Assessing Optimal Sampling Using Rarefaction

MAGGIE CELENTANO, LANE D'ALESSANDRO

Goal

Develop rarefaction curves

Develop a method to find the optimal amount of sampling

Investigate difference by month and location

Investigate causes of those differences

Data Set

Abundance Data

- Number of times each species appears

Surber Sampling

- 12 Samples collected, pooled in sets of four and only a ¼ of the pooled set kept

8 Months: February – November

- Skips August and October

White Clay Creek

2 Locations: Meadow and Woods

Taxa (Family)	MEADOW 10FEB04	MEADOW 11MAR04	MEADOW 15APR04
PLANARIIDAE	0	24	8
NEMERTEA	0	0	4
NEMATODA	0	40	28
OLIGOCHAETA	8	52	48
NAIDIDAE	0	0	0
ISOPODA	0	4	24
ASELLIDAE	0	0	0
AMPHIPODA	0	0	4
TALITRIDAE	0	0	0
CAMBARIDAE	0	0	0
PLECOPTERA	8	0	44
PTERONARCYIDAE	0	0	0
PELTOPERLIDAE	0	0	0
TAENIOPTERYGIDAE	16	0	12
NEMOURIDAE	24	56	0

Rarefaction Curves

Rarefaction:

- the statistical expectation for the accumulation curve, a step-wise function that plots the number of new species found after sampling m more individuals

Rarefaction Curve:

- Curve which estimates the number of species at a given sample

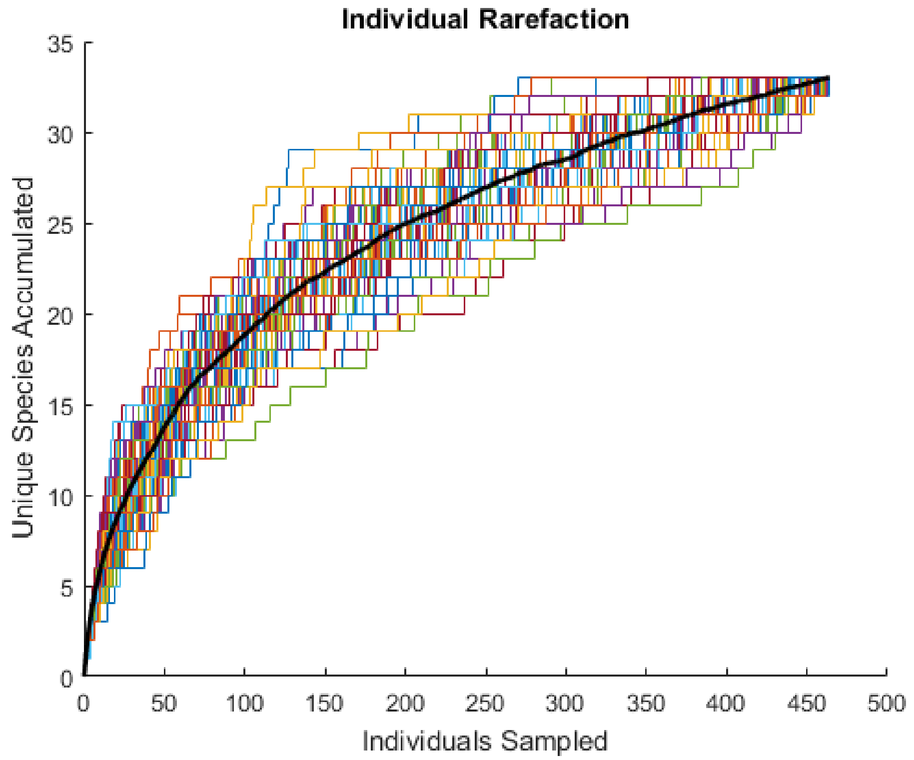
Individual Based Rarefaction

Bootstrap Method

- Artificially perform sampling
- Make a sample set of all the organisms present in data, each species given a unique numerical identifier
- Randomly permute the order of the sample set
- Count unique numbers up to x

Adaptive Method to find the end of repetitions

- Max change in estimated species richness falls below a specific tolerance
  - Largest amount of repetitions = 500



Individual Based Rarefaction

Combinatorics

- $\tilde{S}_{ind}(m) = S - \sum_{i=1}^{S}(1 - p_i)^m$
- $\tilde{S}_{ind}(m) = S_{obs} - \sum_{Xi>0}\left[\frac{\binom{n-X_i}{m}}{\binom{n}{m}}\right]$
- $\tilde{S}_{ind}(m) = S_{obs} - \sum_{k=1}^{n}\left[\frac{\binom{n-k}{m}}{\binom{n}{m}}\right] * f_k$
- Minimum Variance Unbiased Estimator model
  - Hypergeometric model and multinomial model
  - Assumes sampling without replacement

