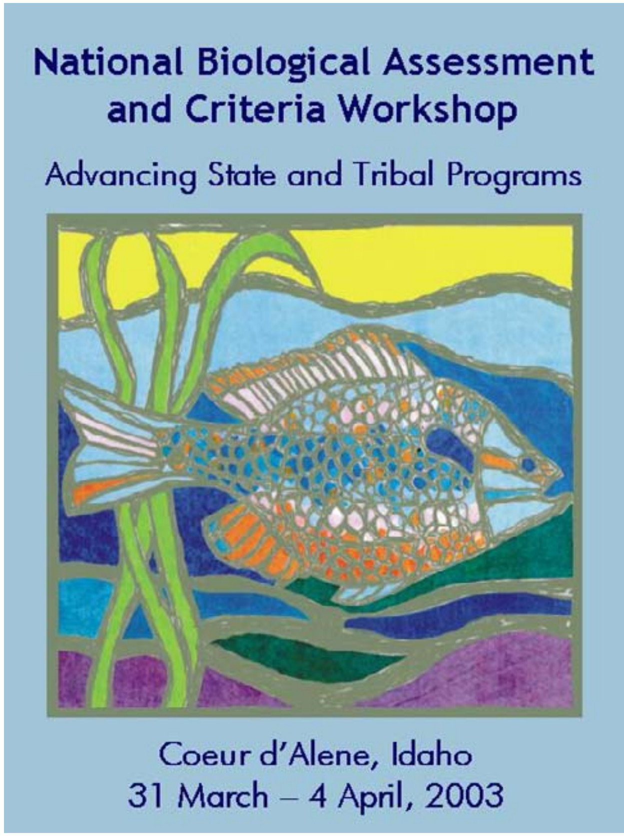


RIVPACS

Presented by
Chuck Hawkins, Utah State University;
Rick Hafele, Oregon Dept. of Environmental Quality;
Mike Paul, Tetra Tech, Inc.

Quick Review of 101

- Understanding the units of measure (O/E).
- *Predicting the expected taxa.*
- *Calculating O/E, the biological condition value.*
- Determining if an assessed site is impaired.

Focus of 201

- Mechanics
  - Predicting the expected taxa.
  - Calculating O/E.
- Application / Case Example

The accuracy and precision of RIVPACS-type assessments are completely dependent on how well we estimate the probabilities of capture of all individual taxa in the regional taxa pool.

Remember this example from 101?
(Units of Measure & the Expected Taxa)

	Replicate Sample Number										Freq
Species	1	2	3	4	5	6	7	8	9	10	(P_c)
A	*	*	*	*	*	*	*	*	*	*	1.0
B	*	*		*	*	*		*	*	*	0.8
C	*		*		*	*			*		0.5
D		*	*				*		*	*	0.5
E					*						0.1
Sp Count	3	3	3	2	4	3	2	2	4	3	2.9

Species Richness is the Currency.

$E = \sum P_c = 0$ number of species / sample = 2.9.

How do we estimate probabilities of capture from single samples at a site?

The basic approach to modeling pc's and estimating E was worked out by Moss et al.*

River InVertebrate Prediction and Classification System (RIVPACS)

*Moss, D., M. T. Furse, J. F. Wright, and P. D. Armitage. 1987. The prediction of the macro-invertebrate fauna of unpolluted running-water sites in Great Britain using environmental data. Freshwater Biology 17:41-52.

RIVPACS-type Models: 8 Basic Steps

1. Establish a network of reference sites.
2. Establish standard sampling protocols.
3. Classify sites based on their biological similarity.
4. Estimate individual probabilities of capture by relating environmental setting to the biological classification (multivariate statistics).

For each assessed site:

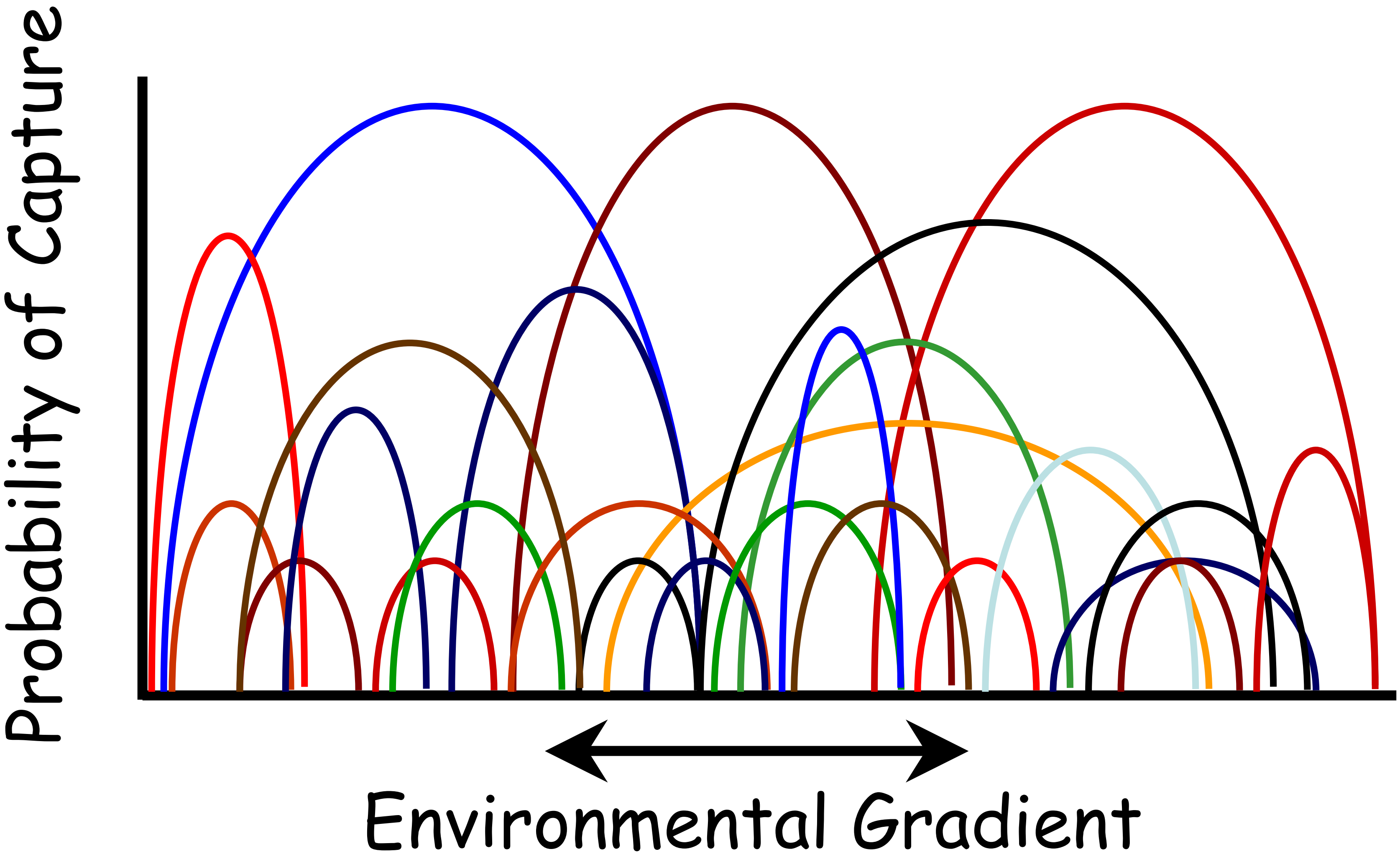
5. Sum p_c's to estimate E.
6. Count O
7. Calculate O/E.
8. Determine if observed O/E is different from reference?

The 'Complicated' Steps

3. Classify sites based on their biological similarity.
4. Estimate individual probabilities of capture by relating environmental setting to the biological classification (multivariate statistics).

In RIVPACS models, site classification is really just a clever mathematical shortcut toward predicting the continuous biological response that occurs along natural environmental gradients.

Remember, we ultimately want to be able to estimate the probabilities of capture of every taxon in the regional taxa pool at any location.



There are at least two approaches to modeling probabilities of capture

1. Logistic regression avoids classification and models each taxon separately. The output of these separate models can be combined to estimate E, the expected number of taxa, but…..

   many models would be necessary, and rare taxa are difficult to model!

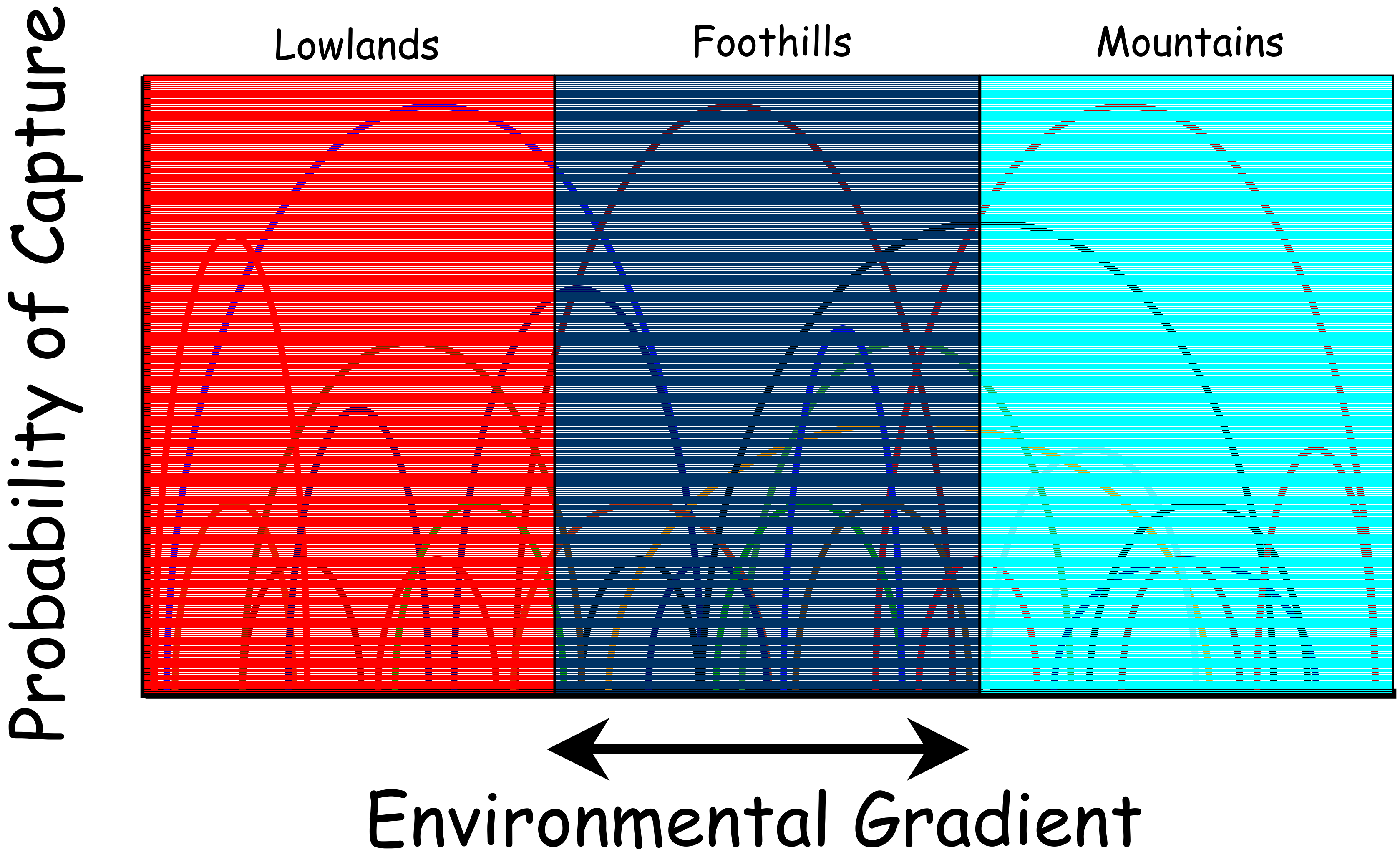
2. The RIVPACS approach creates 1 model and in doing so also potentially circumvents the rare taxa problem, but…..

   it requires some statistical machinations that are a bit complicated, including the *biological* classification of sites.

In RIVPACS, reference sites are classified based on their compositional similarity to one another

- This type of classification involves two steps:
  1. Calculation of a pairwise similarity matrix among all sites, followed by
  2. Cluster analysis to identify biologically similar 'classes'.
- A variety of methods exist for conducting both steps, but we would like to use the methods that result in the most precise predictions.

How can we let the biology define a classification that will allow us to later predict species composition at a site?



But how do we actually get the organisms to tell us where to 'draw the lines'?

Two Commonly Used Similarity (Distance) Measures

- Jaccard Distance = 1-(2W/(A+B-W))
- Sorensen (Bray-Curtis) = 1-(2W/(A+B))

  In both measures, W is the sum of shared abundances and A and B are the sums of abundances of taxa found only in individual sample units. Values of both measures range from 0 to 1. The Jaccard measure can be interpreted as % of taxa shared, but in the Sorenson measure, shared taxa are weighted.

- The Sorensen measure has generally been shown to be superior to the Jaccard measure for RIVPACS applications.

