FW 599 Special Topics: Multivariate Analysis of Ecological Data in R

Lecture 11: Statistical Inference – Part 2

Tuesday, November 5, 2024

Lecture 11: Statistical Inference

- Analysis of Similarity (ANOSIM)

- Multivariate Analysis of Variance (MANOVA) and Permutational MANOVA (PERMANOVA)

- Similarity Percentage (SIMPER)

- Mantel Test

- Procrustes Analysis

- Nested Study Design

Recap: Permutation

Indirect Comparison: Permutation

A **permutation test** is a **non-parametric** method to assess the statistical significance of a test statistic by shuffling data and recalculating the statistic for each new arrangement.

- Instead of relying on assumptions of normality, uses the **distribution of the test statistic under the null hypothesis** generated by permuting the data.
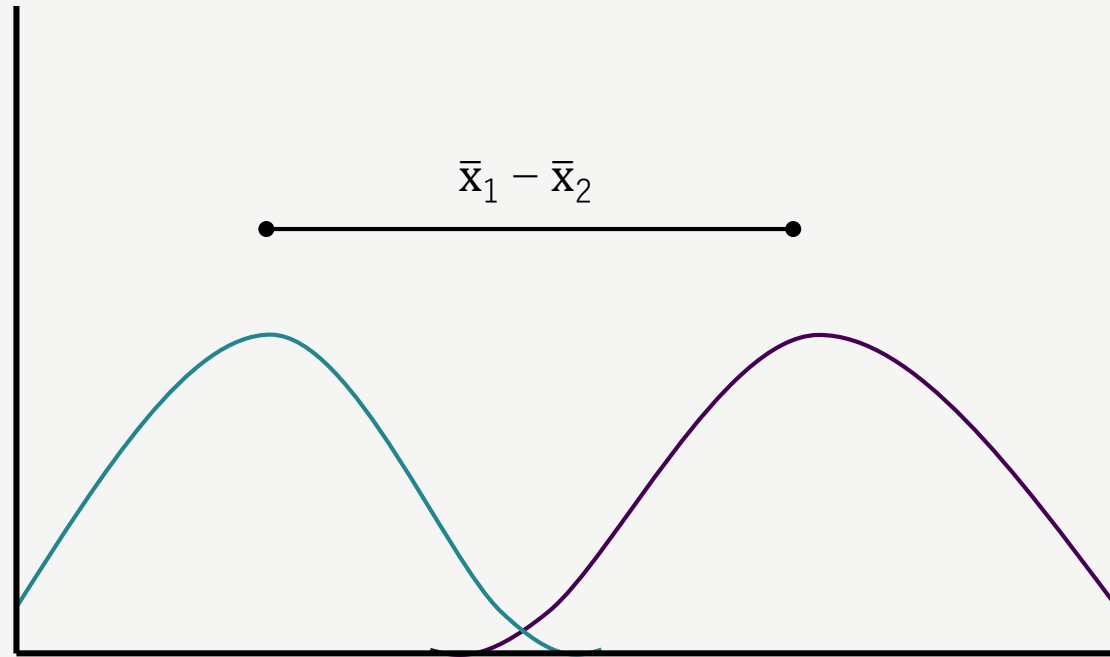
Indirect Comparison: Permutation

A **permutation test** is a **non-parametric** method to assess the statistical significance of a test statistic by shuffling data and recalculating the statistic for each new arrangement.

- Answers the question: "Is the observed result likely under random conditions?"

Indirect Comparison: Permutation

Step 1) Calculate the **test statistic** of interest for the original dataset (e.g., correlation coefficient, difference in group means)

Indirect Comparison: Permutation

Step 2) **Permute** or randomize the data

- Randomly shuffle a vector of continuous variables

- Randomly assign values to classes/groups

Indirect Comparison: Permutation
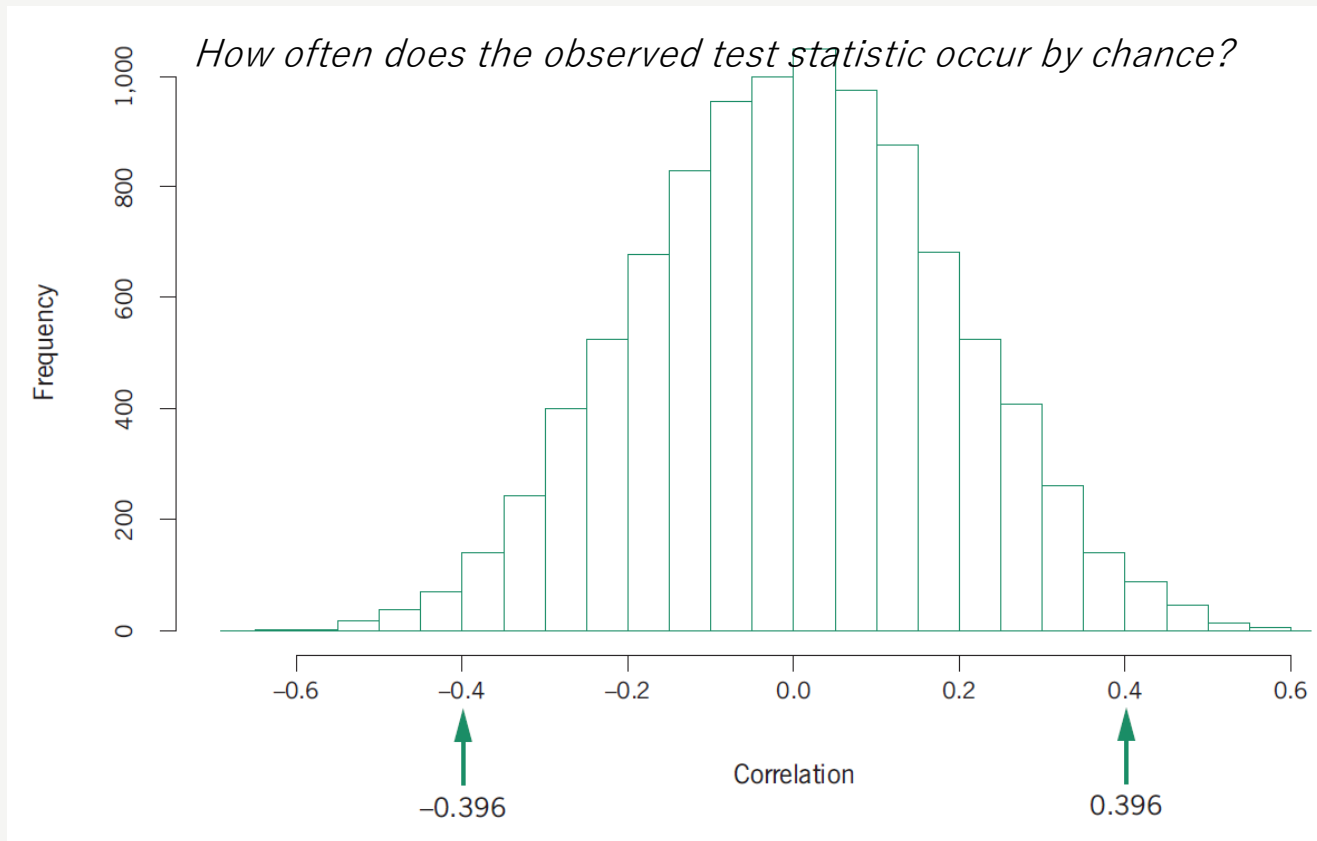
Step 2) **Permute** or randomize the data