Estimator for Species Richness

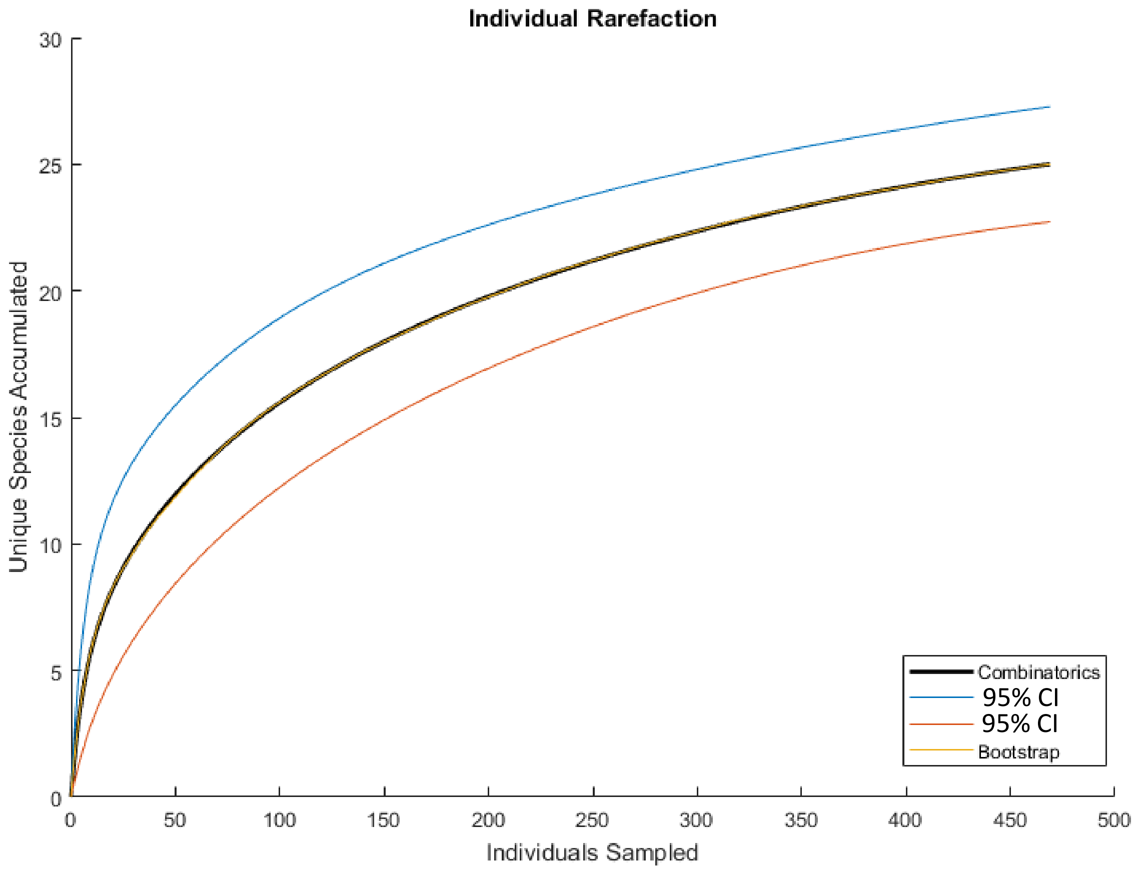
Chao1

- Classic Form: $S_{Chao1} = S_{obs} + \frac{f_1^2}{2f_2}$
- Bias Corrected: $S_{Chao1} = S_{obs} + \frac{f_1(f_1-1)}{2(f_2+1)}$

