**MMI – 2024 DOCUMENTATION**

**SLH = lead**

[**SITE LEVEL INFO:**](file:///C:\Users\shubler\Desktop\BioMonORDEQ\bugs%20analyses\MMI\Bio_MlocIDs_AWQMS_Most%20disturbed.SYB.xlsx)

* AWQMS interface was down, so as in O/E models, we based the site pulls off of work done outside of the Stations table
* Ref/Disturbed Designations – we used what was in this file, originally pulled from Stations/One.Table\_Rule.All
  + But we recognized the Most Disturbed population never had any sort of BPJ oversight.
  + Sabine visually scanned all sites originally labeled as Most Disturbed AND bugs were available (n = 286) and brought to the Reference Council for review any site she was uncertain of.
  + 23 sites were changed by Ref Council BPJ to moderately disturbed. Final Most Disturbed population n = 263.
  + SLH manually edited the Bio\_MlocIDs\_AWQMS\_Most disturbed.SYB.xlsx file to update the 23 sites to Moderately Disturbed.
    - These sites need to be updated in the Stations Table. (see file in email from SYB)
* COMID: use same “COMID” and “Nearby\_COMID” fields for associated WS /CAT metrics, as was done in O/E model.
* Reference Sites: use the same dataset as used for O/E
  + SLH: confirmed, included in R-code to limit to same ref pop as O/E
* Most disturbed:
  + Need to winnow down the sites to one per stream and eliminate spatial overlap
  + SLH: confirmed. Done in Excel, similarly to O/E.

**BUG DATA**

* AWQMS queries broken at this time, so using the data download that Lesley complied from AWQMS through (?) SQL queries (?). (‘raw\_bugs.Rdata’)
* NEED METRICS, not raw data
  + Use BCG attribute table?
  + BCG table is vetted and maintained. Better to use that one.
    - Requires adapting current METRICS code
    - DECISION: use BioMonTools from Erik Leppo?
* Tetra Tech PNW BCG attributes/taxonomy
* USU verified that bugs needs to be rarified. At least for richness metrics.
  + Begin, start with rarified bugs for all metrics. If poor model performance, we can explore running metrics with and without rarified bugs.
  + Travis fix for rarify: Need to change object to data.frame to work with rarify, otherwise this error comes up
* Error in Ops.data.frame(inbug[, sample.ID], isamp) :
* ‘==’ only defined for equally-sized data frames
* Drop all samples with < 200 total count (after rarifying).

**BUG METRICS**

* **New process: used BioMonTools and PNW Attribute table**
  + R package developed by Erik Leppo (Tetra Tech, available on GitHub)
  + PNW Attribute table: available through BioMonTools
    - Managed by Sean Sullivan (Rhithron)
    - Discovered HABIT was missing, so Sean created a working copy for me to use until it was able to be uploaded to BioMonTools
  + Aligned DEQ bugs to BioMonTools taxonomy.
  + Used that linkage to join with attributes, which are used for metric calculations
  + “Excluded” function allows for calculation of richness metrics (nt = # of taxa, pt = % of taxa), without need for data to have labs submit “Unique taxon”.
* Hundreds of metrics calculated.
  + Several dropped because of “infinite” values, or no non-zero results.

**PREDICTORS**

* Used same StreamCat predictors as O/E.
* Similarly used WS metrics at sites with matching COMID, and CAT metrics for sites without matching COMID.