A simple example of calculating a similarity matrix: the raw data

Sites	Species A	B	C	D	E	F
1	1	1	1	1	0	0
2	1	1	1	0	0	0
3	1	1	0	0	1	1
4	0	0	0	0	1	1
5	1	1	1	1	1	1
6	0	0	0	1	1	1

The distance matrix based on the Sorensen Measure

	1	2	3	4	5	6
1	0.00					
2	0.14	0.00				
3	0.50	0.43	0.00			
4	1.00	1.00	0.33	0.00		
5	0.20	0.33	0.20	0.50	0.00	
6	0.71	1.00	0.43	0.20	0.33	0.00

A similarity or distance measure is the intermediate step to classification

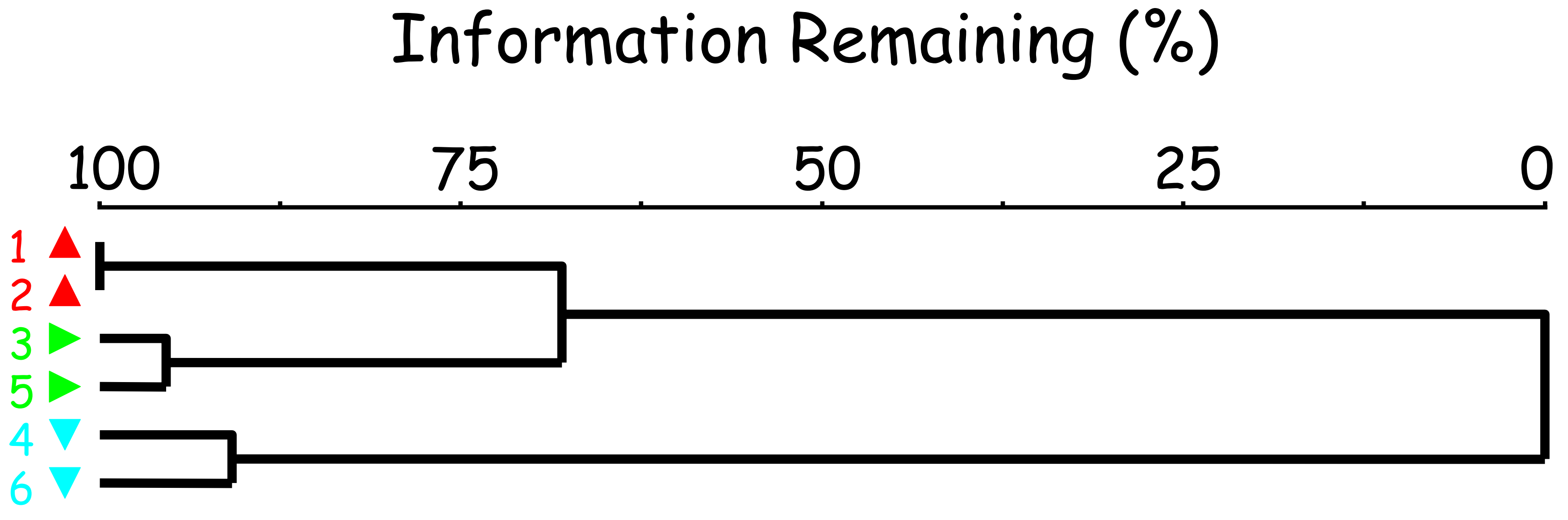
- The next step is to create a cluster diagram, which is produced by applying one of several possible clustering algorithms to the matrix. The different algorithms may produce different looking dendrograms and thus different classifications.
- Experience has shown that two methods produce better models:

  flexible beta and Ward's

The dendrogram produced from the practice data by flexible beta clustering.

So how many classes are there?

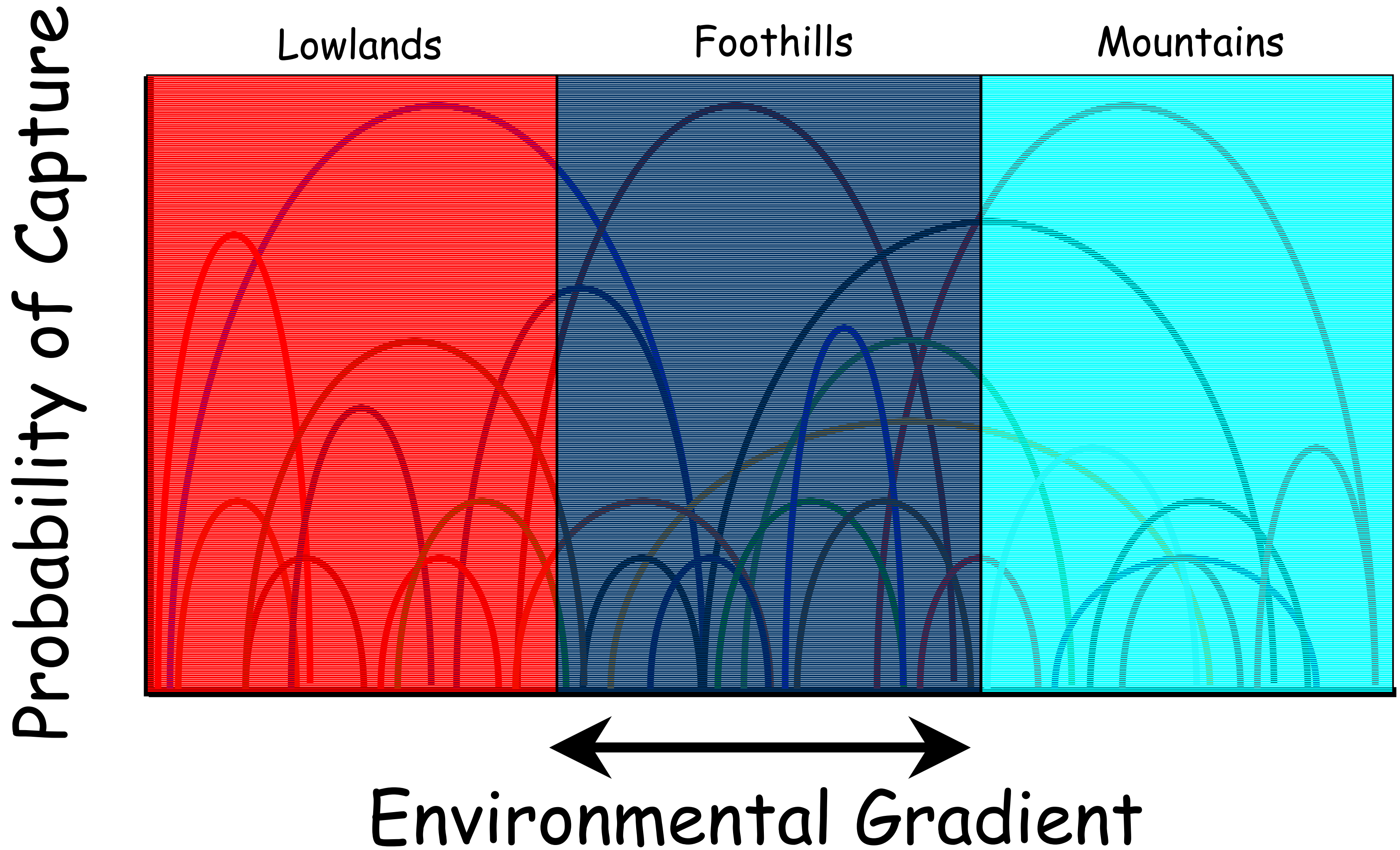
In general, for RIVPACS, classes should be defined as finely as possible as long as ≥ 5 sites occur within classes.



What do we do with the classification?

- If classes were truly discrete, we could calculate frequencies of occurrence of different taxa within classes, and use these values as estimates of probabilities of capture, but
- We know the classes are not discrete, they are simply the artifact of our chopping up a continuous world into chunks.

Some species occur only in one class, but not at all sites; other species occur in more than one class; no species occurs everywhere.



How do we apply this classification to new sites?

- This is the modeling part, and…
- how we predict continuous gradients from the 'discrete' classification that we produced.

The next 'step' is actually a series of 4 linked calculations

1. Calculate the frequencies of occurrence of each taxon within each class.
2. Estimate the probability that a new site belongs to each of the classes.
3. Use these probabilities of class membership to weight the frequencies of occurrence within classes.
4. Sum the weighted frequencies of occurrence for a taxon to estimate the probability of capturing that taxon at that site.

Estimate frequencies of occurrence of each taxon in each biotic class as (n_i/N).

Class	Sp 1	Sp 2	Sp 3	Sp 4	Sp 5	Sp 6
A	0.33	0.89	0	0.25	1.00	0
B	0.80	0.99	0.21	0.36	0.87	0
C	0.60	0	0.16	0.28	0.98	0.05
D	0.10	0.54	0.09	0.29	1.00	0

Derive a model to predict (from environmental features) the probabilities (P_G) that a new site belongs in each of the biologically-defined classes.