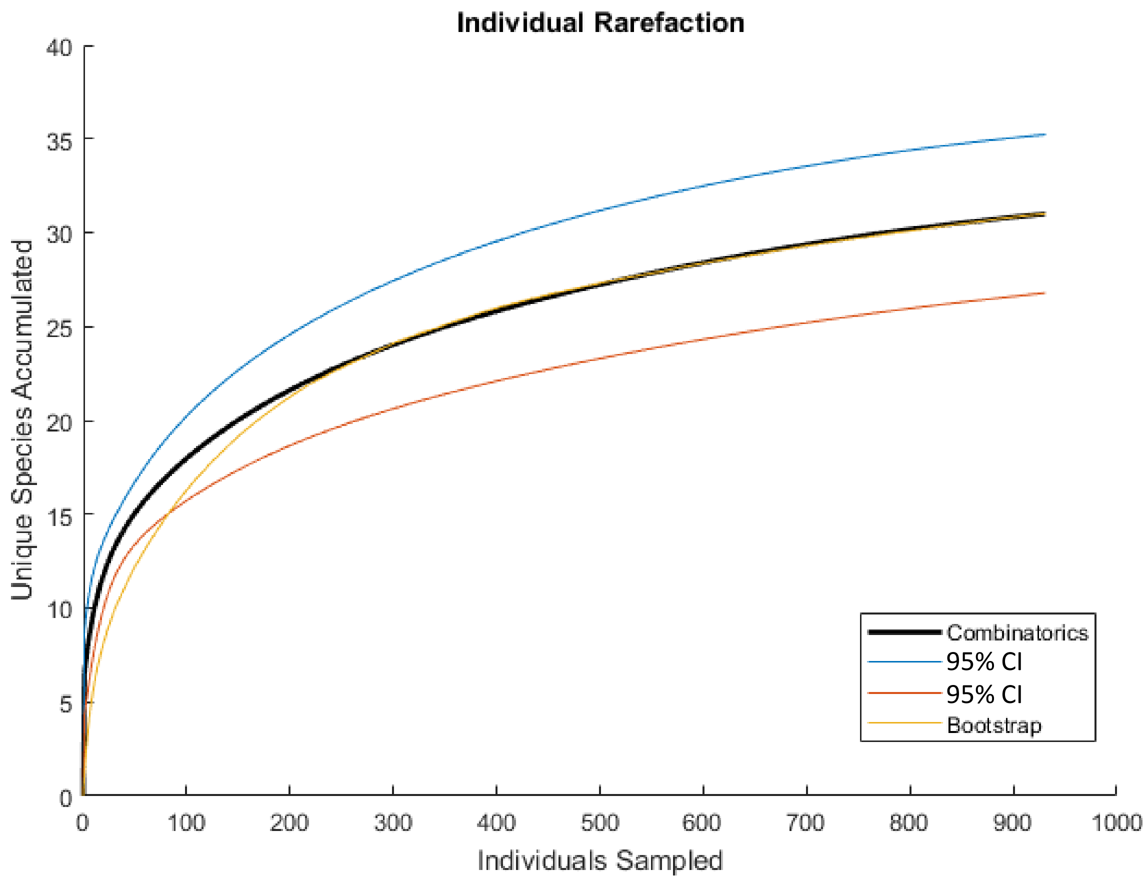
Variance

- $\alpha_{km} = \frac{\binom{n-k}{m}}{\binom{n}{m}}$
- $\sigma^2(m) = \sum_{k=1}^{n}(1 - \alpha_{km})^2 f_k - \tilde{S}_{ind}(m)^2/S_{est}$

Combinatorics



Combinatorics



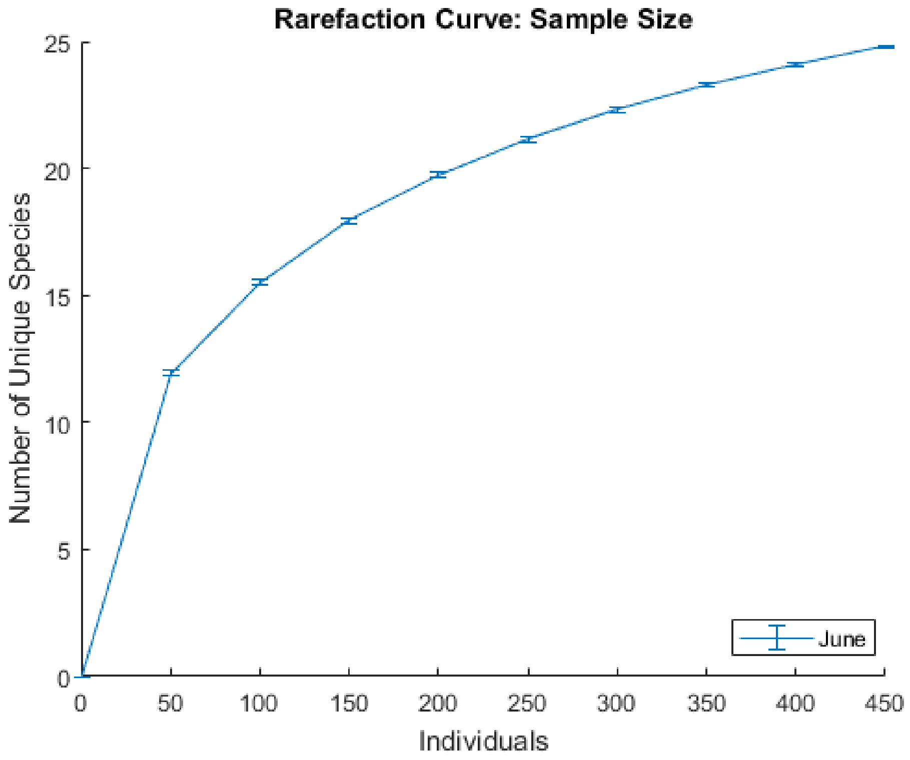
Sampling Rarefaction

Bootstrap Method

- Artificially perform sampling
- Make a sample set of all the organisms present in data, each species given a unique numerical identified
- Randomly permute the order of the sample set
- Count unique numbers up to x members in the sample set, increasing by the sample size.