- Randomly shuffle a vector of continuous variables

- Randomly assign values to classes/groups

Step 3) For **each** permutation, recalculate the test statistic

Indirect Comparison: Permutation

Step 4) Generate the permutation distribution

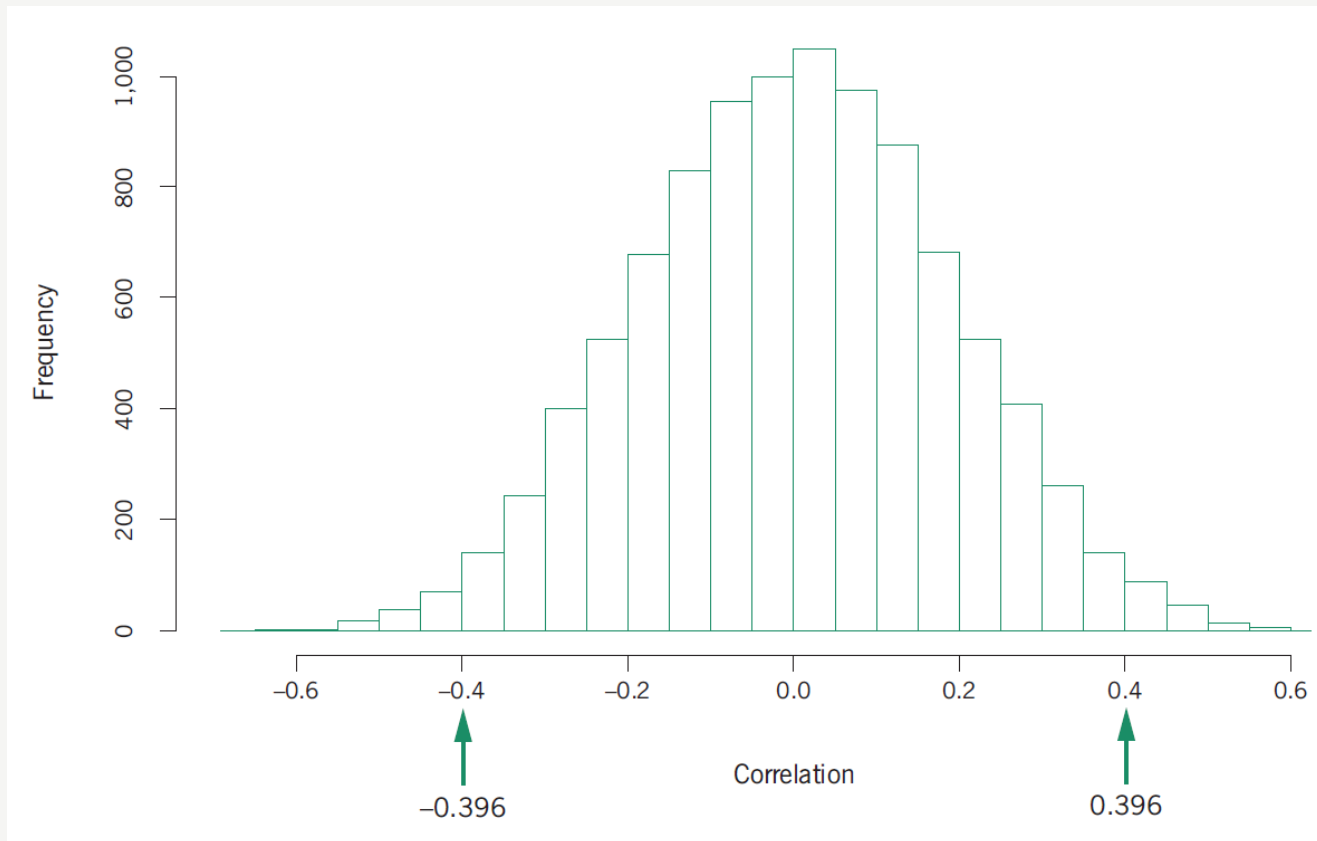*How often does the observed test statistic occur by chance?*

Greenacre & Primicerio 17.3

Indirect Comparison: Permutation

Step 5) Compare the observed test statistic to the permutation distribution to calculate a *P*-value

Greenacre & Primicerio 17.3

Indirect Comparison: Permutation

Advantages:
- Non-parametric: no assumptions about distribution of data
- Flexible across many test statistics
- Can be used for small sample sizes

Disadvantage:
- May be computationally intensive
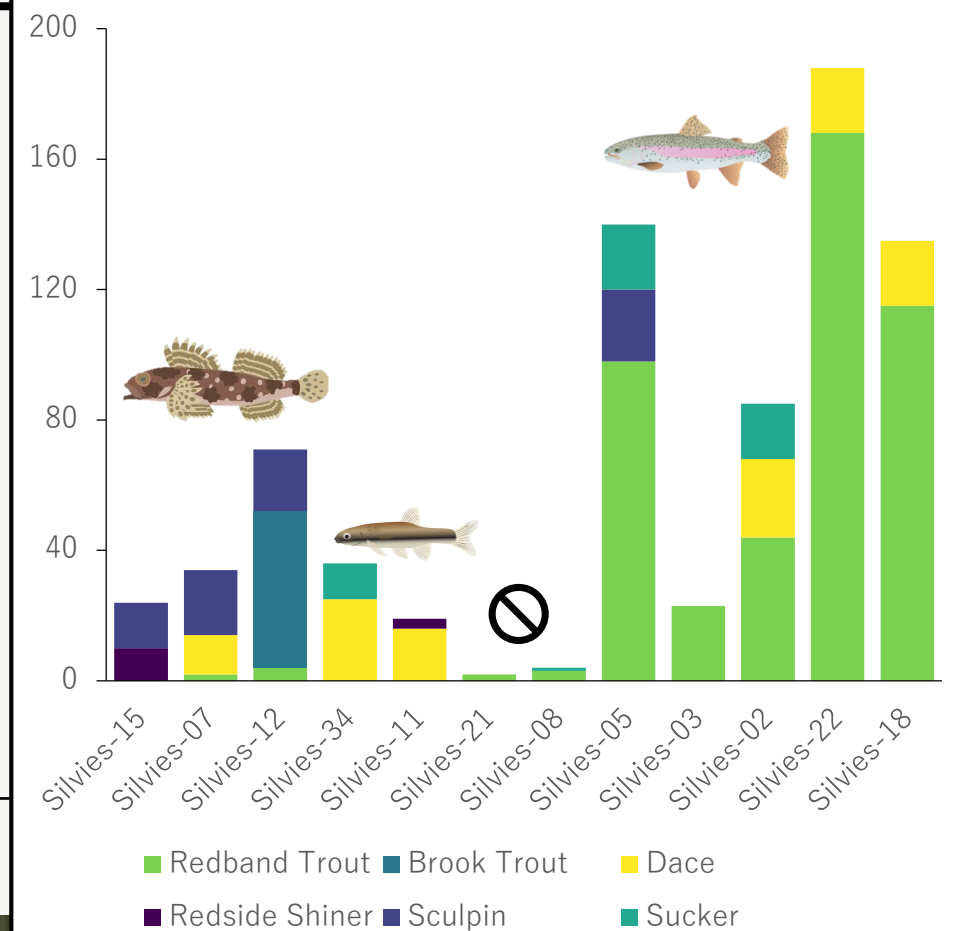
Analysis of Similarity

Analysis of Similarity

Analysis of Similarity (ANOSIM) is a non-parametric statistical test used to compare the similarity of multivariate data among pre-determined groups.