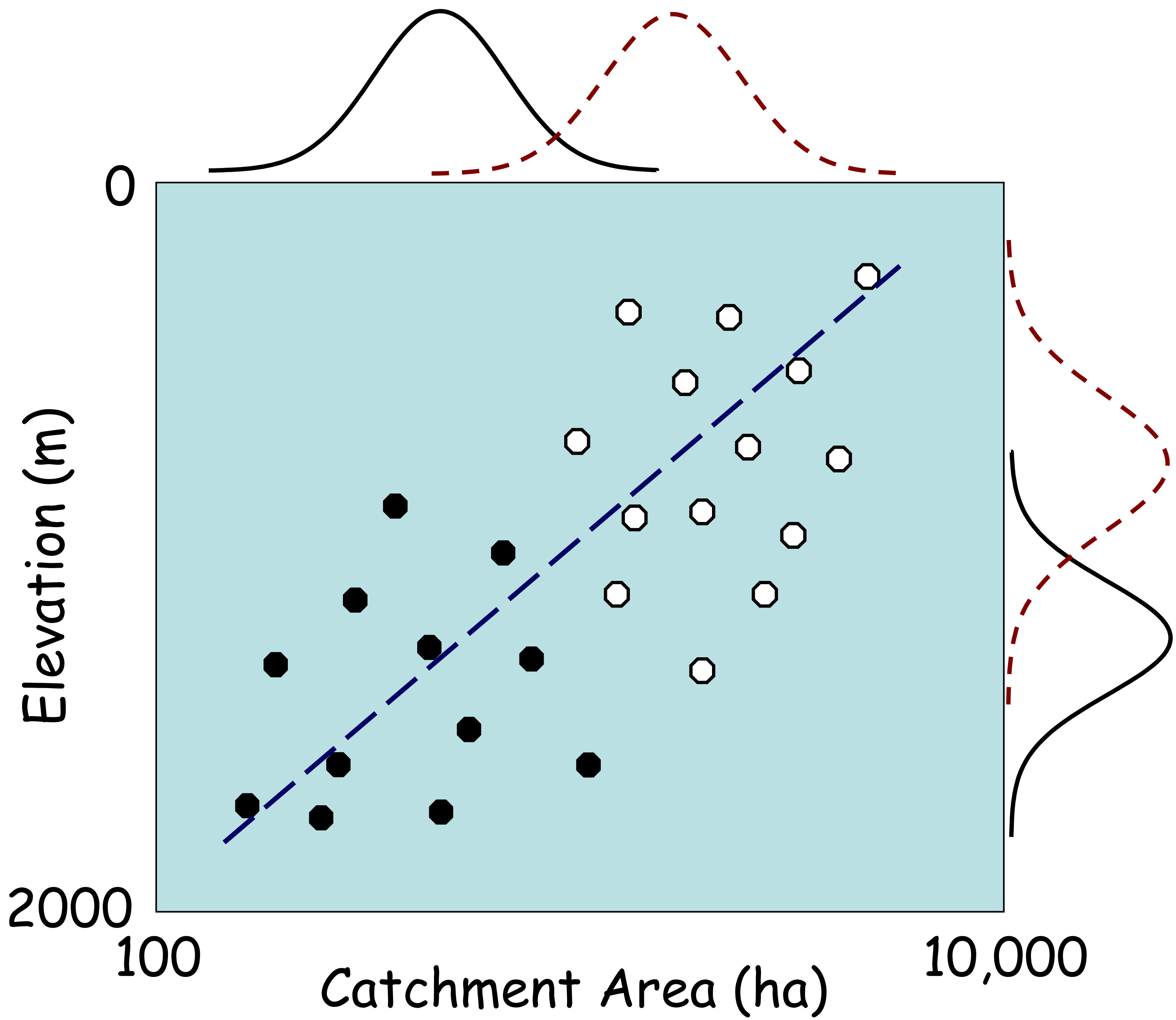
Discriminant functions, e.g.,

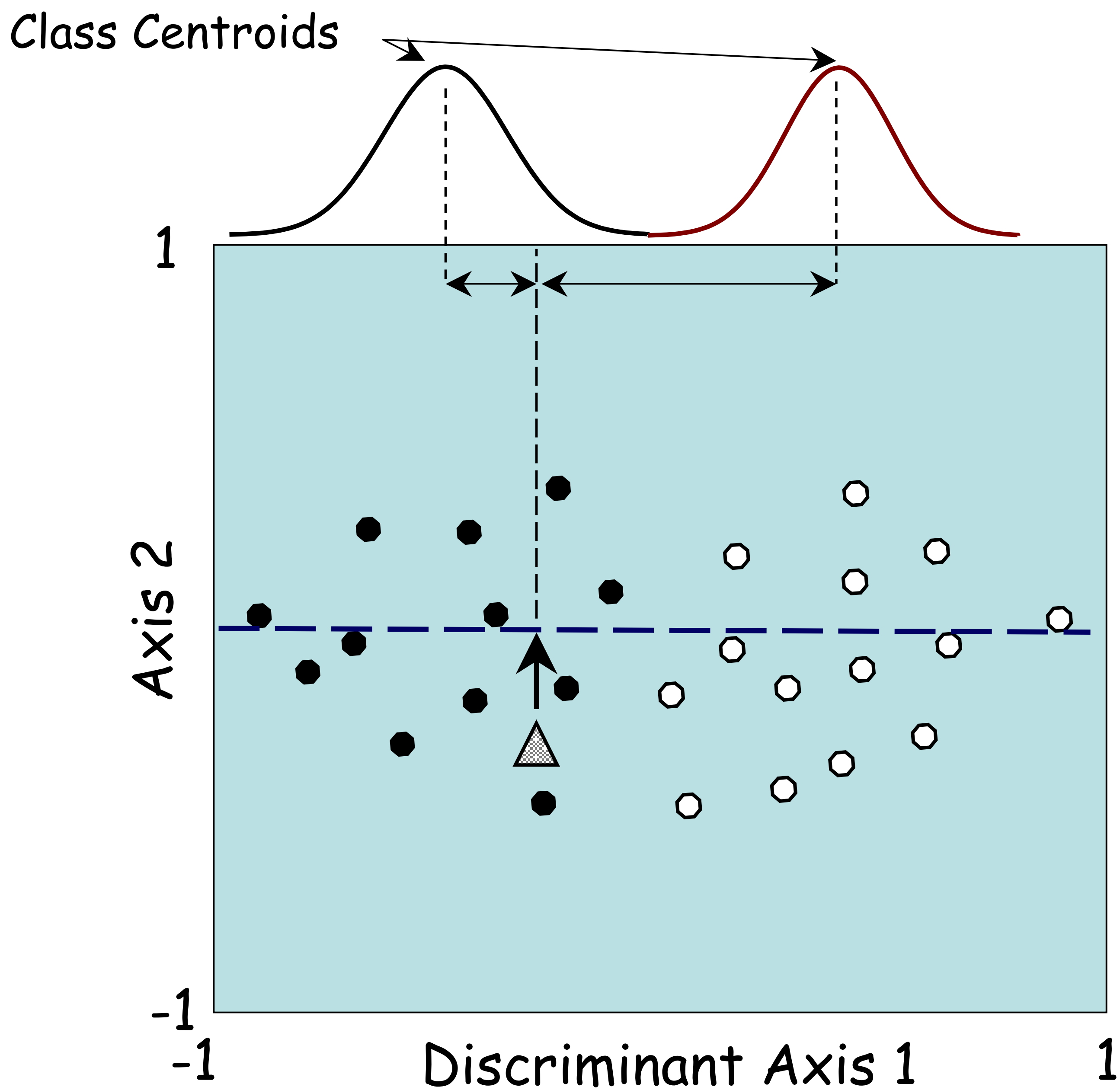
P_g = f(elevation, watershed area, geology)

Predictors should be insensitive to human alteration

This is not a class in multivariate statistical procedures, but… let's take a quick graphical look at how discriminant functions models work.

A simple graphical explanation of discriminant models

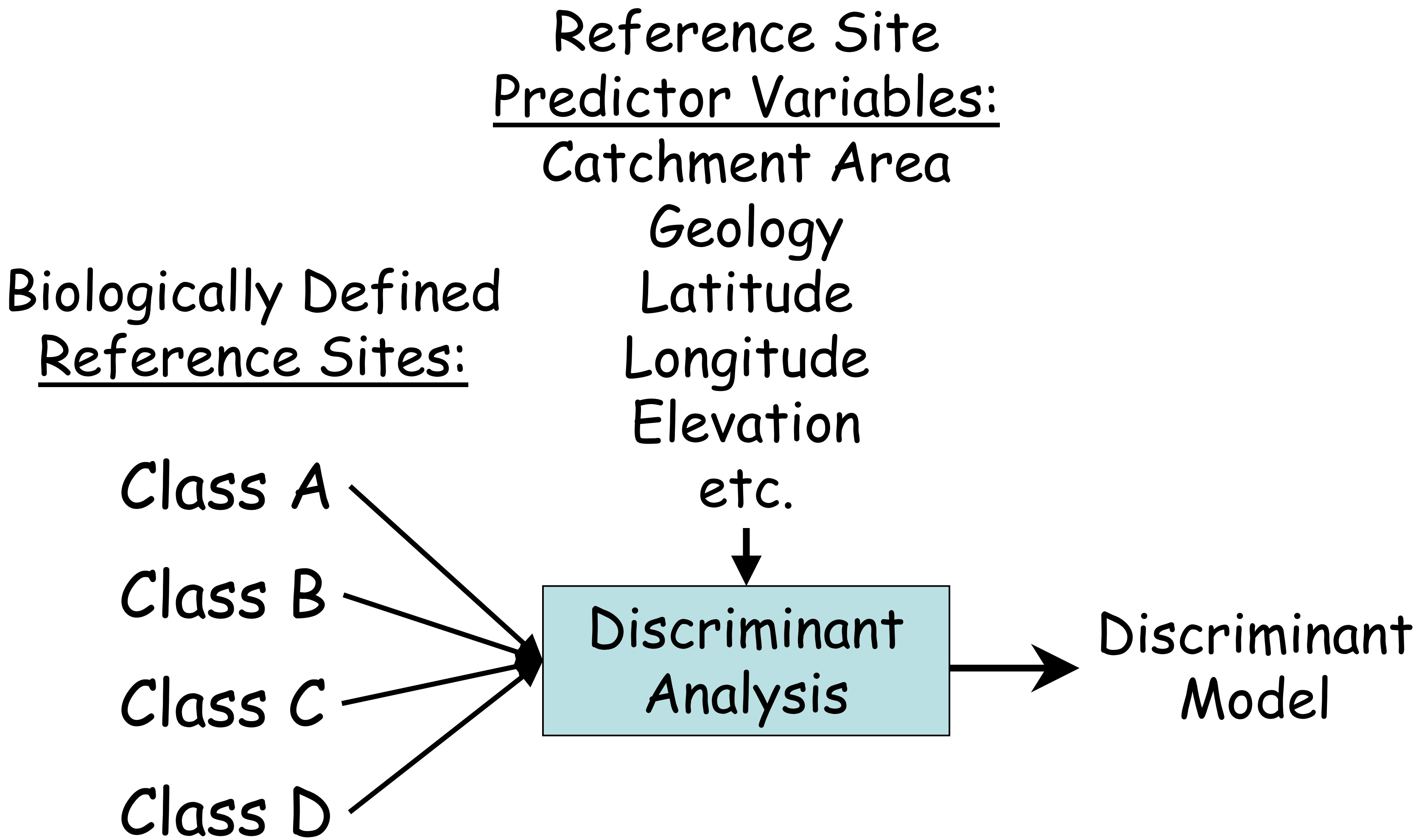




Centroids are the combination of predictor variables that represent the average site in a class. The taxa at a centroid will therefore represent the best estimate of the taxa expected at a site classified into a discrete class.

We can refine estimates of the taxa expected at individual sites by recognizing that nature is seldom discrete and using probabilities of class membership.

The Discriminant Model



Combining the Discriminant Model + Frequencies of Occurrence Provides Estimates of Probabilities of Capture

Test Site
Predictor Variables

↓

Discriminant Model

Class Probabilities		Taxon 1 Reference Site Frequencies	
50% Class A	*	60% Class A	= 0.30
40% Class B	*	20% Class B	= 0.08
10% Class C	*	0% Class C	= 0.00
0% Class D	*	0% Class D	= 0.00

$P_c = 0.38$

Weight frequencies of occurrence of taxa within classes ($F_{i,g}$) by (P_g) and sum to calculate p_c's for the new site.

Sp 1

Class	P_g	$F_{i,g}$	$P_g \times F_{i,g}$
A	0.50	0.60	0.30
B	0.40	0.20	0.08
C	0.10	0.00	0.00
D	0.00	0.00	0.00

$P_C = \Sigma (P_g \times F_{i,g}) = 0.38$

We have to do this for every taxon in the regional taxa pool!

Now that we have estimates of probabilities of capture, we can estimate O/E.

Sum p_c's to estimate the number of taxa (E) that should be observed at the site based on standard sampling.

Species	P_c
1	0.70
2	0.92
3	0.86
4	0.63
5	0.51
6	0.32
7	0.07
8	0.00
E	4.01

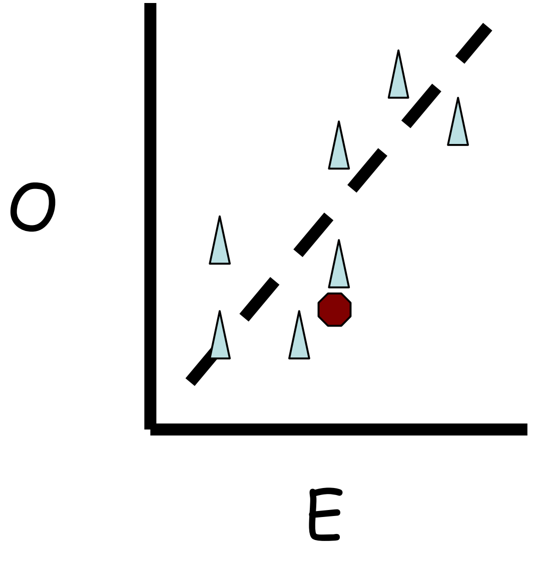
Determine O, the number of predicted taxa that were collected.

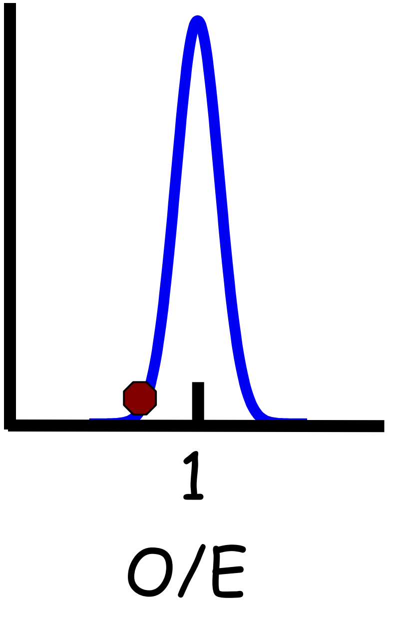
Calculate O/E.

Species	P_c	O
1	0.70	*
2	0.92	*
3	0.86	
4	0.63	
5	0.51	*
6	0.32	
7	0.07	
8	0.00	
E	4.01	3

$$O/E = 3 / 4.01 = 0.75$$

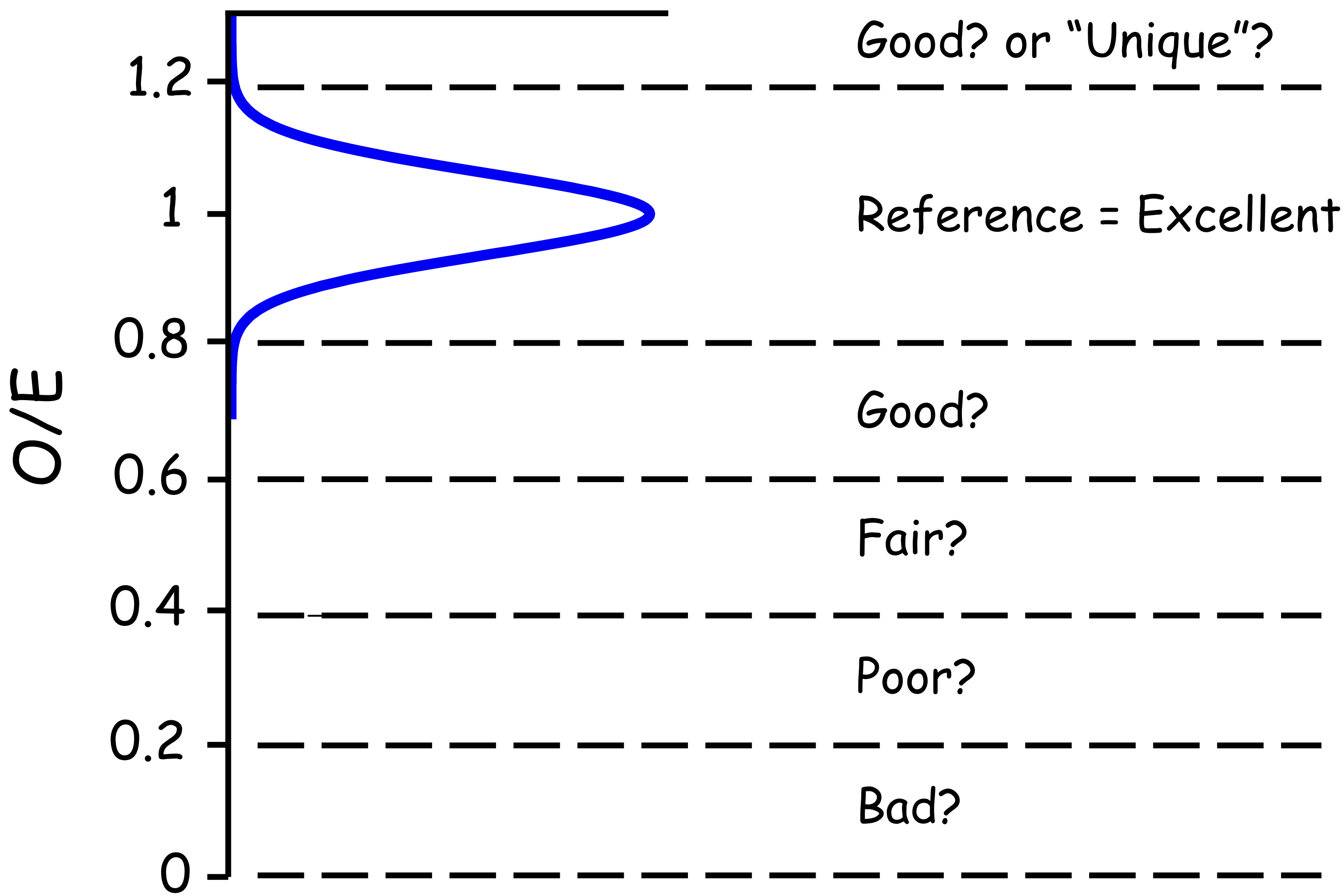
Determine if the O/E value is significantly different from the reference condition by comparing against model predictions and error.