Recommended by liaison from Stroud Water Research Center

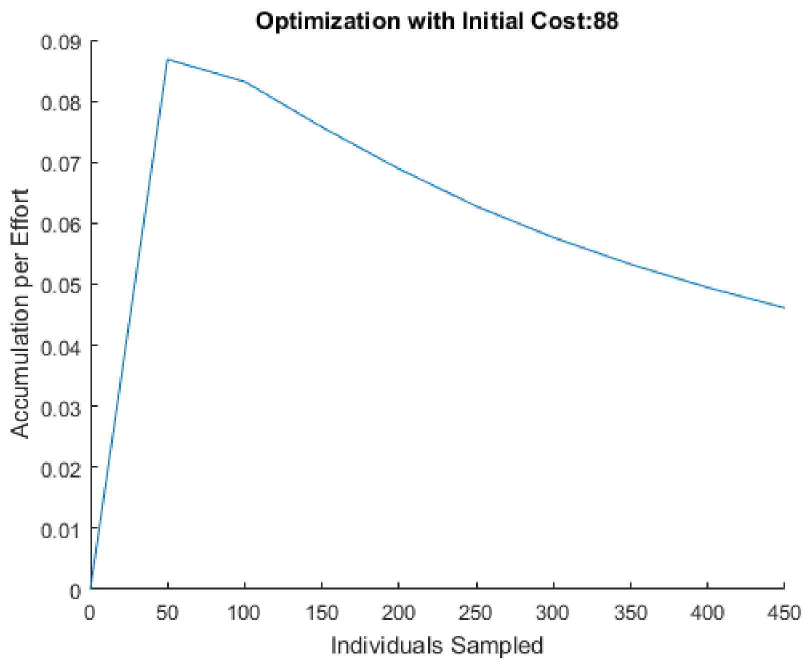
All future data will use sample sizes of 50