Analysis of Similarity

Site ID	Redband Trout	Brook Trout	Dace	Redside Shiner	Sculpin	Sucker
Silvies-15	0	0	0	10	14	0
Silvies-07	2	0	12	0	20	0
Silvies-12	4	48	0	0	19	0
Silvies-34	0	0	25	0	0	11
Silvies-11	0	0	16	3	0	0
Silvies-21	2	0	0	0	0	0
Silvies-08	3	0	0	0	0	1
Silvies-05	98	0	0	0	22	20
Silvies-03	23	0	0	0	0	0
Silvies-02	44	0	24	0	0	17
Silvies-22	168	0	20	0	0	0
Silvies-18	115	0	20	0	0	0

Analysis of Similarity

Step 1) Calculate the distance matrix

	S-15	S-07	S-12	S-34	S-11	S-21	S-08	S-05	S-03	S-02	S-22
Silvies-07	0.52										
Silvies-12	0.73	0.72									
Silvies-34	1.00	0.62	1.00								
Silvies-11	0.88	0.55	1.00	0.54							
Silvies-21	1.00	0.87	0.88	1.00	1.00						
Silvies-08	1.00	0.88	0.89	0.93	1.00	0.33					
Silvies-05	0.75	0.72	0.76	0.80	1.00	0.90	0.81				
Silvies-03	1.00	0.91	0.93	1.00	1.00	0.80	0.73	0.36			
Silvies-02	1.00	0.64	0.95	0.50	0.58	0.86	0.74	0.42	0.41		
Silvies-22	1.00	0.76	0.96	0.81	0.75	0.90	0.86	0.38	0.38	0.41	
Silvies-18	1.00	0.73	0.95	0.79	0.70	0.88	0.83	0.31	0.29	0.34	0.10

Analysis of Similarity

Step 1) Calculate the distance matrix

	S-15	S-07	S-12	S-34	S-11	S-21	S-08	S-05	S-03	S-02	S-22
Silvies-07	11										
Silvies-12	19	18									
Silvies-34	36	15	36								
Silvies-11	29	13	36	12							
Silvies-21	36	28	29	36	36						
Silvies-08	36	29	30	33	36	4					
Silvies-05	21	18	22	24	36	31	25				
Silvies-03	36	32	33	36	36	24	19	6			
Silvies-02	36	16	34	10	14	27	20	9	8		
Silvies-22	36	22	35	25	21	31	27	7	7	8	
Silvies-18	36	19	34	23	17	29	26	3	2	5	1

Analysis of Similarity

Step 2) Define Groups

Site ID	Max Depth (m)	Gradient (%)	Elevation (m)	Canopy (%)	Herb (%)
Silvies-11	0.45	0.3	1439	0.0	55.1
Silvies-34	0.78	1.1	1487	0.0	0.0
Silvies-02	0.71	0.4	1372	29.6	0.0
Silvies-15	0.40	0.2	1471	41.1	0.0
Silvies-07	0.50	1.3	1547	52.3	0.0
Silvies-08	0.40	0.6	1492	51.4	0.0
Silvies-22	0.42	0.9	1555	54.7	0.0
Silvies-18	0.42	0.5	1510	46.2	0.0
Silvies-12	0.52	3.2	1658	51.9	0.0
Silvies-21	0.18	2.4	1713	37.5	0.0
Silvies-05	0.45	5.5	1565	46.7	0.0
Silvies-03	0.20	3.3	1634	59.0	0.0

Analysis of Similarity

Step 3) Calculate the R statistic

"**R**" compares the mean of *ranked* dissimilarities between groups to the mean of *ranked* dissimilarities within groups

PCA of Hellinger-Transformed Fish Data
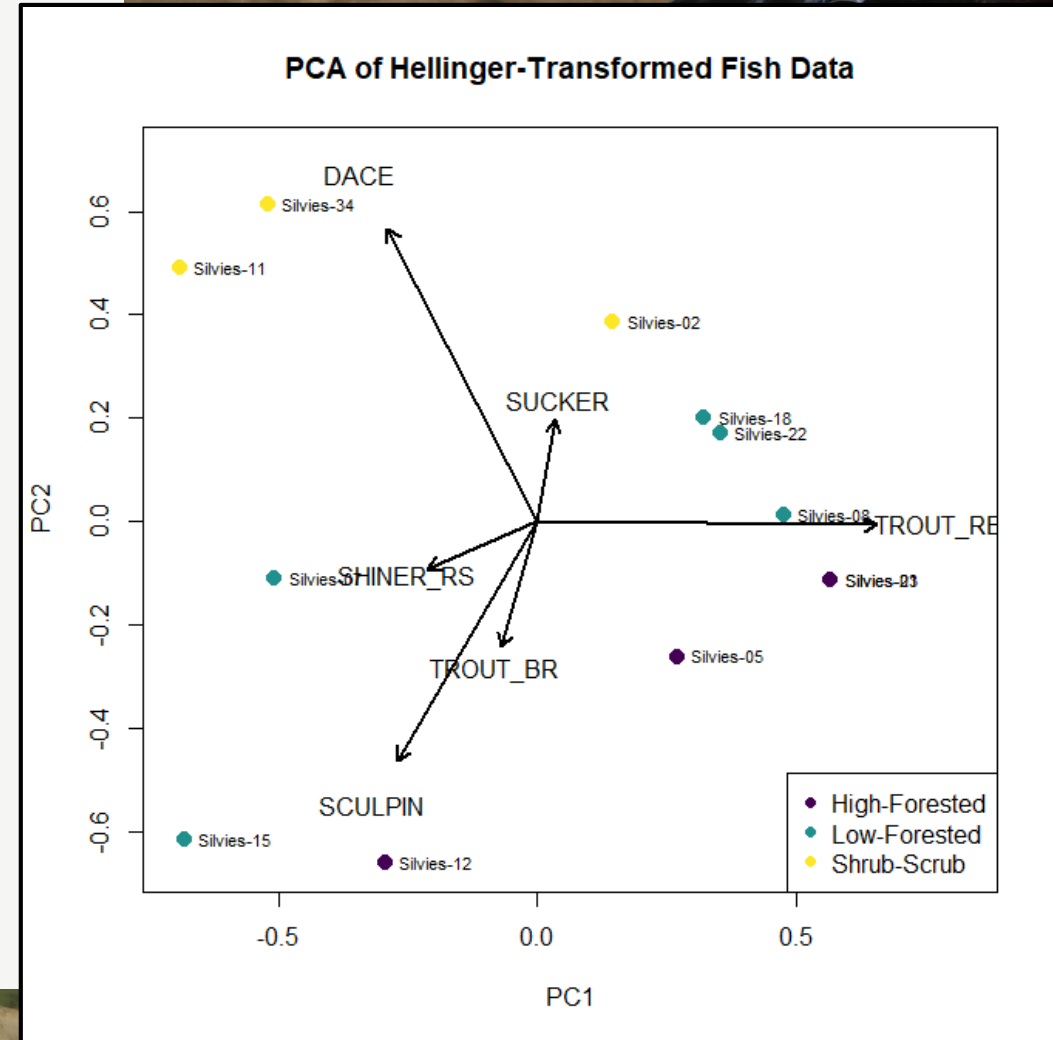
Analysis of Similarity