Relating Numbers and Narratives: Some Cautionary Comments

Numeric | Narrative



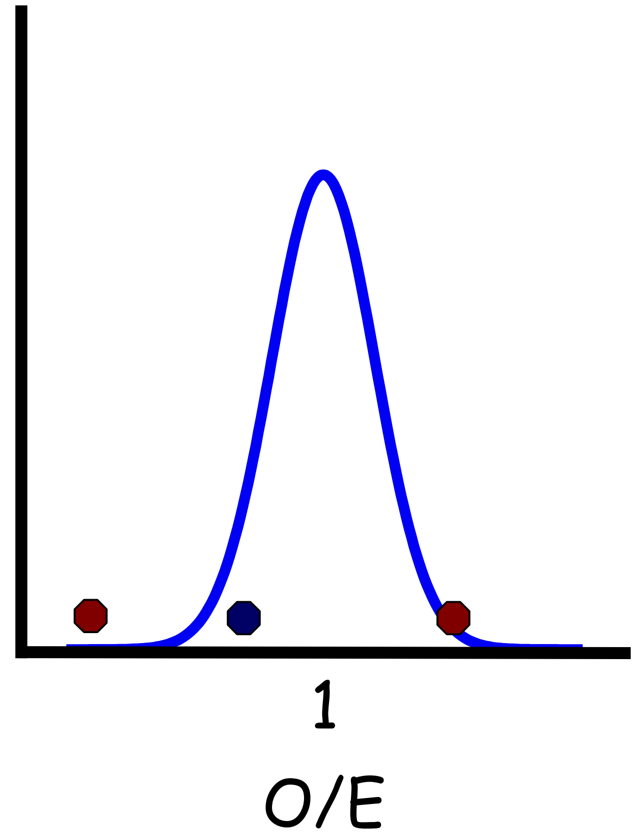
Do narrative terms convey an accurate sense of our numerical assessment of a site?

We need to think carefully about what narrative terms imply about the condition of the biota.

Statistical Issues Regarding Inferences of Impairment

Single Sites/Samples

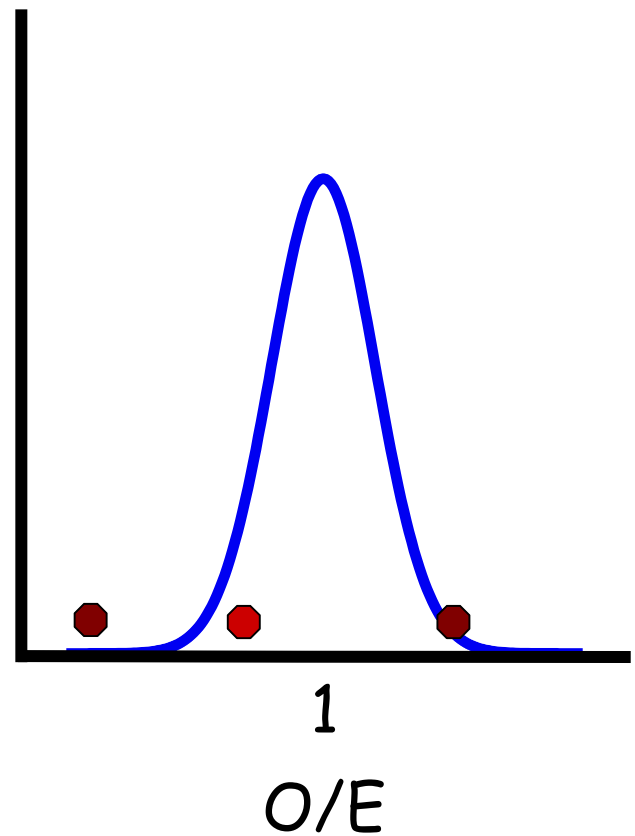
- Hypothesis: the observed O/E value is from the same distribution of values estimated for reference sites, i.e., the site is equivalent to reference.



Statistical Issues Regarding Inferences of Impairment

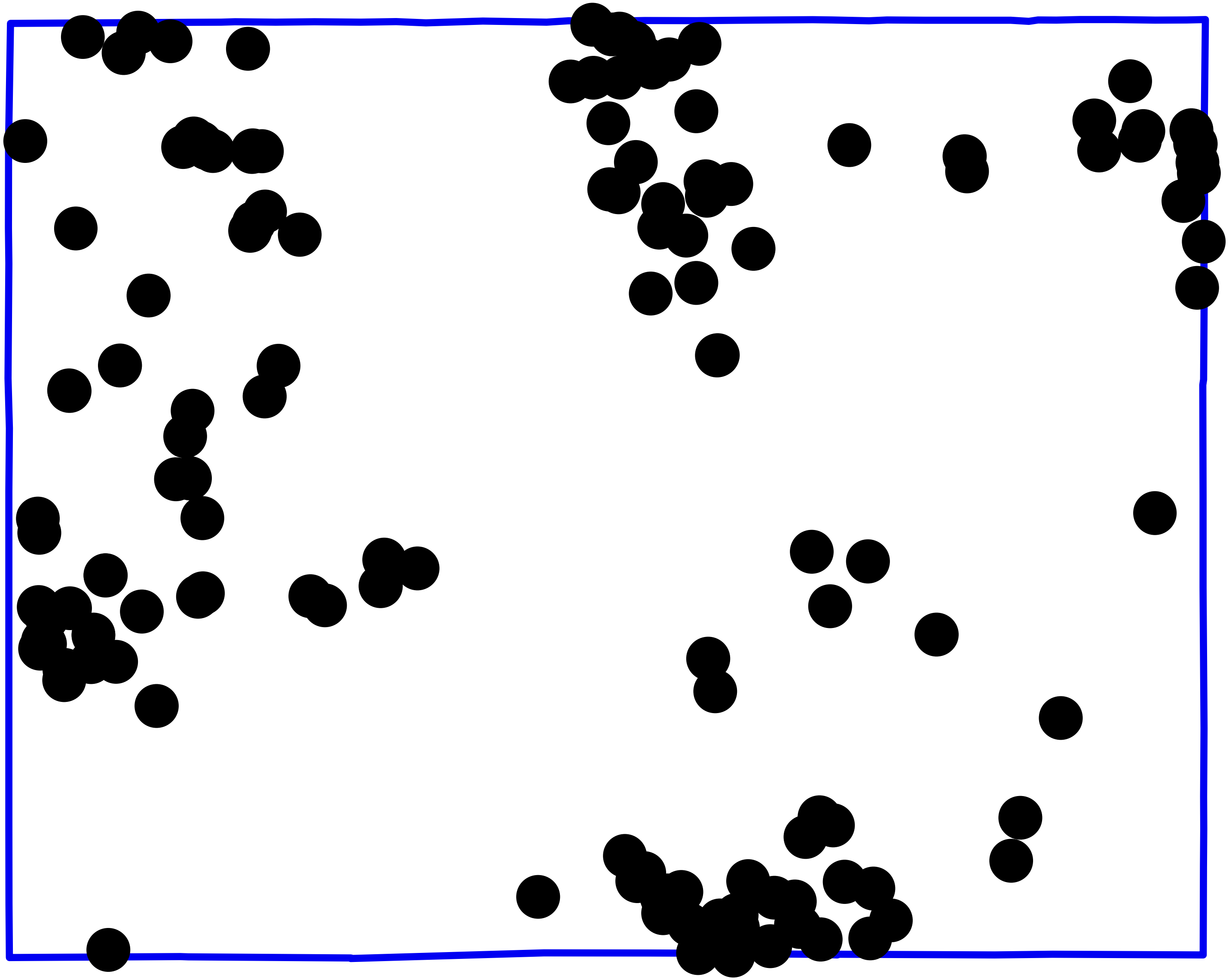
Multiple Sites or Replicated Samples at a Site

Hypothesis: the observed mean is different from 1 (the reference mean).

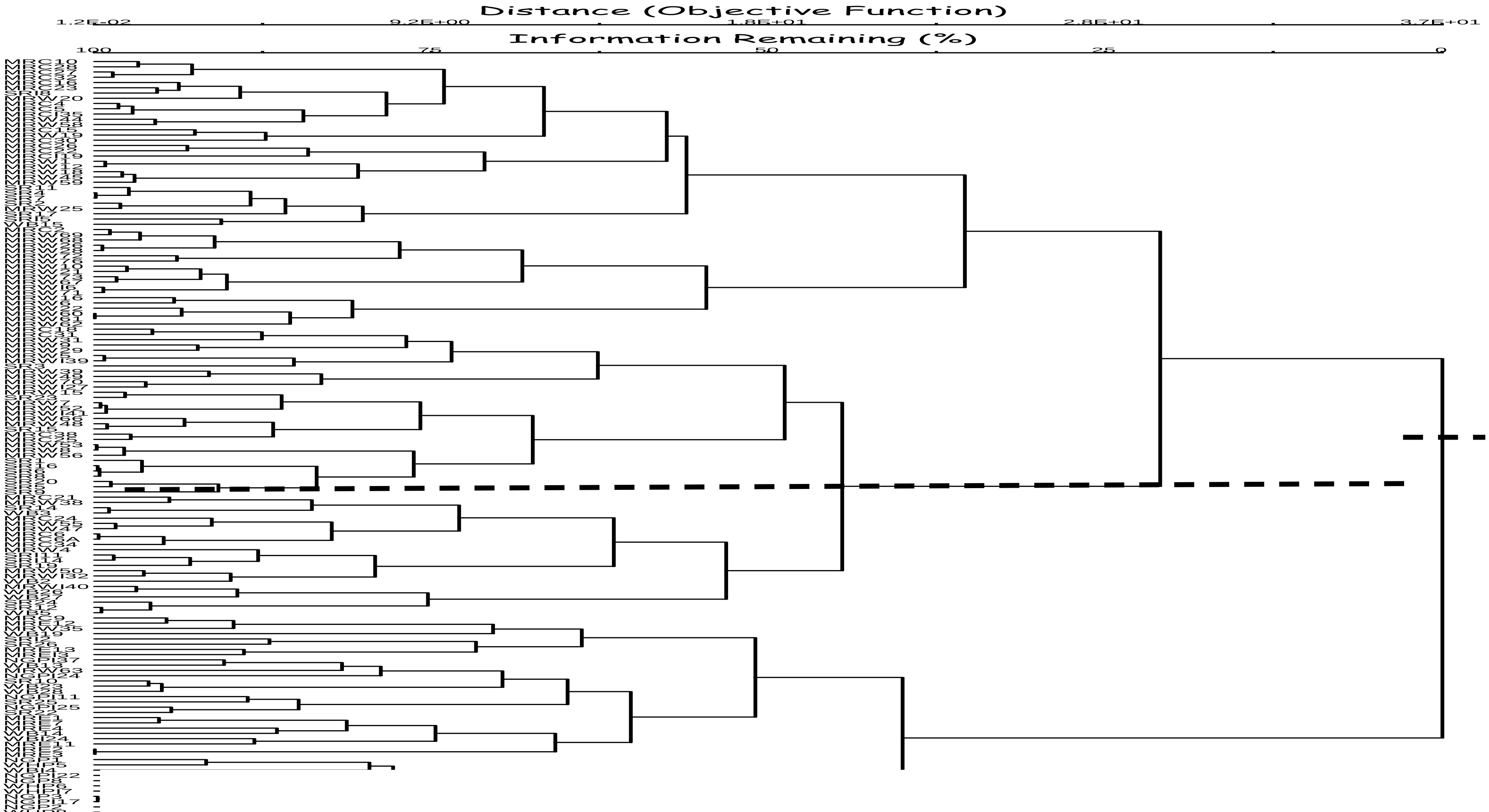


To illustrate the application of RIVPACS to real systems, we will use a case study from Wyoming