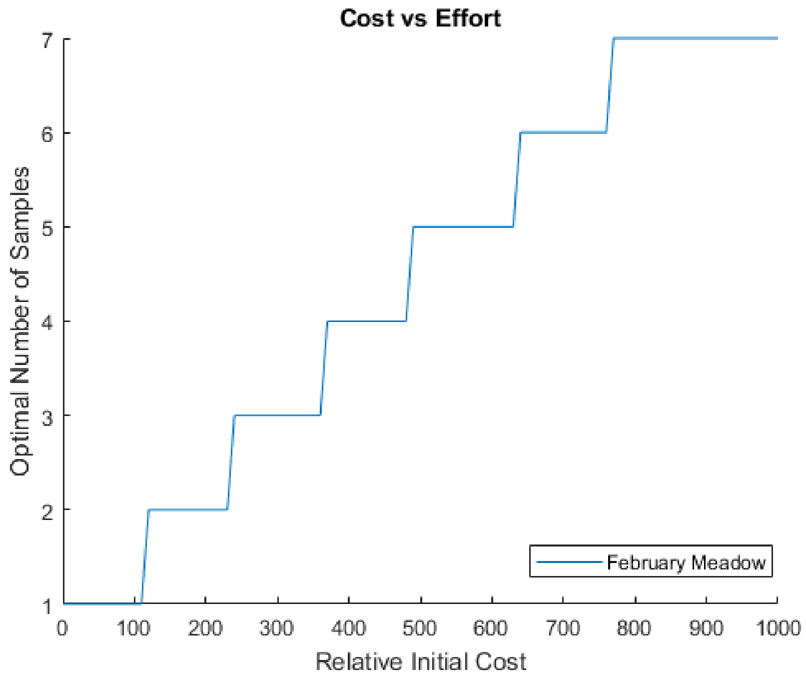


Accumulation with Initial Cost

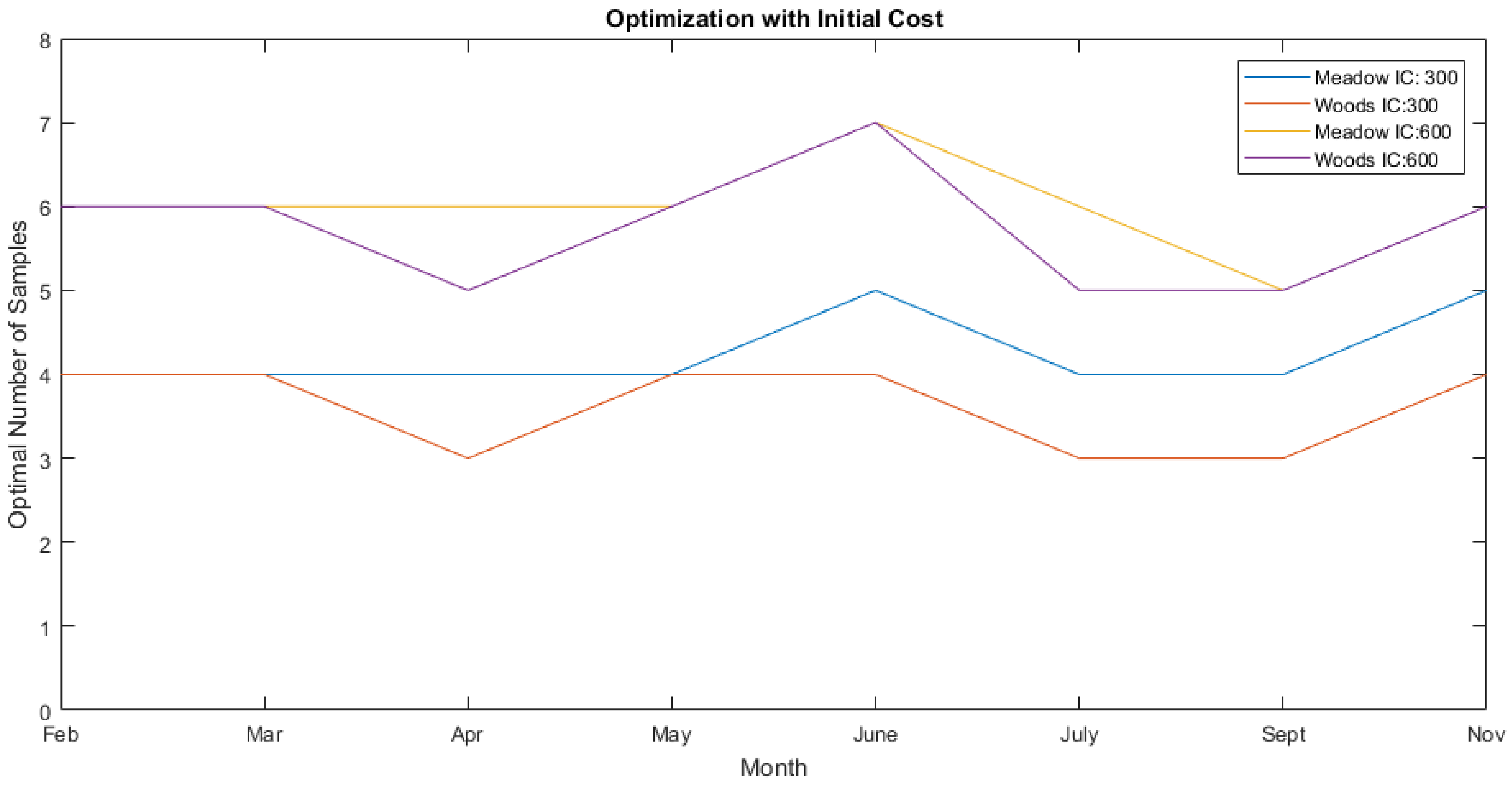
$$AE = \frac{\tilde{S}_{ind}(m)}{IC+m}$$

- Optimize the accumulation rate to the effort put forth





Initial Cost By Month



Coverage Based Rarefaction

Coverage

- The percent of organisms represented by species present in the sample

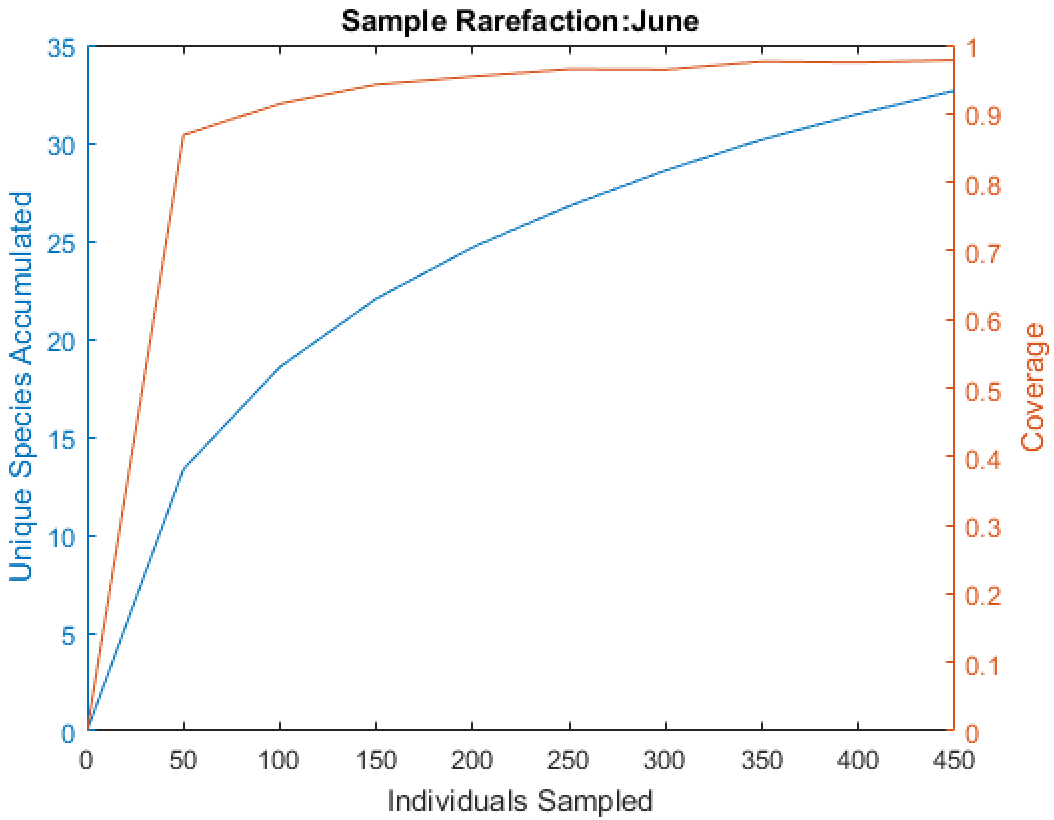
Coverage can be used to identify sampling effort.