Step 3) Calculate the R statistic

"**R**" compares the mean of *ranked* dissimilarities between groups to the mean of *ranked* dissimilarities within groups
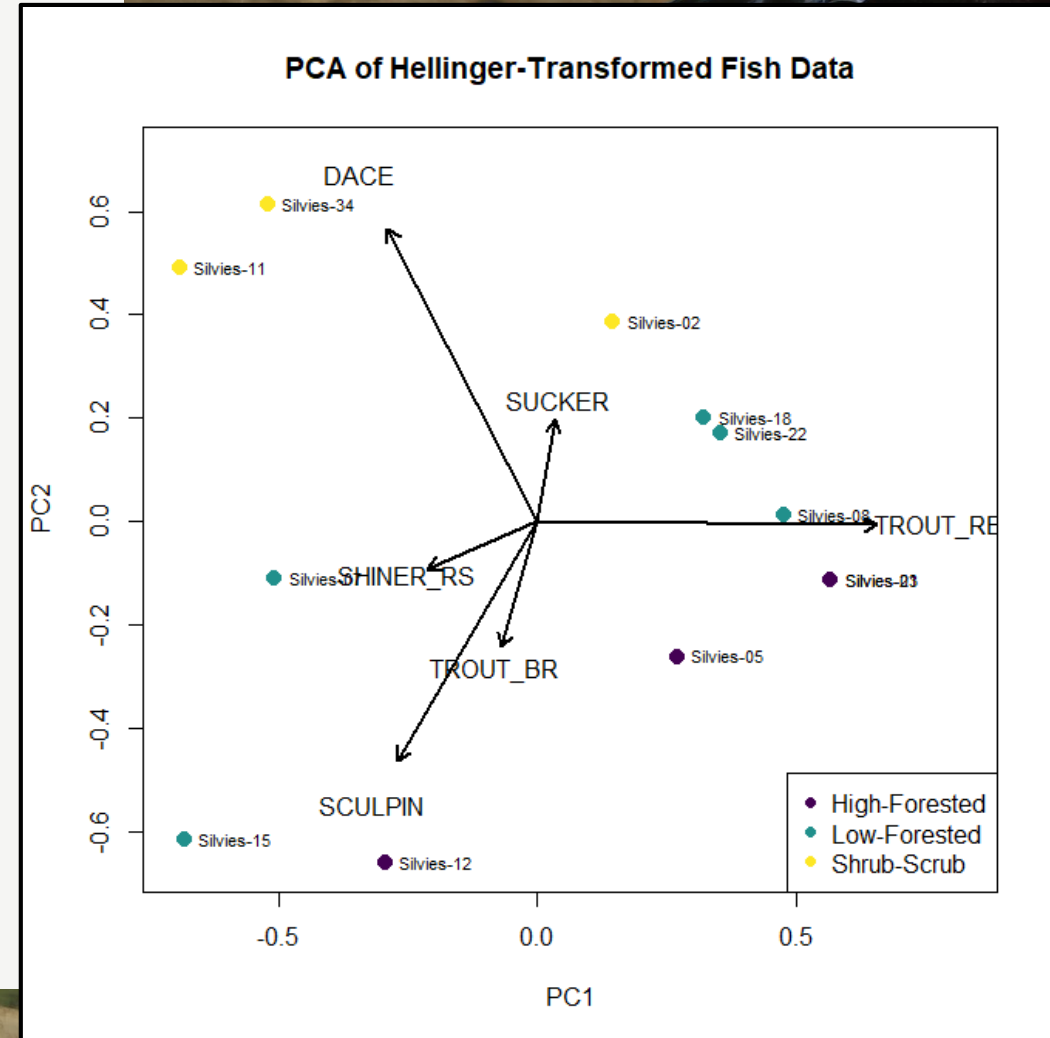
- **R = 1:** all within-group dissimilarities are less than all between-group dissimilarities.
- **R = 0:** no difference in dissimilarity.
- **R < 0:** more within-group than between-group dissimilarities.

PCA of Hellinger-Transformed Fish Data

Analysis of Similarity

Step 4) Perform permutation tests to assess the significance of the observed **R** value.

ANOSIM (observed vs expected R)

Observed R = 0.1 , Expected R = 0.001 , P = 0.235
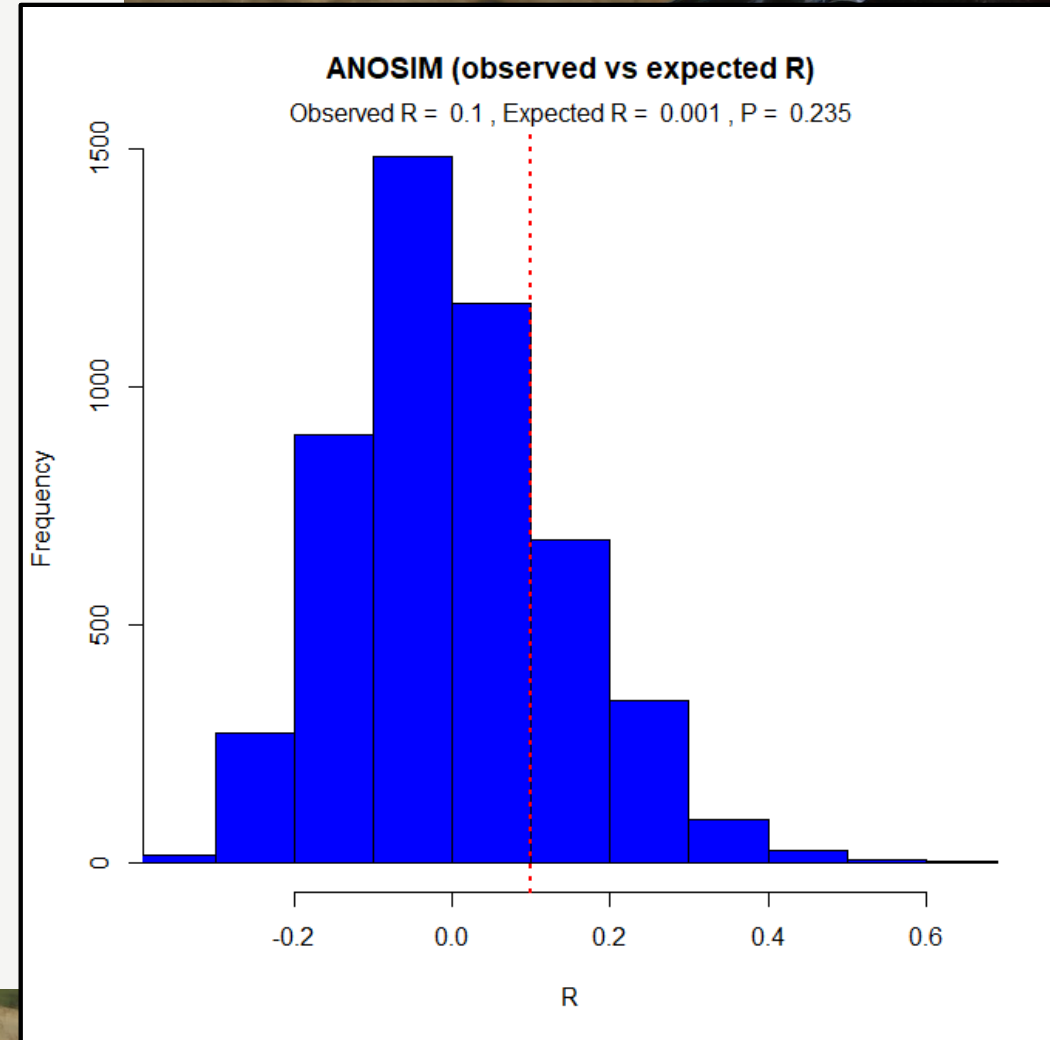
Analysis of Similarity