142 Reference Sites in Wyoming

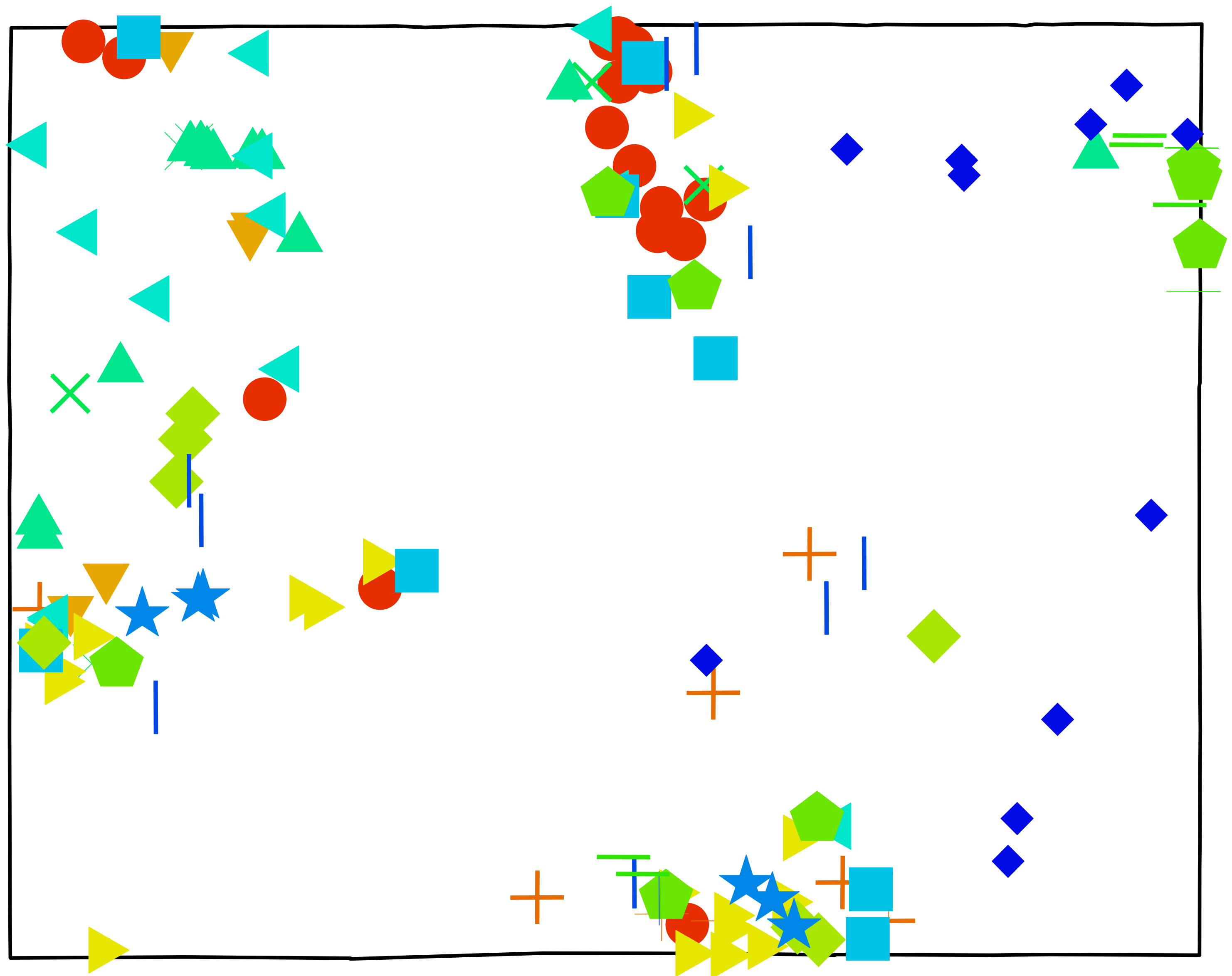


What did the dendrogram look like for the Wyoming data?



119 taxa were used to classify sites and 14 “classes” were identified.

Spatial Distribution of Reference Sites Coded by Biotic Class



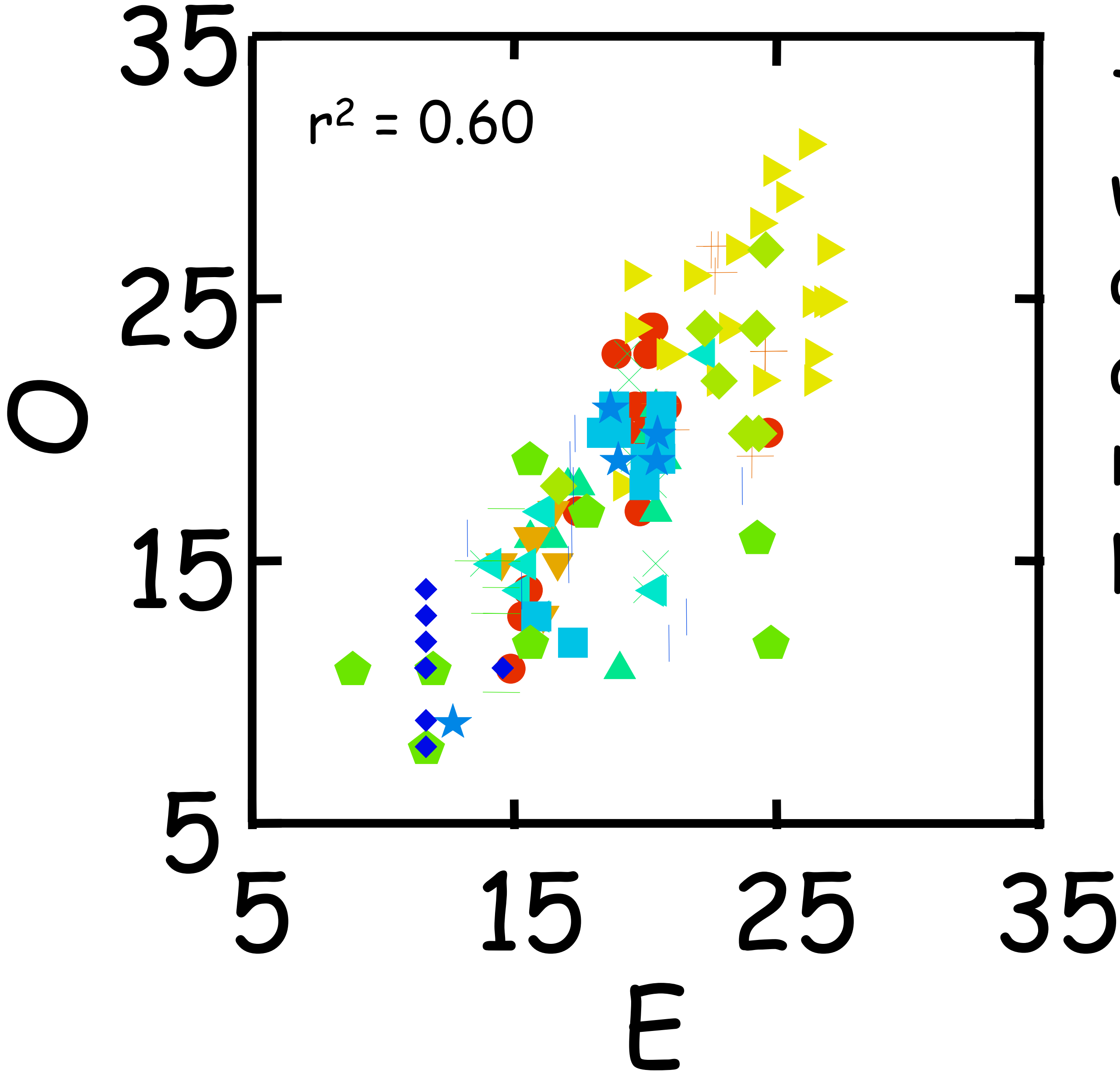
Two Discriminant Models

Continuous Variables

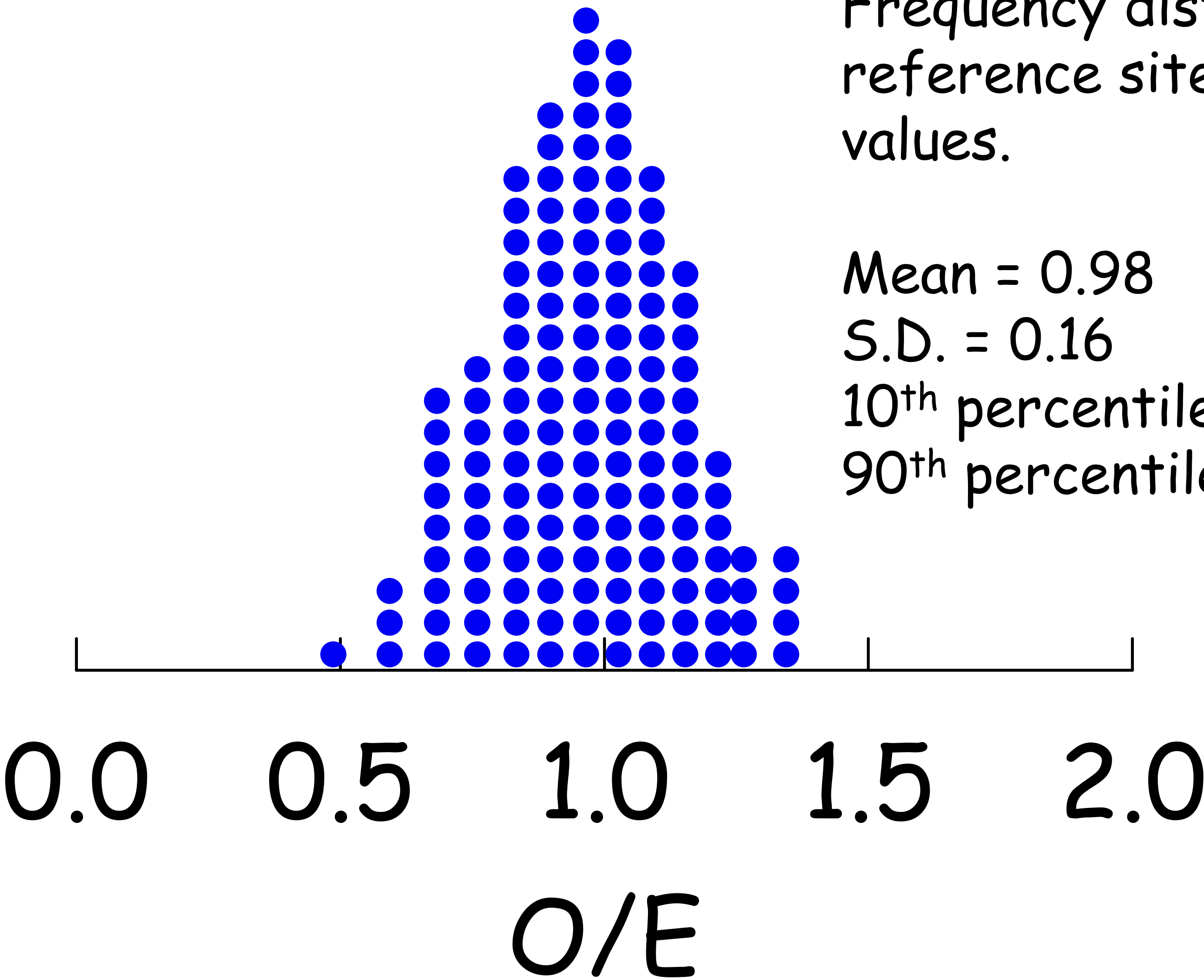
% Cobble	9.39
Log WS Area	6.54
Latitude	6.39
Longitude	5.13
Elevation	2.88
Velocity	2.60
Date	2.49
Log Alkalinity	2.33

Mixed Variables

Wyoming Basin ER	7.75
Log WS Area	5.77
Plains landscape	4.89
Mid-Rockies	4.41
Longitude	4.39
Latitude	4.26
Date	3.89
% Cobble	3.86
TWP geology	3.47
NG-Montane	3.39
Elevation	3.31
PPM geology	2.86
Velocity	2.73
MD geology	2.40
Log Alkalinity	2.17



The model was globally accurate and reasonably precise.

