Data Analysis Workflow: Habitat Modeling

Executive Summary

This document outlines the workflow used to determine the level at which measured catchment attributes are associated with habitat integrity across Conservation Units (HUC8s) in North Carolina. This workflow presumes that HUC 6 level databases have been generated using procedures outlined in the [Script Metadata](https://github.com/Duke-NSOE/EEP/blob/master/Scripts/SCRIPT%20METADATA.docx?raw=true) document. In particular, this workflow requires the table (derived from Mark Endries’ species occurrence data) listing NHD catchments in NC and indicators for each species whether it was observed in that catchment or not. It also requires the feature class for all NHD catchments in the HUC 6 tagged with the ~120 attributes determined in the Script Metadata document.

Contents

[🡪Creating the statistical model data file 1](#_Toc425862288)

🡪Creating the statistical model data file

[STATS\_CreateDataFile.py](https://github.com/Duke-NSOE/EEP/blob/master/Scripts/STATS_CreateDataFile.py)

This script merges species presence records (extracted from the *AquaticSpeciesPoints* Table in the NC geodatabase) with NHD catchment attributes (*ResponseVars* feature class generated for each HUC 6 in earlier steps) to create a full table of all the catchments within all HUC8s in which a user selected species occurs.

The script outputs include a new folder in which all habitat model results will be stored (typically in the MaxEnt folder) and the statistical modeling table. The output folder is named with the species’ name (e.g. “Moxostoma\_collapsum”), and the output table is named “AllHUC8Records.csv”.

The output table includes a column listing, for each catchment, whether the species was recorded in that catchment and values for all response variables calculated for the catchments. However, any response variable containing one or more null values for any catchment will be removed. Null values cannot be processed in the statistical models and, given the paucity of species occurrence records, we felt it better to remove response records with null values rather than species occurrence points.

🡪Screening response variables for correlations with presence/absence

[STATS\_SHCorrelate.py](https://github.com/Duke-NSOE/EEP/blob/master/Scripts/STATS_SHCorrelate.py)

This script identifies response variables that can be eliminated from the analysis from lacking a significant correlation with whether a species occurs in a catchment or not. It computes the Pearson product moment correlation between a given response variable and the binary presence/absence variable. If the correlation is not significant (p < 0.05), the response variable is considered more noise than signal and is tagged for eliminated from the habitat modeling.

The output of this script is a CSV file listing only significant response variables (p < 0.05) along with their Pearson product moment correlation value, the absolute value of the correlation value (for sorting), and the P value of the correlation. This output is typically created in the species’’ habitat modeling output folder (created above), and is called “Correlations.csv”.

*This script uses the Python “SciPy” module which must be installed separately. Furthermore, a legacy version of the SciPy module (v.0.12.0) must be installed as that is the one compatible with other Python modules installed with ArcGIS 10.2. This version of SciPy can be downloaded here:* <http://bit.ly/1IOE9oy>.

🡪Calculating cross correlated response variables

STATS\_CalculateCrossCorrelation.py

This script calculates the Pearson product moment correlation between pairs of response variables. It produces a CSV formatted table listing all variable pairs with a correlation coefficient above a threshold of 0.7, i.e., variables considered redundant with each other.

Inputs to the script include the complete list of response variable values for all catchments within the identified HUC8s (output of the STATS\_CreateDatafile.py script), the list of response variables determined to be correlated with presence/absence (output of the STATS\_SHCorrelate.py script), and a threshold value, typically set to 0.7.

🡪Calculating cross correlated response variables

STATS\_VisualizeCorrelations.py

This script generates an HTML document that, when viewed in a web browser, enables the user to visualize redundant variables and select which one should be kept for subsequent habitat modeling. This script makes use of a JavaScript visualization plugin hosted by <http://visjs.org> to display graphs interactively.

The page generated displays each response variable as a set of points, or nodes. The size of the nodes reflect the variable’s correlation strength with presence/absence and the node’s color reflects the response variable’s importance in linking it with management activities (described later in this document).

The script’s inputs include the species’ name (to display on the web page), the table listing each response variable’s correlation with presence absence