Step 4) Perform permutation tests to assess the significance of the observed **R** value.
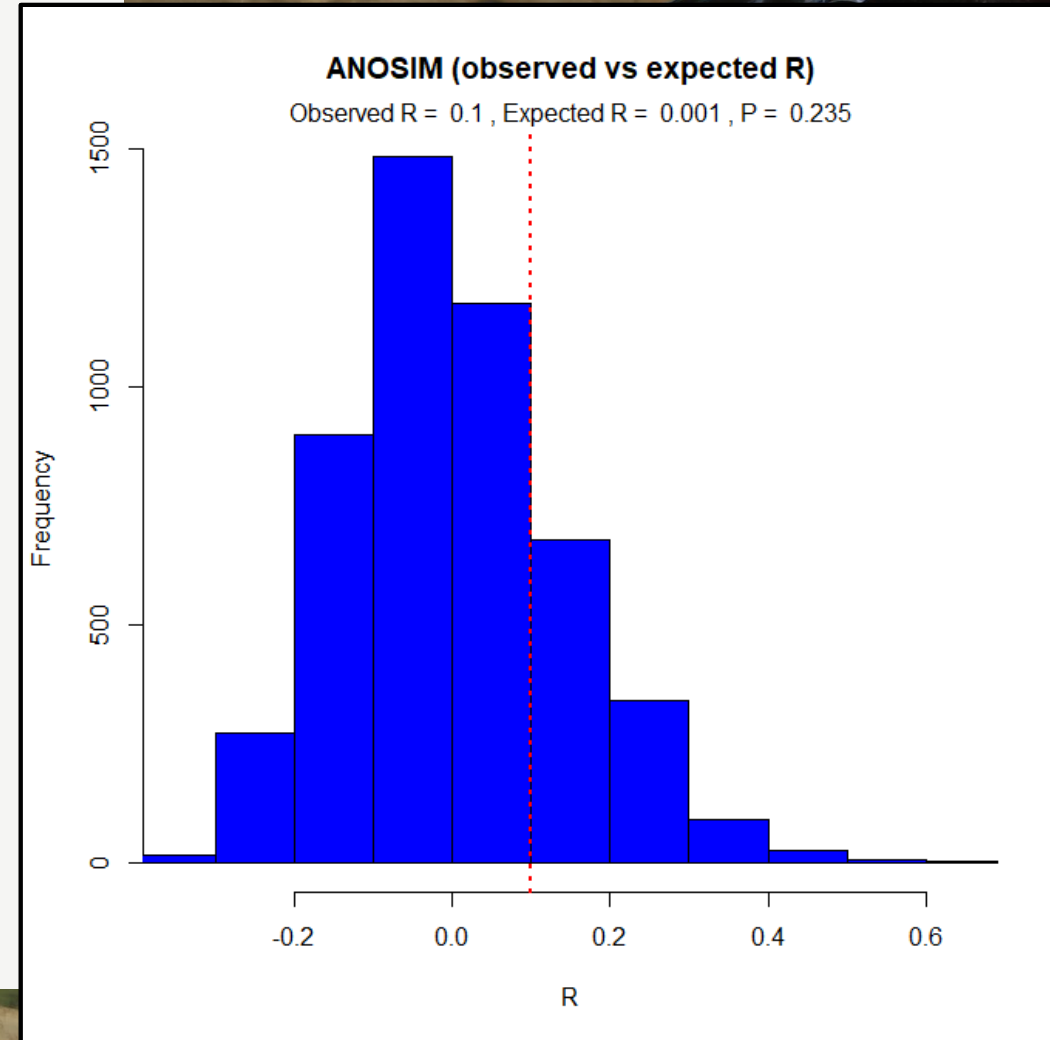
ANOSIM statistic R: 0.010
Significance: 0.235

Upper quantiles of permutations (null model):

90%	95%	97.5%	99%
0.192	0.250	0.301	0.384

Dissimilarity ranks between and within classes:

	0%	25%	50%	75%	100%	N
Between	2	18.5	34	55.00	59	47
High-Forested	6	29.5	36	43.25	47	6
Low-Forested	1	23.5	36	54.00	59	10
Shrub-Scrub	12	13.0	14	15.00	16	3

ANOSIM (observed vs expected R)
Observed R = 0.1 , Expected R = 0.001 , P = 0.235

Analysis of Similarity

Disadvantages:

- Provides no information on which groups differ

- Sensitive to sample size, where larger sample sizes can lead to more significant results

(Permutational) Multivariate Analysis of Variance

(Permutational) Multivariate Analysis of Variance

Multivariate Analysis of Variance (MANOVA) is a statistical test used to compare means of multiple dependent variables across different groups.

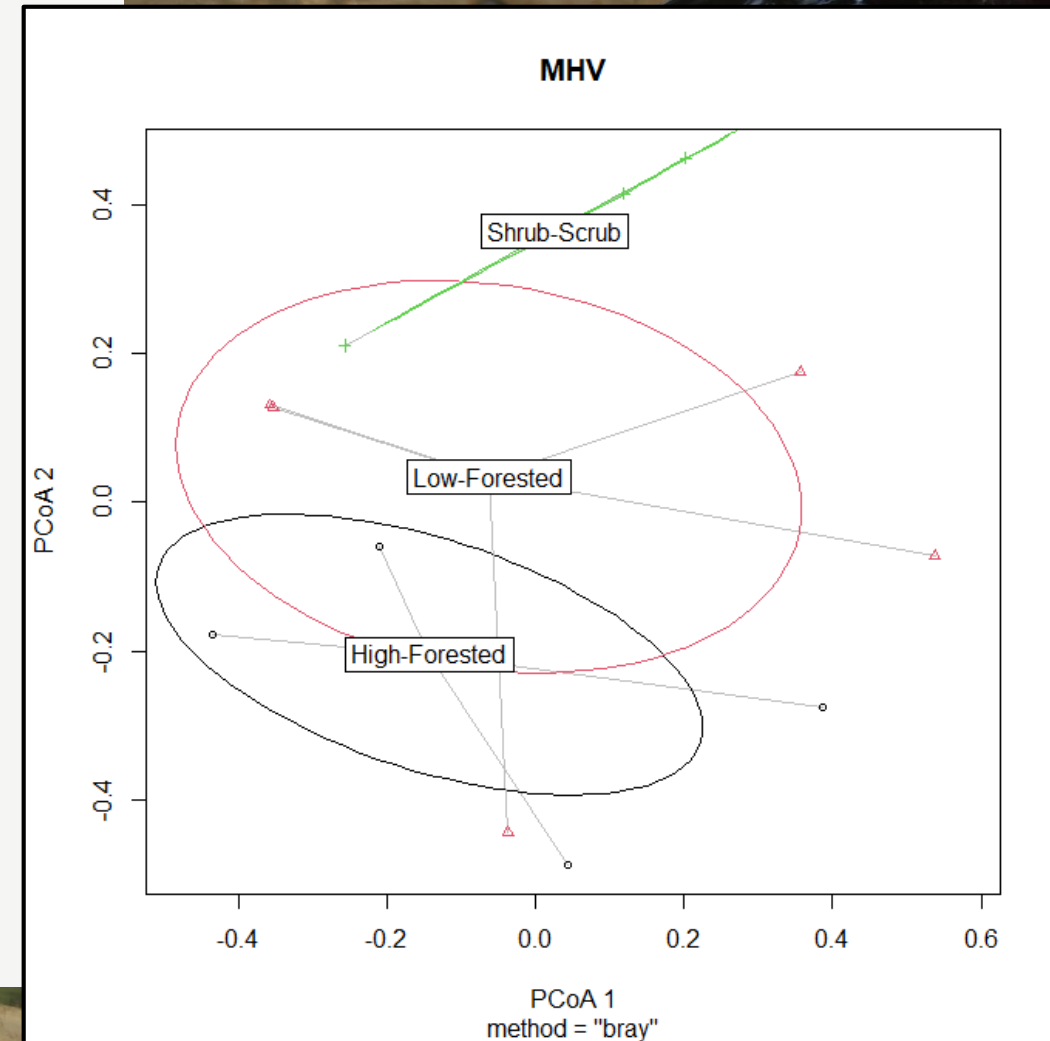# (Permutational) Multivariate Analysis of Variance