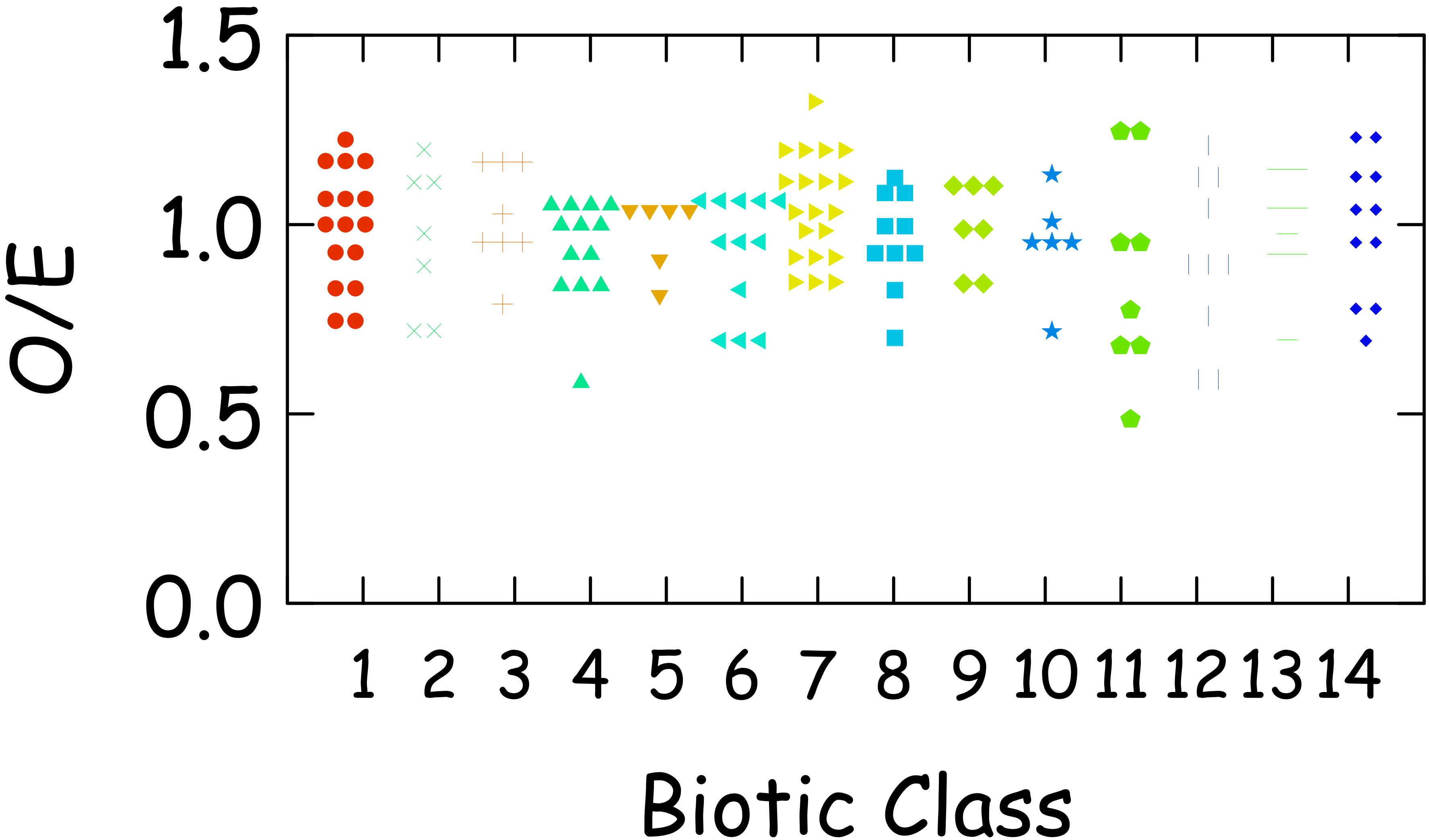


Frequency distribution of reference site O/E values.

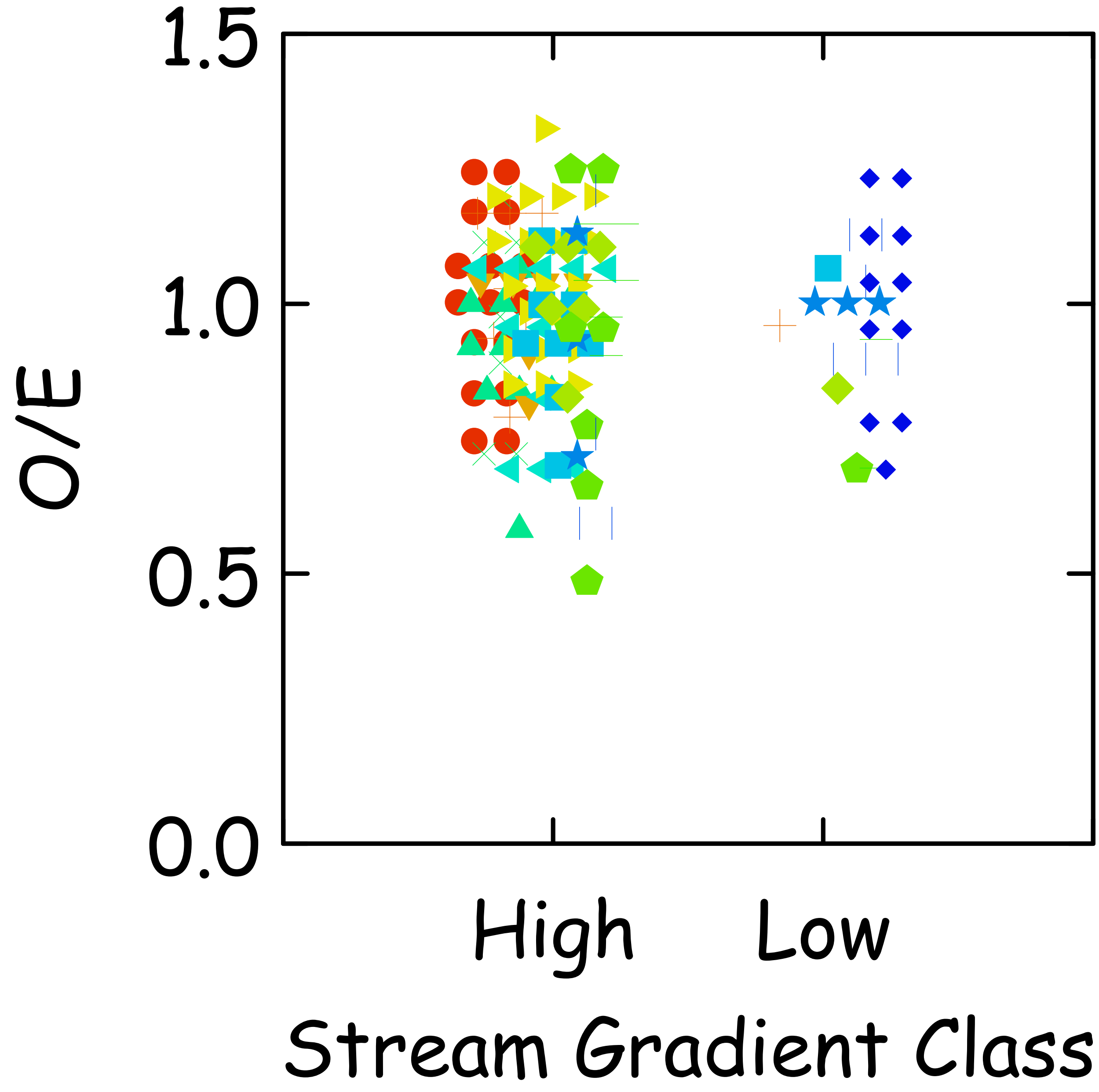
Mean = 0.98
S.D. = 0.16
10th percentile = 0.73
90th percentile = 1.19

Models can potentially be globally accurate, but locally biased, so we need to check if model predictions are biased under various local conditions.

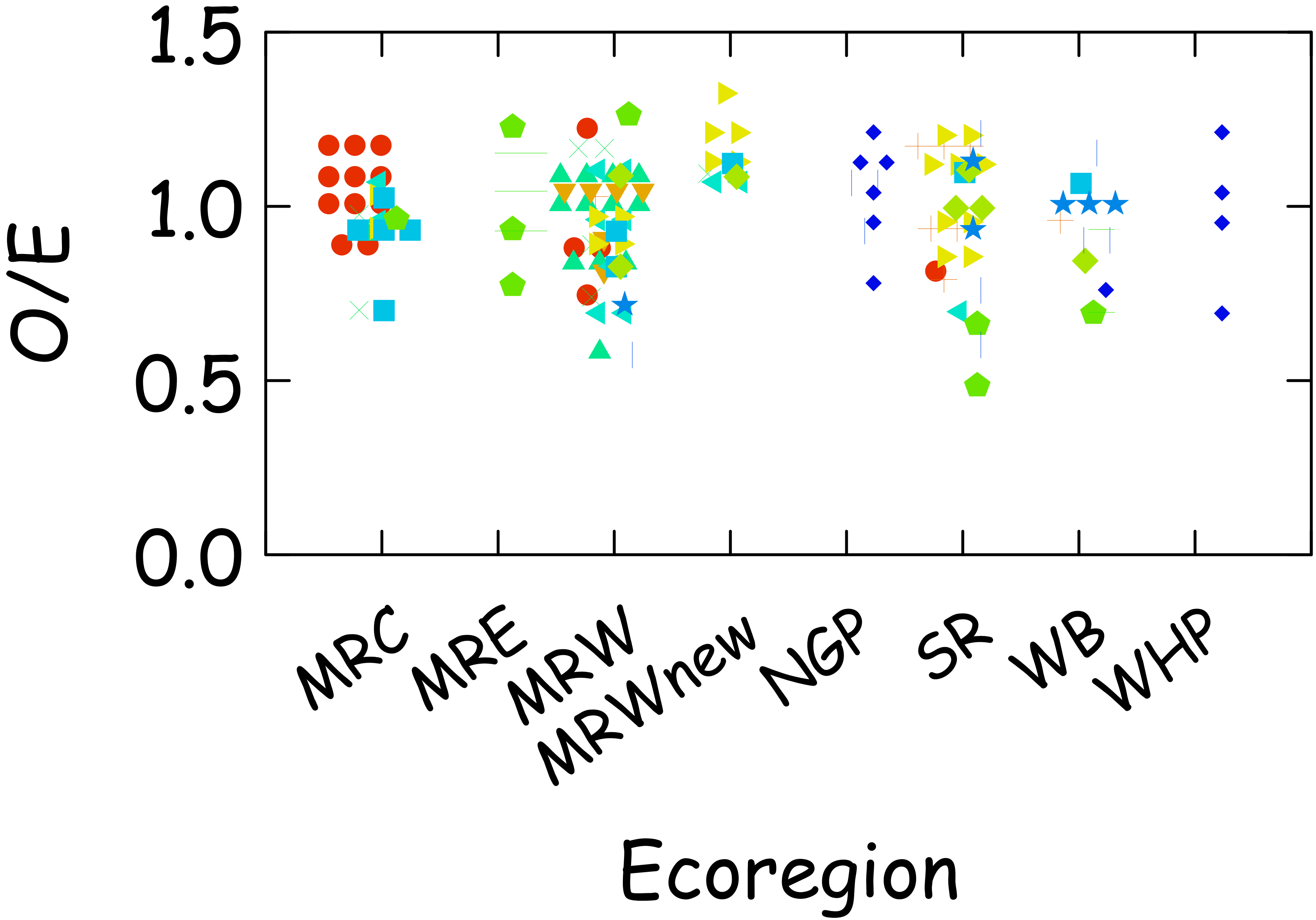
O/E values were not associated with the biotic class to which reference sites were assigned.



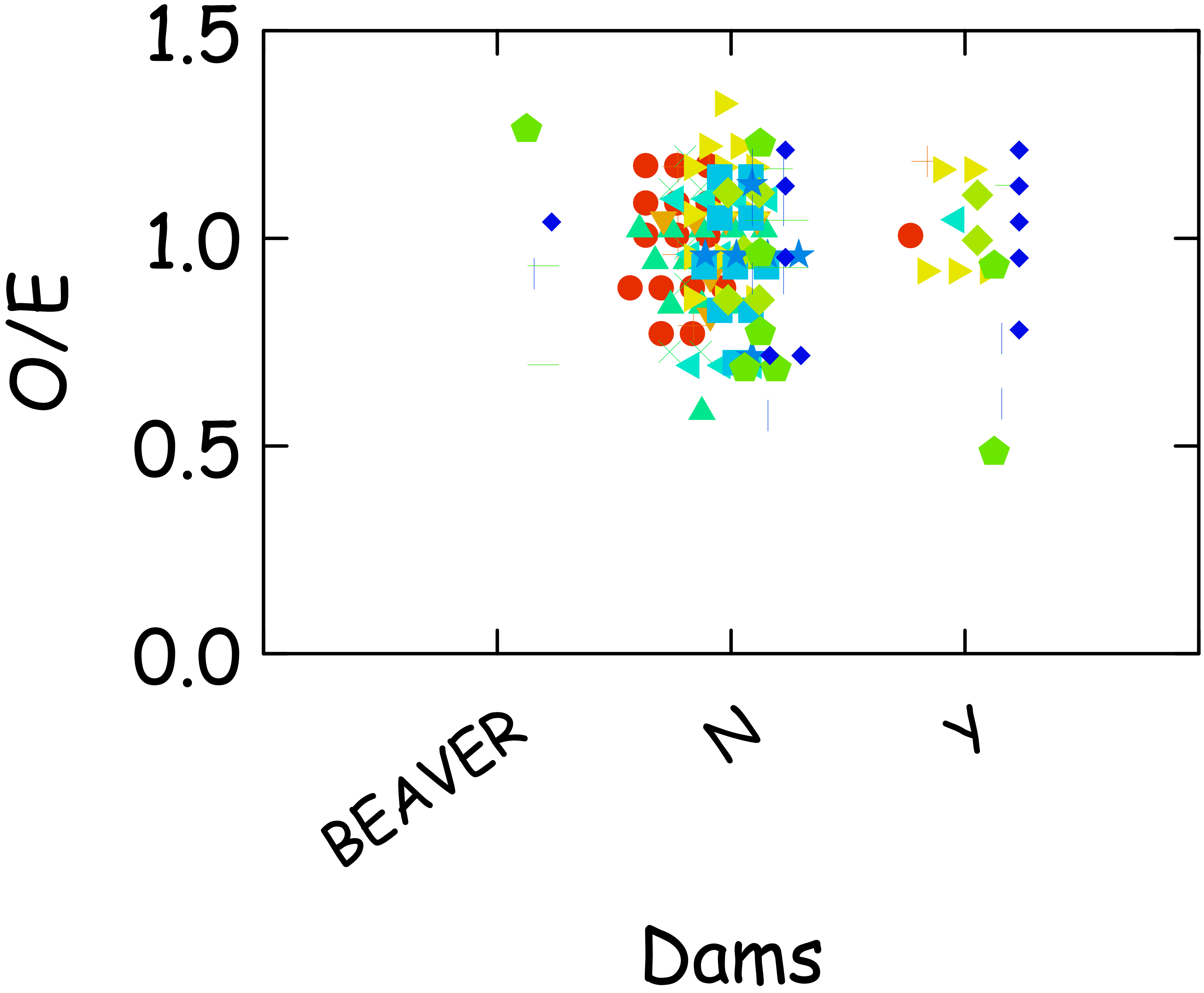
O/E values were not associated with stream gradient.