**MMI Modeling**

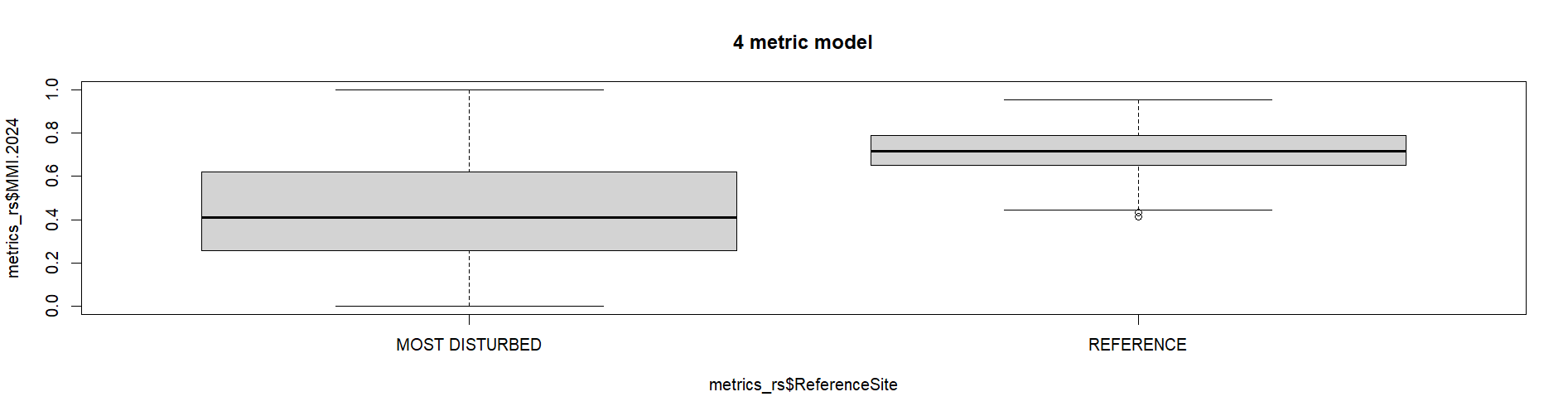
* Followed standard procedures used by USU
* See *Hawkins, C.P., Cao, Y. and Roper, B., 2010. Method of predicting reference condition biota affects the performance and interpretation of ecological indices. Freshwater Biology, 55(5), pp.1066-1085.* (Saved in IR 2026/Methodology/FW BioCriteria)
* RF models for each metric
  + Criteria for using RF modeled metrics: >= 10% variance explained (good.mods)
  + If didn’t meet this criteria, used the raw metric in the next steps (poor.mods)
* Get predictions for each metric, based on rf models
* Get residuals
* T-tests between reference and most disturbed
  + Residuals for good.mods
  + Raw metrics for poor.mods
* PCA (principal components analysis) used to identify metrics with similar axis loadings (correlated metrics)
* Selecting final MMI
  + Combine PCA and t-vals: ‘pca.tval\_select FINAL FINAL metrics\_5.8.24.xlsx’
  + Used absolute values for each PCA to identify the metrics with the strongest axis loadings (> 0.7, or 0.6 for last axis)
  + Then sorted by absolute value of t-values
  + Identified what “type” of metrics were included on each axis.
  + Optimal choice is take the highest t-value for each PCA axis
    - Higher ‘t’ = greater discrimination between reference and most disturbed samples
  + However, also had the goal of maintaining ecological independence in the metrics chosen for each axis.
    - In other words, we wanted different metric types selected for each axis
    - E.g., if intolerance is highest in Axes 1 and 4, only select one axis to include an intolerance metric.
    - This helps reduce redundancy in the metrics and expands the types of disturbances potentially identifiable by the MMI.
  + Ultimately, selected two candidate models

|  |  |
| --- | --- |
| **1st choice** | **2nd choice** |
| pt\_habitat\_rheo\_resid | pt\_tv\_intol\_resid |
| nt\_habit\_cling\_resid | nt\_habitat\_rheo\_resid |
| pt\_ti\_stenocold\_cold\_cool\_resid | pt\_ti\_stenocold\_cold\_cool\_resid |
| pi\_tv\_intol\_resid | pi\_POET |
| pi\_Pleco\_resid |  |

* + Final decision = t-tests between reference and most for the final MMI score (average of all metrics in MMI)
    - 5 metrics:
      * t = -12.616, p < 2.2e-16
      * mean Most = 0.438, mean REF = 0.680
    - 4 metrics:
      * t = -13.412, p < 2.2e-16
      * mean Most = 0.461, mean REF = 0.732
  + Final Model
    - Pt\_tv\_intol: 36% variability explained
      * TMAX8110
      * CLAY
      * OM
      * KFFACT
      * PRECIP8110
    - nt\_habitat\_rheo: 18% variability explained
      * TMAX8110
      * ELEV
      * OM
    - pt\_ti\_stenocold\_cold\_cool: 34% variability explained
      * MSST\_mean08.14
      * KFFACT
      * TMAX8110
      * PERM
      * CLAY
      * AREASQKM
      * SLOPE
    - pi\_EPTNoHydro: 12% variability explained
      * PRECIP8110
      * KFFACT
      * CLAY
      * P205

**MMI PRECISION and SENSITIVITY**

* Responsiveness
  + Delta Mean MMI = 0.271
    - Ref = 0.732
    - Most = 0.461
  + T-value = -13.412
* Precision: standard deviation (SD) of reference MMI values
  + Reference (n = 221) SD = 0.103
  + Most Disturbed (n = 158) SD = 0.245
* Reference 10th percentile = 0.60
  + 5th = 0.55
* Sensitivity: % of most disturbed sites below reference 10th percentile
  + 113/158 = 70%
* Bias in final MMI?
  + Ran a randomForests model of the final mmi scores, against a suite of 28 ‘natural’ predictors
  + REFERENCE SITES only
  + RF results:
    - % var explained = -26.7%



**###**

**# Standardize MMI scores to compare to O/E models**

**###**

ref.X <- mean(mmi.ref$MMI.2024) # 0.727

mmi.ref\_stand.mean <- (mmi.ref$MMI.2024)/ref.X

mean(mmi.ref\_stand.mean) # 1

sd(mmi.ref\_stand.mean) # 0.142

quantile(mmi.ref\_stand.mean, probs = c(0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95)) # 0.82

**RUNNING MMI on new bug data**

* Pull bugs
* Associate StreamCat predictors to samples (stations)
* Subsample bugs to 300 ct
* Calculate metrics using BiomonTools
* Get EXPECTED metric values
  + Run metrics through saved randomForests models
* Calculate RESIDUALS (Observed metric – Expected metric)
* RESCALE residuals to 0-1
* FINAL MMI = average of rescaled-residual metrics