Hypothesis Testing:

- **Null hypothesis (H_0):** the means of the groups are equal for all descriptors

- **Alternative hypothesis (H_1):** at least one group mean is different

(Permutational) Multivariate Analysis of Variance

Assumptions:

- Multivariate normality

- Homogeneity of variance-covariance matrices

- Independence of observations

(Permutational) Multivariate Analysis of Variance

MANOVA	PERMANOVA
Multi-variate normal distribution	No assumed distribution
All groups have same variance	Allows for differences in between-group variation
Sensitive to multicollinearity	Insensitive to multicollinearity
Need more samples than variables	Allows more variables
Highly sensitive to many 0's	Insensitive to many 0's

(Permutational) Multivariate Analysis of Variance

Permutational Multivariate Analysis of Variance (PERMANOVA) is a non-parametric statistical test used to assess the differences in multivariate data across groups.
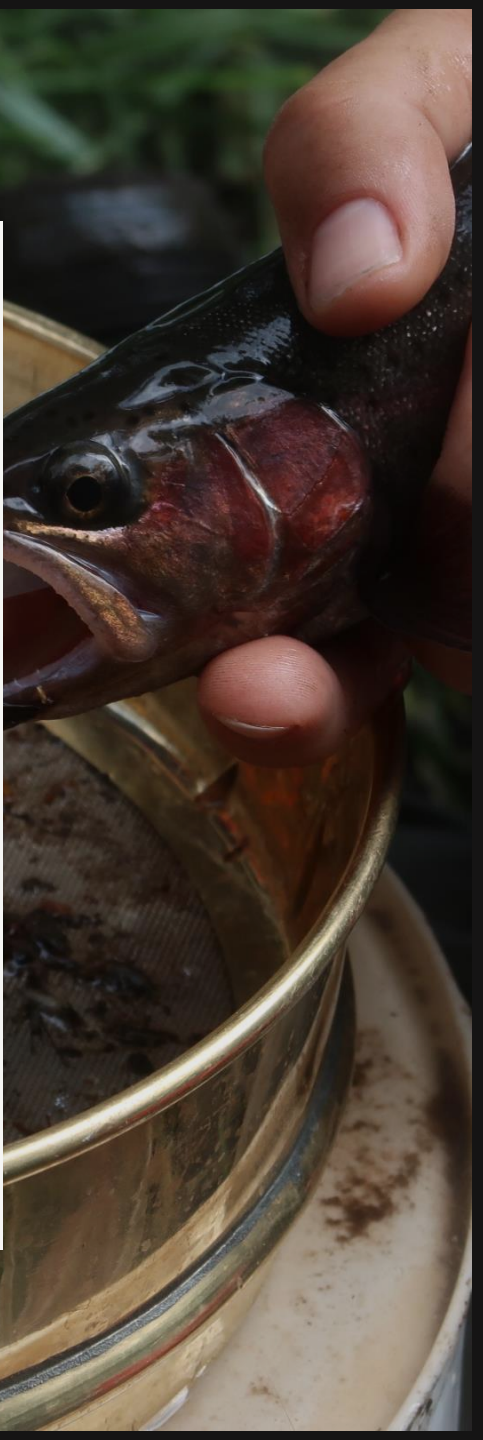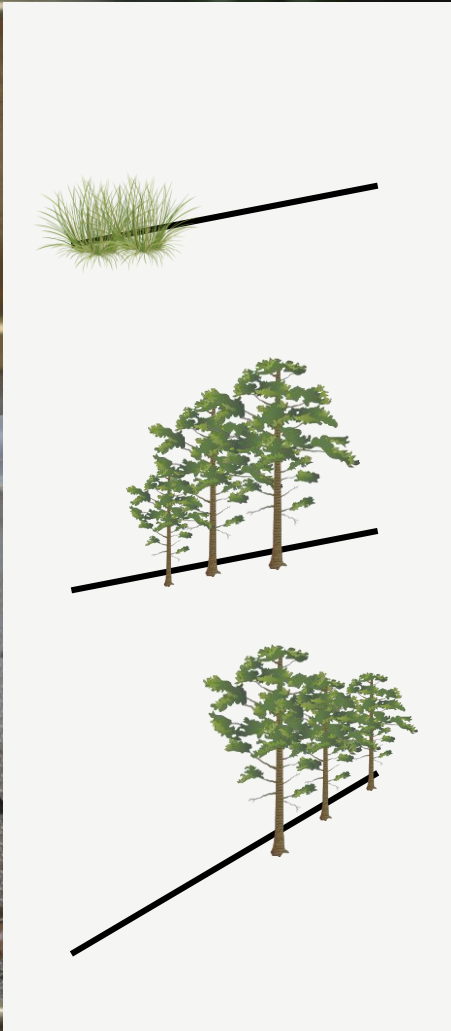
PERMANOVA

Step 1) Calculate the distance matrix

	S-15	S-07	S-12	S-34	S-11	S-21	S-08	S-05	S-03	S-02	S-22
Silvies-07	0.52										
Silvies-12	0.73	0.72									
Silvies-34	1.00	0.62	1.00								
Silvies-11	0.88	0.55	1.00	0.54							
Silvies-21	1.00	0.87	0.88	1.00	1.00						
Silvies-08	1.00	0.88	0.89	0.93	1.00	0.33					
Silvies-05	0.75	0.72	0.76	0.80	1.00	0.90	0.81				
Silvies-03	1.00	0.91	0.93	1.00	1.00	0.80	0.73	0.36			
Silvies-02	1.00	0.64	0.95	0.50	0.58	0.86	0.74	0.42	0.41		
Silvies-22	1.00	0.76	0.96	0.81	0.75	0.90	0.86	0.38	0.38	0.41	
Silvies-18	1.00	0.73	0.95	0.79	0.70	0.88	0.83	0.31	0.29	0.34	0.10

PERMANOVA

Step 2) Define Groups

Site ID	Max Depth (m)	Gradient (%)	Elevation (m)	Canopy (%)	Herb (%)
Silvies-11	0.45	0.3	1439	0.0	55.1
Silvies-34	0.78	1.1	1487	0.0	0.0
Silvies-02	0.71	0.4	1372	29.6	0.0
Silvies-15	0.40	0.2	1471	41.1	0.0
Silvies-07	0.50	1.3	1547	52.3	0.0
Silvies-08	0.40	0.6	1492	51.4	0.0
Silvies-22	0.42	0.9	1555	54.7	0.0
Silvies-18	0.42	0.5	1510	46.2	0.0
Silvies-12	0.52	3.2	1658	51.9	0.0
Silvies-21	0.18	2.4	1713	37.5	0.0
Silvies-05	0.45	5.5	1565	46.7	0.0
Silvies-03	0.20	3.3	1634	59.0	0.0

PERMANOVA

Step 3) Calculate the test statistic (F) based on the variation between and within groups, using permutation methods to generate a null distribution.

PERMANOVA

Step 4) Assess significance by comparing the observed test statistic to the distribution generated from random permutations of the data.