O/E values were not associated with ecoregion.



Upstream dams did not affect O/E values at reference sites.

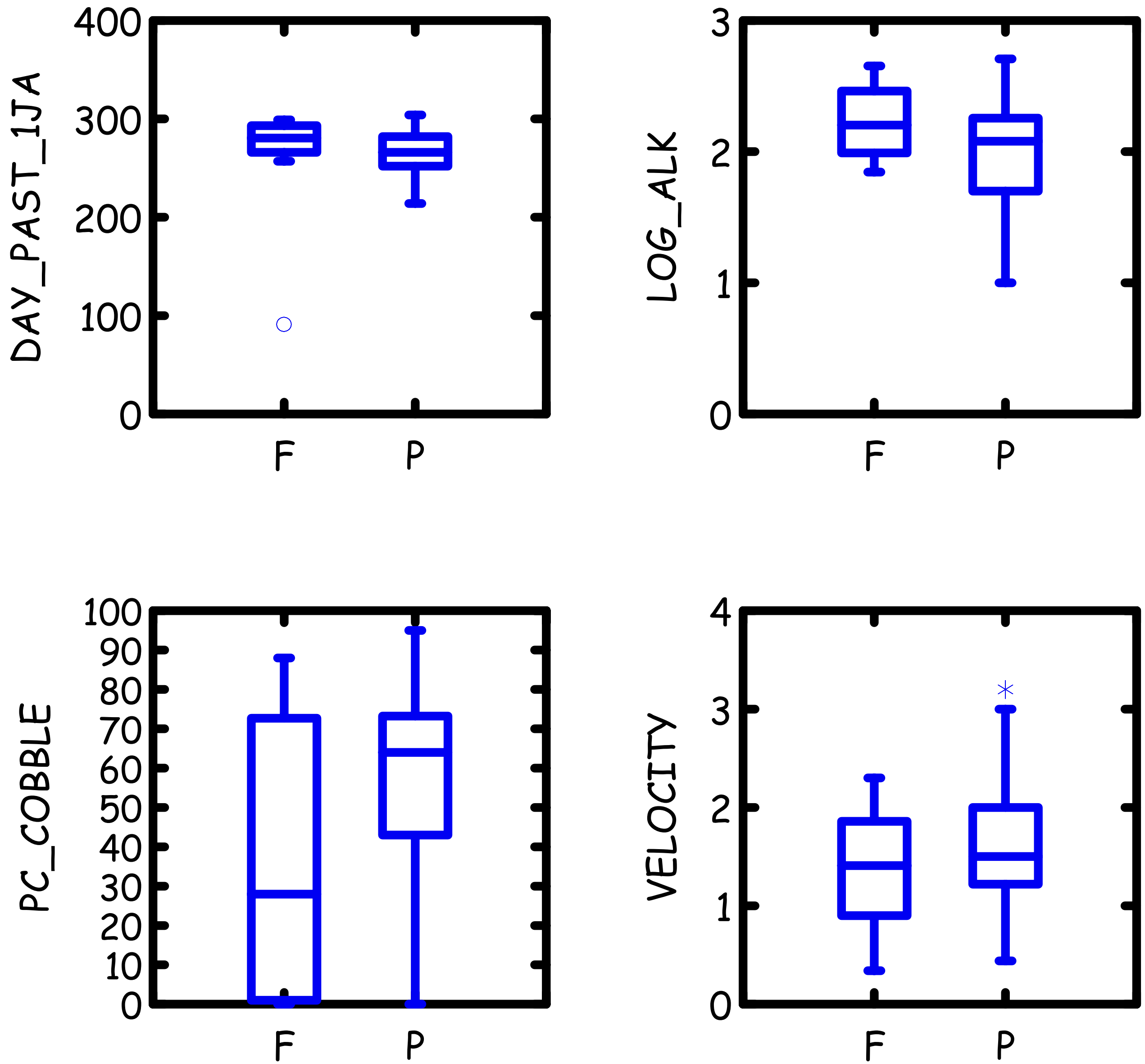


Applying the Model to Test Sites

Simple statistical tests can be applied to the predictor variables measured at a new site to determine if the model applies.

If it doesn't, the program is prevented from conducting an assessment.

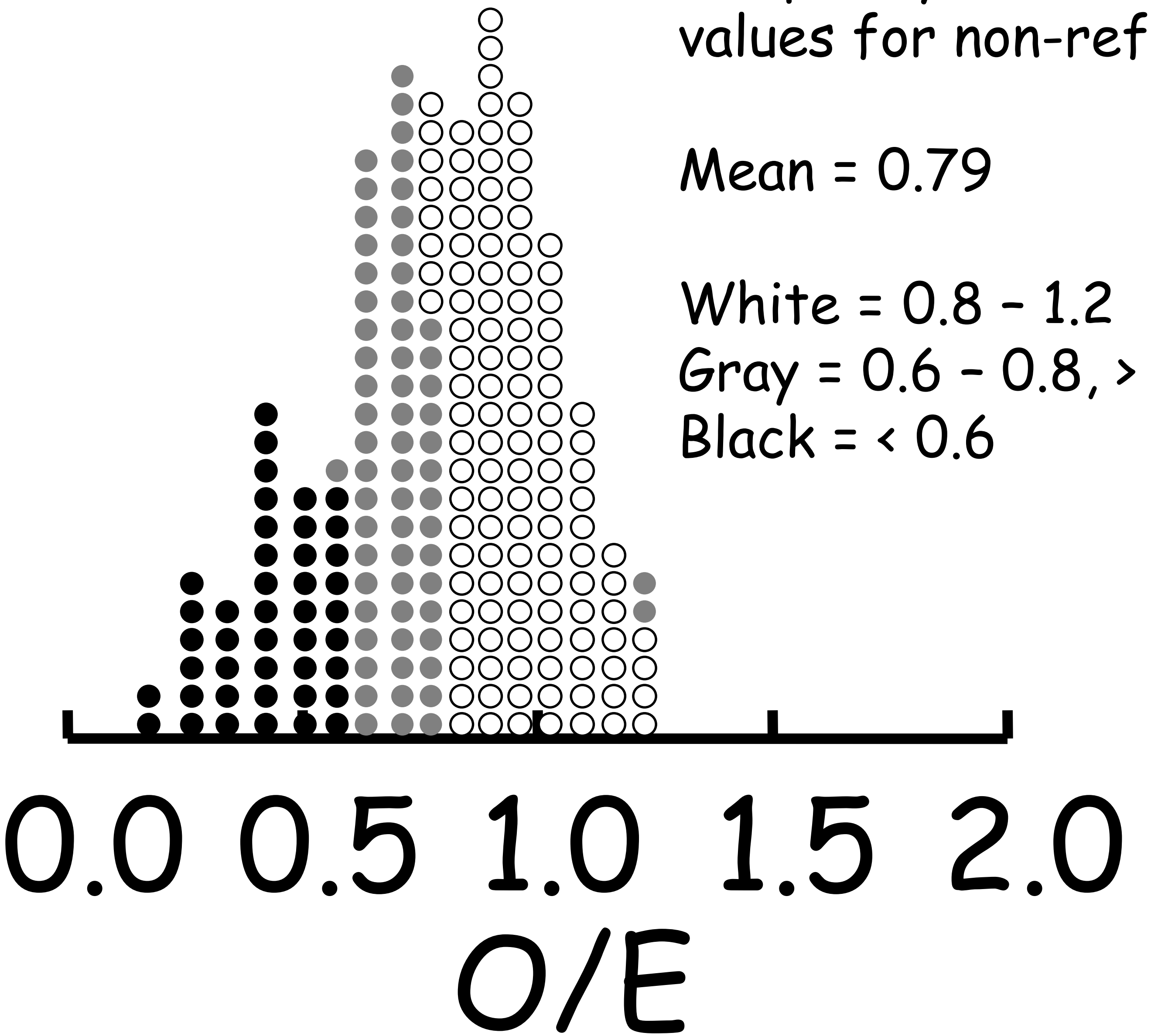
Of 241 non-reference sites, 14 (6%) were outside of the experience of the model and an assessment was not calculated.



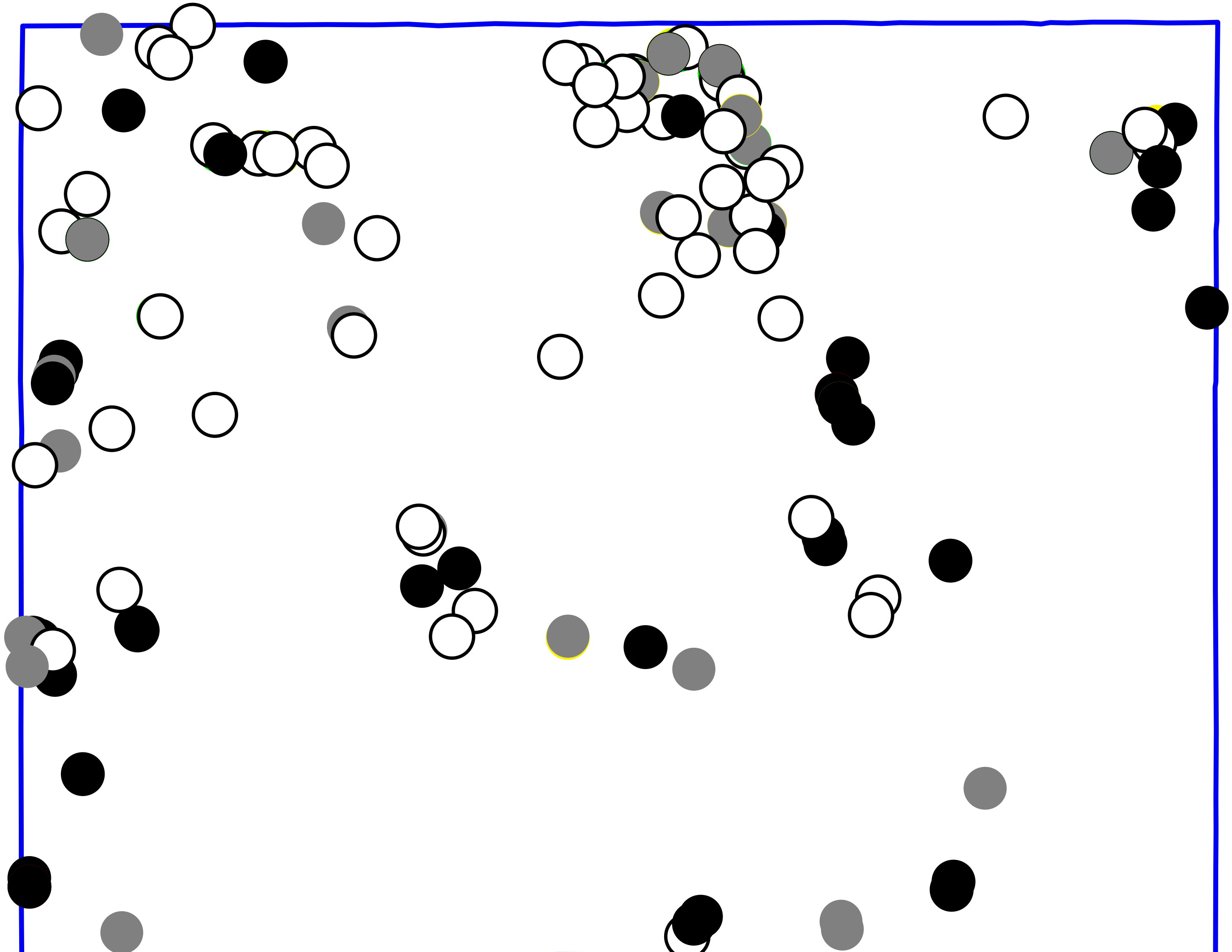
Frequency distribution of O/E values for non-reference sites.