Identify the desired coverage to find recommended sampling

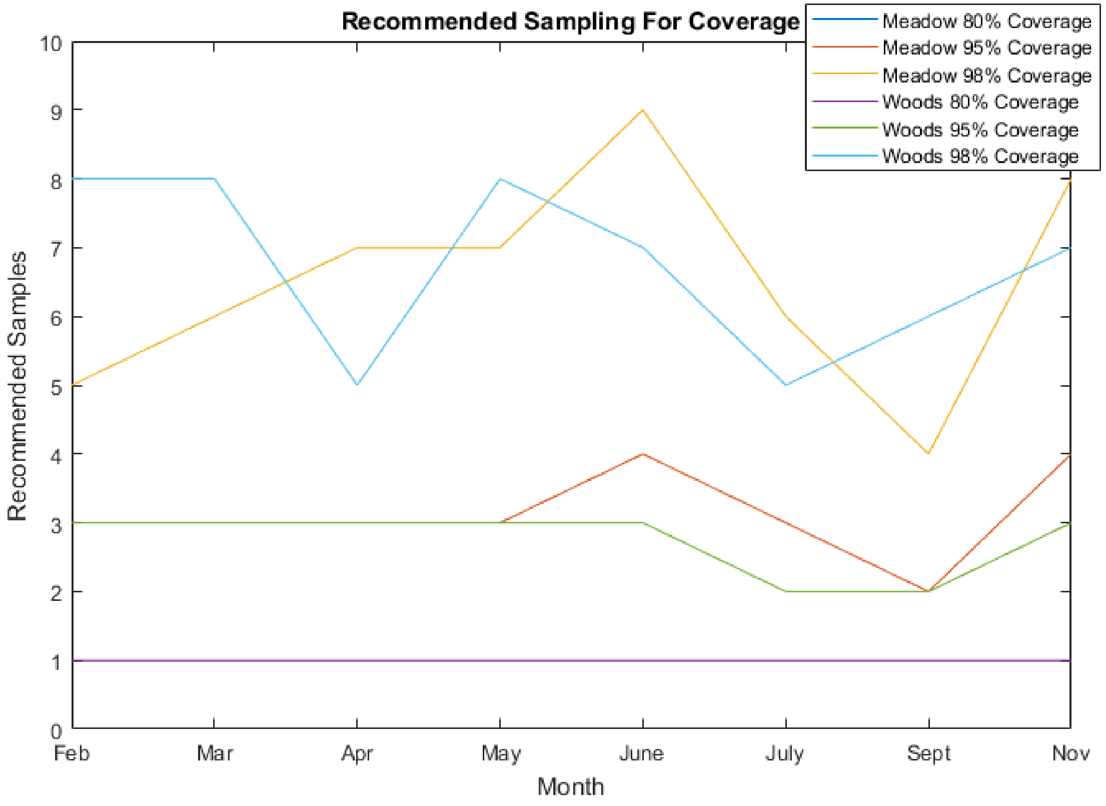
Estimating Coverage

- *Formulaic*
  - $1 - \hat{C}_m = \hat{S}_{m+1} - \hat{S}_m$
- Unbiased Algorithm
  - $\hat{C}_m = 1 - \frac{f_1(m+1)}{m+1}$
    - Repeated a sufficiently large number of times and averaged