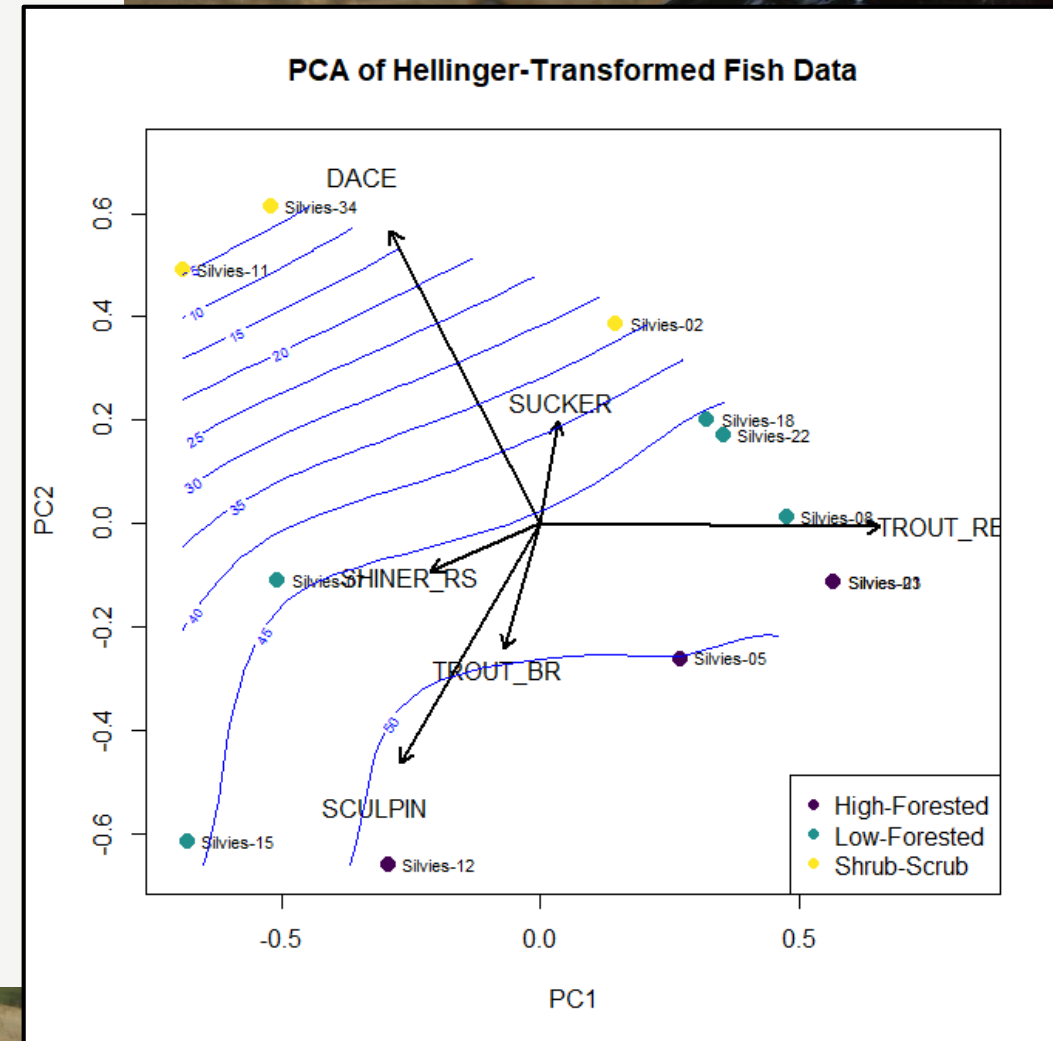|	Df	SS	R2	F	Pr(>F)
silv_dat$HAB2	2	0.898	0.260	1.578	0.139
Residual	9	2.560	0.740		
Total	11	3.458	1.000		

PERMANOVA

The `adonis2()` function in R can handle continuous data and interaction effects.

e.g., fish community structure is significantly associated with canopy cover in our example:

	Df	SS	R2	F	Pr(>F)
silv_dat$Canopy	1	0.718	0.208	2.623	0.030 *
Residual	10	2.739	0.792		
Total	11	3.458	1.000		

PCA of Hellinger-Transformed Fish Data

Similarity Percentage

Similarity Percentage

Similarity Percentage (SIMPER) is a *post-hoc* (usually) method used to identify which descriptors contribute most to the dissimilarity between groups in ecological data.

Similarity Percentage

Similarity Percentage (SIMPER) is a *post-hoc* (usually) method used to identify which descriptors contribute most to the dissimilarity between groups in ecological data.

For each pair of groups, calculates the average dissimilarity and identifies the contributions of individual descriptors to this dissimilarity.

Similarity Percentage

Contrast: High-Forested_Shrub-Scrub

	average	sd	ratio	ava	avb	cumsum	p
DACE	0.324	0.177	1.834	0.000	0.367	0.374	0.002
TROUT_RB	0.259	0.194	1.330	0.359	0.139	0.672	0.727
SUCKER	0.113	0.103	1.095	0.052	0.159	0.803	0.064
TROUT_BR	0.079	0.148	0.535	0.098	0.000	0.894	0.411
SCULPIN	0.066	0.071	0.920	0.101	0.000	0.970	0.889
SHINER_RS	0.025	0.045	0.569	0.000	0.019	1.000	0.631

PCA of Hellinger-Transformed Fish Data
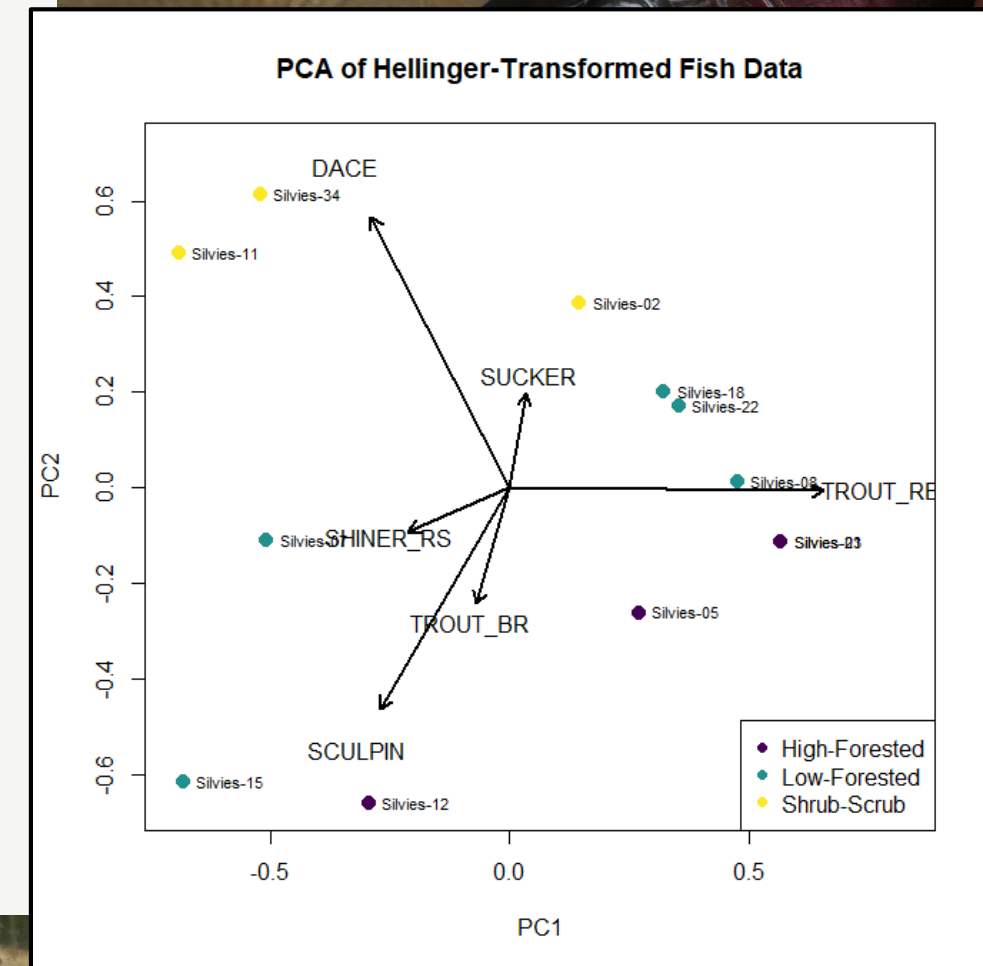
Mantel Test

Mantel Test

A **Mantel Test** assesses the correlation <u>between two distance or similarity matrices</u>.