Mean = 0.79

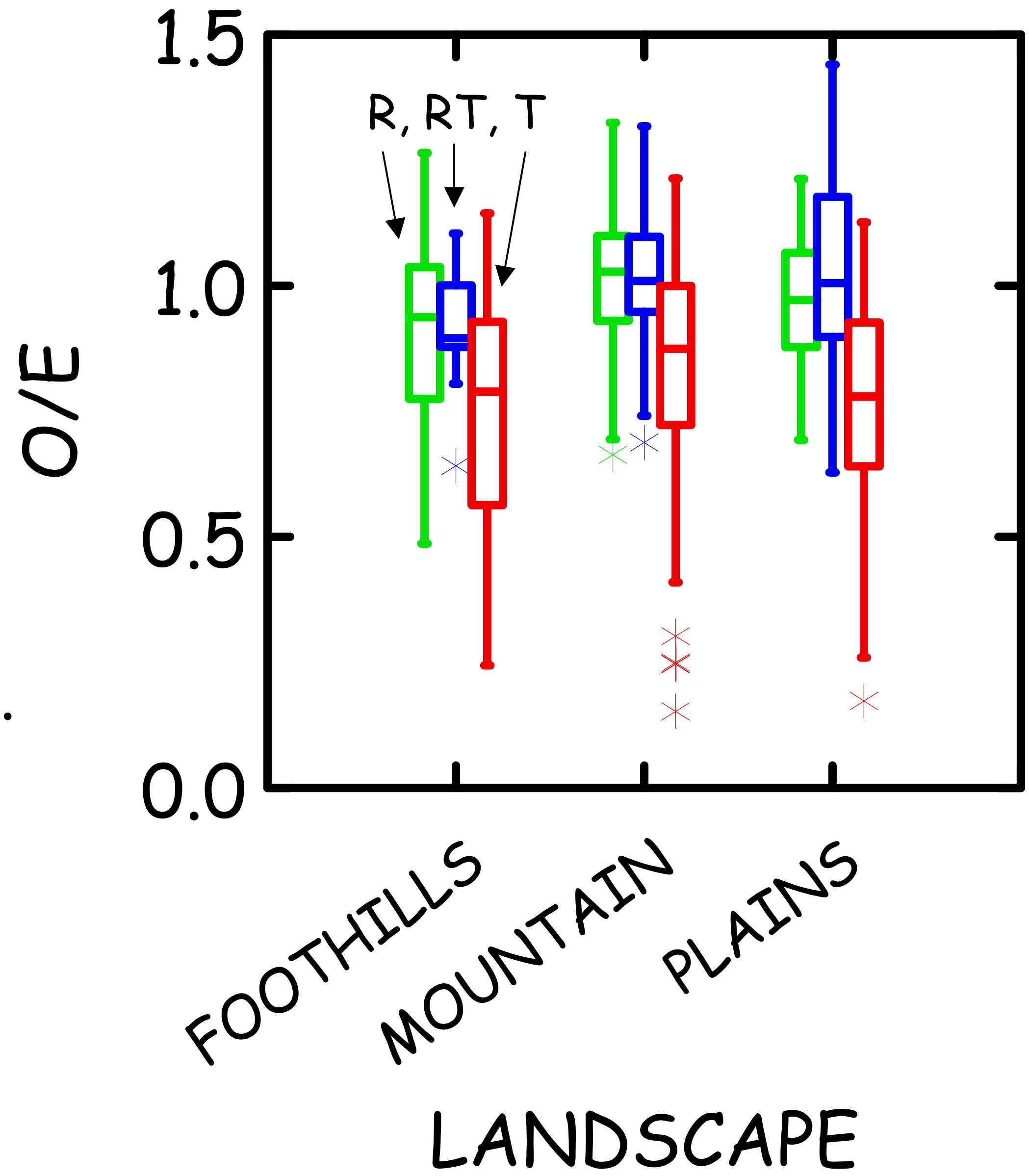
White = 0.8 – 1.2 (53%)
Gray = 0.6 – 0.8, > 1.2 (28%)
Black = < 0.6 (19%)



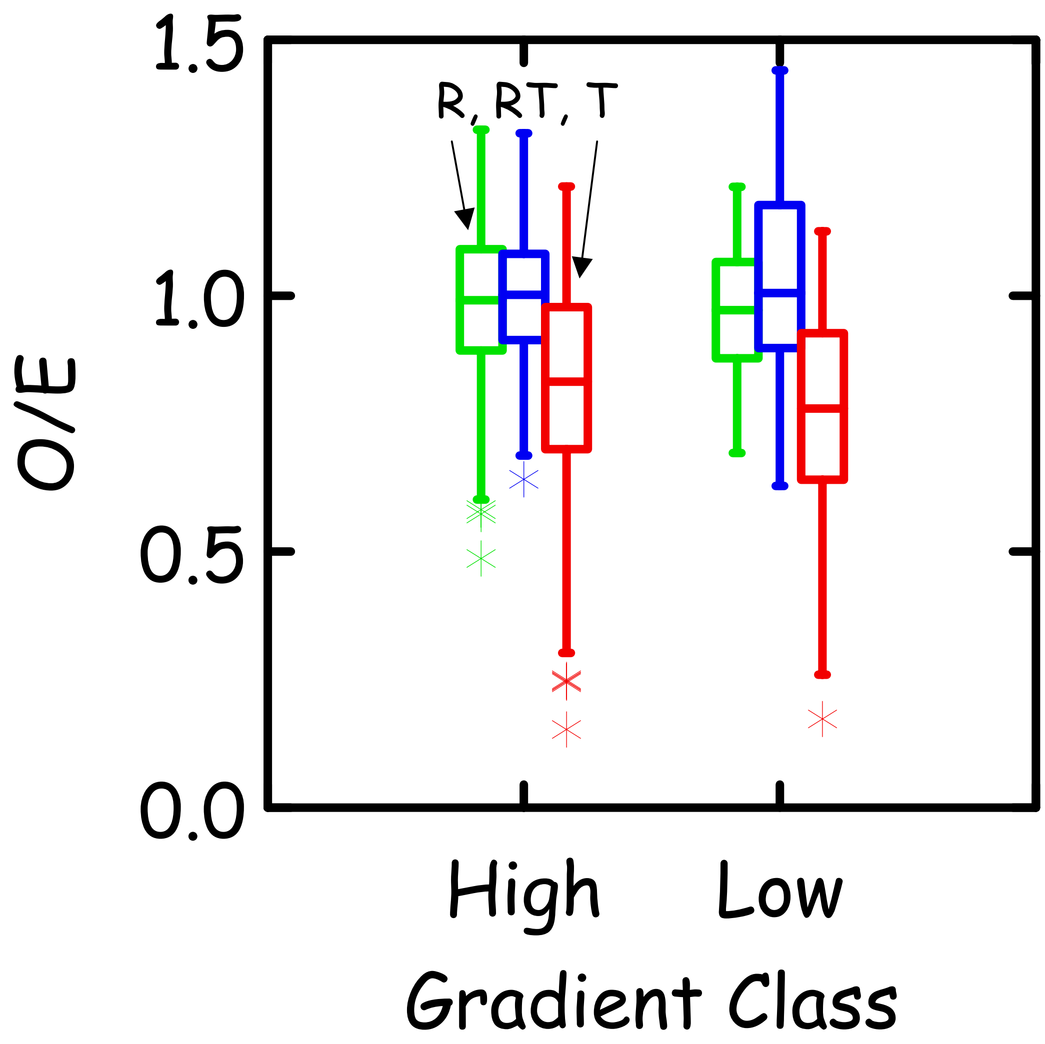
Spatial Distribution of O/E Classes for Non-Reference Sites



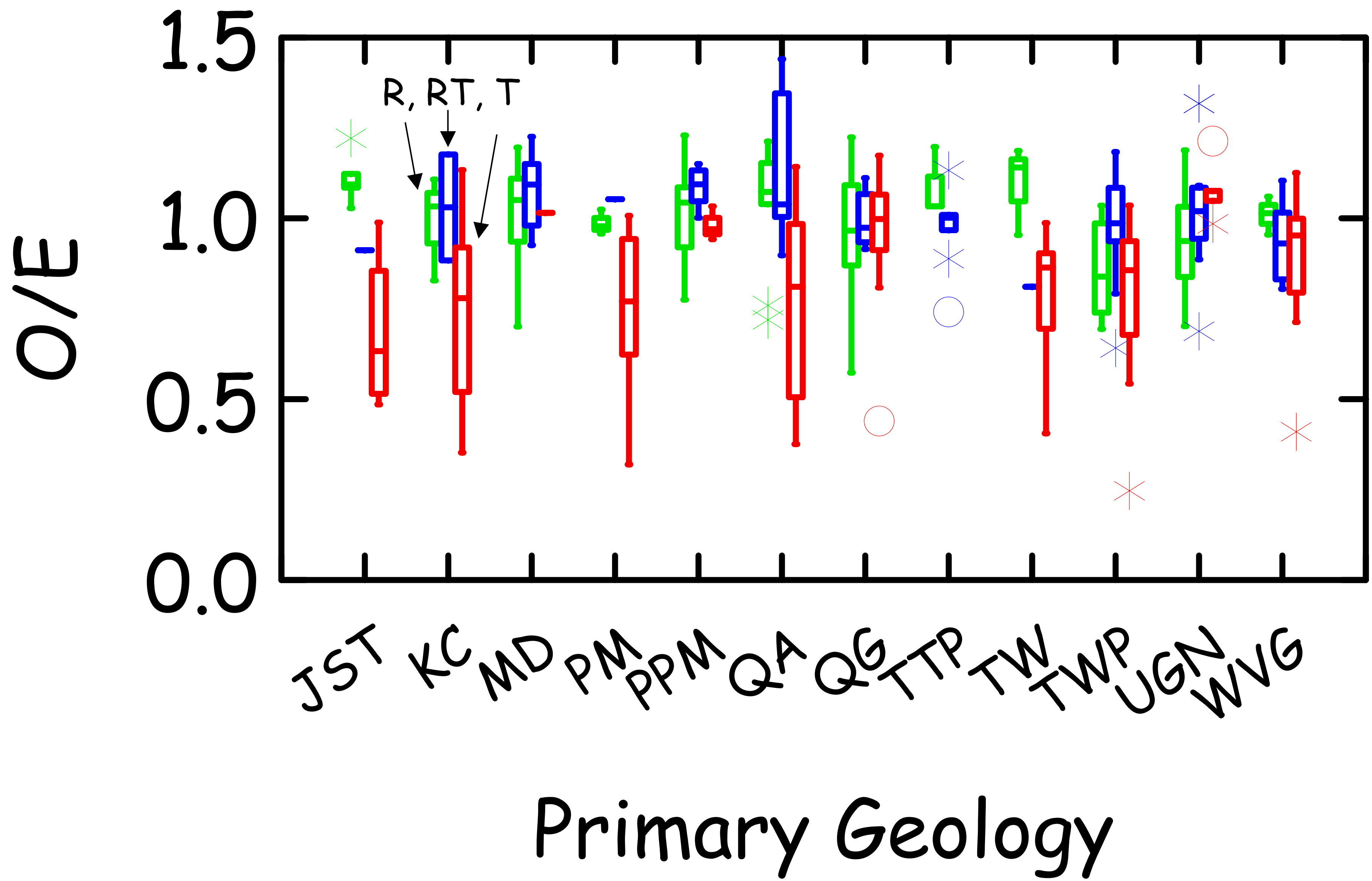
Mountain streams were slightly less impaired than streams in other landscapes.



Low-gradient test sites were no more impaired than high-gradient test sites.



The difference between reference and test site O/E values did depend on geologic setting.



Taxon Specific Responses Can be Used to Help Diagnose Causes of Impairment

From the Test Site Probability Matrix, we can see that across all of the test sites, some taxa decreased, some increased, and others showed little change.

Model outputs can also be used to identify potentially sensitive and tolerant taxa.

Sensitivity Index (SI)

$$\frac{\text{\# sites taxon was observed}}{\text{\# sites taxon was expected}}$$

SI is different than a conventional tolerance value. SI measures 'tolerance' or 'sensitivity' relative to a taxon's natural tolerance/sensitivity.

Wyoming Decreaser Taxa

TAXA	Mean PC	Expected	Observed	SI
Rhyacophila_betteni_grp	0.16	36.22	8	0.22
Deuterophlebia	0.06	13.30	3	0.23
Stempellinella	0.07	15.89	4	0.25
Wiedemannia	0.05	11.53	3	0.26
Rhyacophila_cyalinata_grp	0.08	18.25	5	0.27
Neophylax	0.05	10.98	4	0.36
Dolophilodes	0.12	26.65	10	0.38
Lepidostoma	0.30	68.28	27	0.40
Rhyacophila_pellisa	0.19	42.40	19	0.45
Zapada_columbiana	0.13	29.01	13	0.45
Ecclisomyia	0.08	19.08	9	0.47
Megarcys	0.24	55.14	28	0.51
Tanytarsus	0.07	15.72	8	0.51
Rhyacophila_coloradensis_grp	0.23	52.91	28	0.53
Neothremma	0.20	44.99	25	0.56
Parapsyche_elsis	0.28	63.90	36	0.56
Caudatella	0.05	12.36	7	0.57
Epeorus	0.51	114.76	66	0.58
Doroneuria	0.15	34.63	20	0.58
Drunella_coloradensis_flavilinea	0.33	75.64	44	0.58

Wyoming Increaser Taxa

TAXA	Mean PC	Expected	Observed	SI
Pseudochironomus	0.01	1.53	9	5.88
Nais_variabilis	0.01	3.13	18	5.76
Cryptochironomus	0.02	4.59	21	4.57
Hesperophylax	0.03	6.05	20	3.31
Paratanytarsus	0.01	3.06	10	3.27
Prodiamesa	0.01	3.13	9	2.88
Phaenopsectra	0.02	4.60	12	2.61
Pseudodiamesa	0.02	3.84	10	2.61
Planorbidae	0.02	4.82	12	2.49
Stenonema	0.02	5.26	13	2.47
Hydrobaenus	0.08	18.21	44	2.42
Hydrophilidae	0.03	7.10	16	2.25
Hemerodromia	0.06	13.98	31	2.22
Ceratopogonidae	0.05	10.91	23	2.11
Parametriocnemus	0.04	10.02	21	2.09
Microtendipes	0.06	14.03	28	2.00

It's time for questions and some exercises!