formula influences the parameter estimates (as with any regression, the estimates are order-dependent). While there is no universal rule for how to order the variables, in this application the variables are entered in order of the absolute magnitude of their correlation with species occurrence. This requires sorting the variables in decreasing order of these correlations.

Finally, the reduced and nonrendundant (and ordered, for a GLM) set of variables is used to estimate habitat suitability.

3. Habitat Suitability Modeling

There are several approaches to modeling habitat suitability and in this project we are evaluating three models as part of the pilot study. These include the GLM, maximum entropy modeling (maxent), and a presence-only model based on ecological dissimilarity.