Mantel Test

A **Mantel Test** assesses the correlation between two distance or similarity matrices.

Commonly used to analyze relationships between spatial, ecological, or genetic distance and environmental factors.

Mantel Test

Step 1) Convert two sets of data into distance or similarity matrices (e.g., lat/long, environmental characteristics).

Mantel Test

Step 1) Convert two sets of data into distance or similarity matrices (e.g., lat/long, environmental characteristics).

Step 2) Compute the correlation coefficient (Pearson or Spearman) between the two distance matrices to produce the **Mantel Statistic (r)**.

Mantel Test

Step 1) Convert two sets of data into distance or similarity matrices (e.g., lat/long, environmental characteristics).

Step 2) Compute the correlation coefficient (Pearson or Spearman) between the two distance matrices to produce the **Mantel Statistic (r)**.

Step 3) Perform a permutation test to generate a *P*-value.

Mantel Test

The **Mantel statistic (r)** indicates the strength and direction of the relationship between the two matrices.

- A positive **r** value suggests a positive correlation between the matrices, while a negative **r** indicates an inverse relationship.

- A significant P-value (e.g., < 0.05) indicates that the observed correlation is unlikely to have occurred by chance.

Mantel Test

Limitations:

- Assumes linear relationships and can be sensitive to data distribution and matrix size

- Interpretation can be complicated by the presence of spatial autocorrelation

- Only useful for **distance** matrices (hence, often used in a spatial capacity)

Procrustes Analysis

Procrustes Analysis

Procrustes Analysis is a statistical technique used to compare and align two or more sets of data by removing differences in scale, rotation, and translation.

Procrustes Analysis

Procrustes Analysis is a statistical technique used to compare and align two or more sets of data by removing differences in scale, rotation, and translation.

Produces Procrustes residuals between aligned datasets to assess how similar they are.

Procrustes Analysis

A **Procrustes statistic** $(0 - 1)$ indicates the level of fit between datasets, where a *lower* value indicates better fit.

Procrustes Analysis

A **Procrustes statistic** $(0 - 1)$ indicates the level of fit between datasets, where a *lower* value indicates better fit.

Procrustes Sum of Squares (m12 squared):   0.601
Correlation in a symmetric Procrustes rotation: 0.632
Significance: 0.055
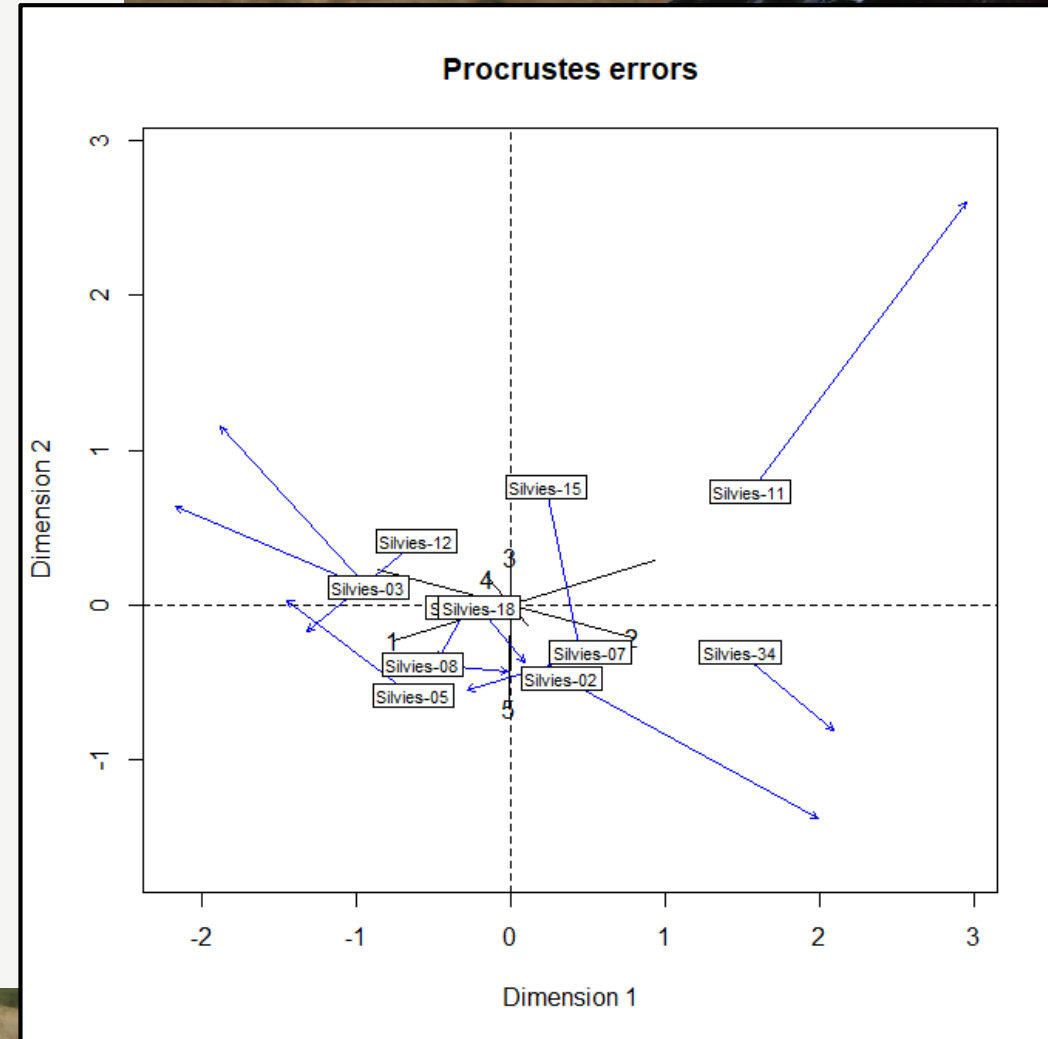
Some Notes on Nested Study Design

Some Notes on Nested Study Design

A **nested study design** involves grouping samples within higher-level factors, where one factor (e.g., sites) is nested within another (e.g., treatments).

Some Notes on Nested Study Design

A **nested study design** involves grouping samples within higher-level factors, where one factor (e.g., sites) is nested within another (e.g., treatments).

PERMANOVA can include nested factors in the model to account for the hierarchical structure.

Some Notes on Nested Study Design

Why account for a nested study design?

- Captures variation at different levels

- Improves statistical power by properly partitioning variance

- Provides insights into both main effects and interaction effects between factors

Conclusion: Summary of Key Points

- Statistical tests used to test for among-group differences in multivariate data:
  - **Analysis of Similarity (ANOSIM)**
  - **(Permutational) Analysis of Variance (PERMANOVA)**

- **Similarity Percentage (SIMPER)** is a post-hoc test used to check for significant drivers of among-group differences

- **Mantel Test** assesses correlation between two distance or similarity matrices

- **Procrustes Analysis** is used to compare and align two or more sets of data by removing differences in scale, rotation, and translation

Questions?