Coverage Based Rarefaction



Samples Required for Specified Coverage: Meadow vs Woods



Degree-Days

Degree-Days

- The difference between the mean temperature of a day and a developmental threshold temperature

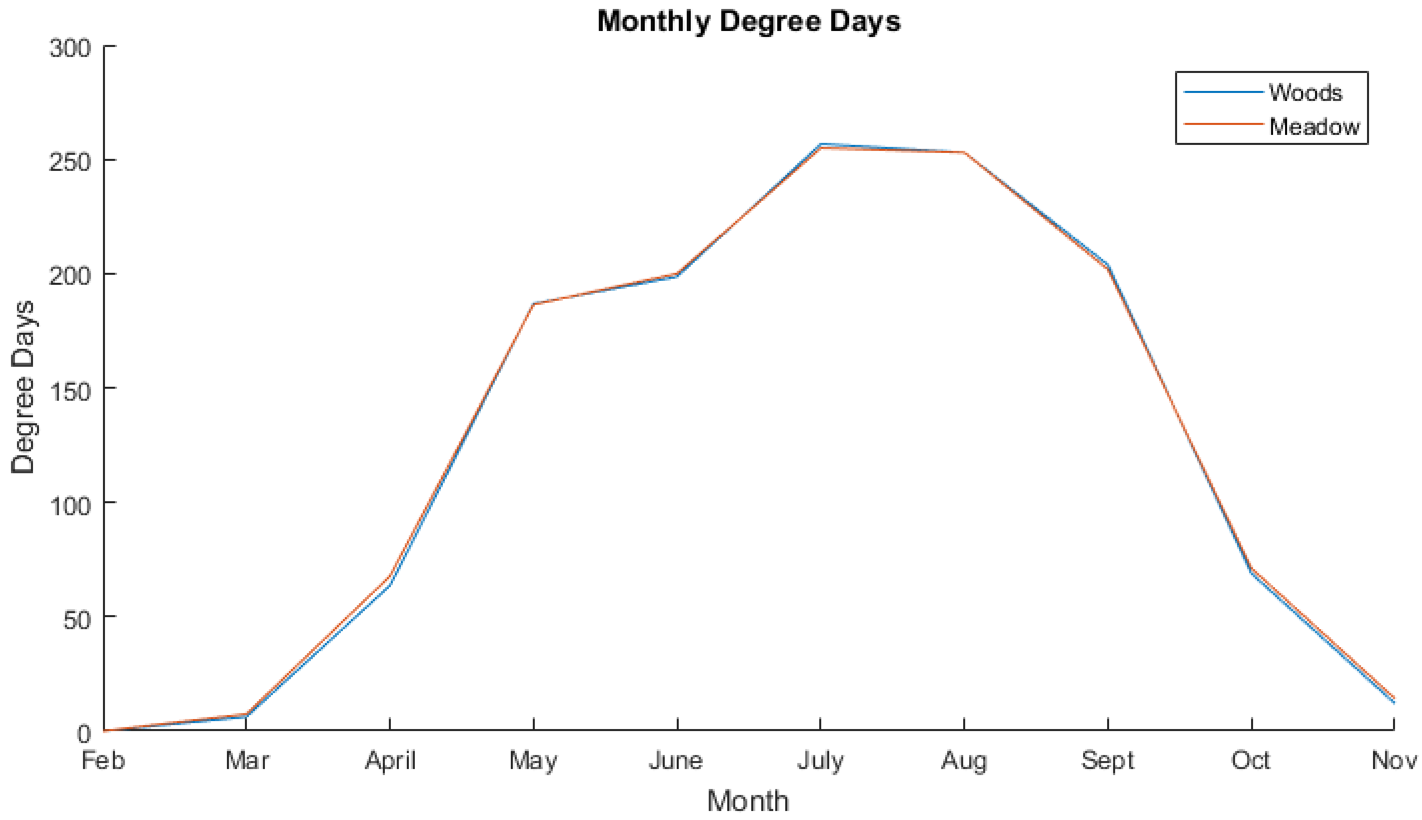
Climate indicator

- Representative of the growth for organisms
  - More degree-days indicates more growth
- Each species has a different developmental threshold temperature

$$DD_{day} = \left[\sum_{i=1}^{24} (T - T_{base})^{+}\right] / \ 24$$

$$DD_{month} = \sum_{j=1}^{N} DD_{day,j}$$

Degree-Days, Meadow vs Woods



Diversity Index: Species Richness

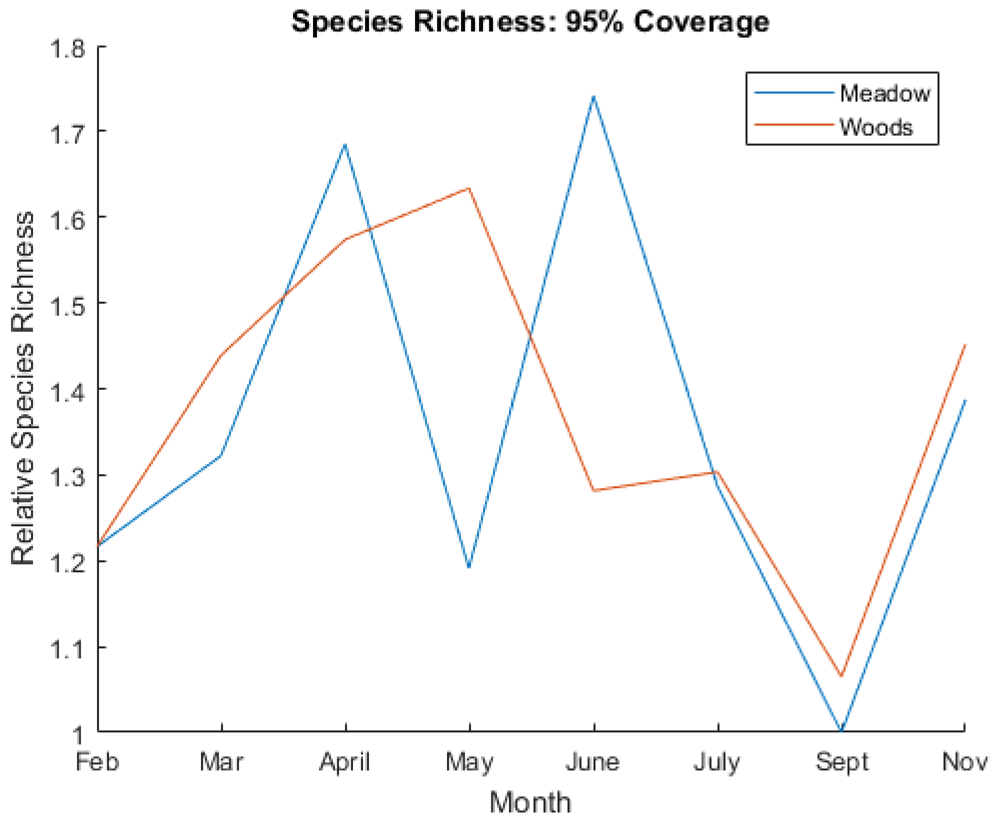
Need to compare species richness across a standardized sample

Standardized samples by coverage, not number of individuals, due to differences in species-abundance distributions

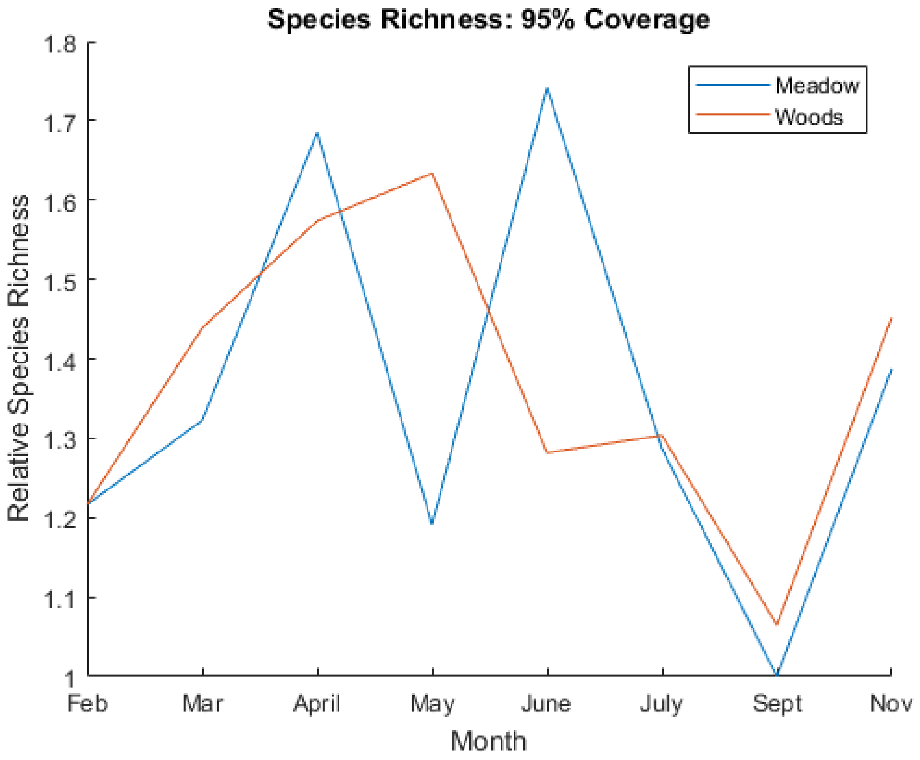
95% Coverage level

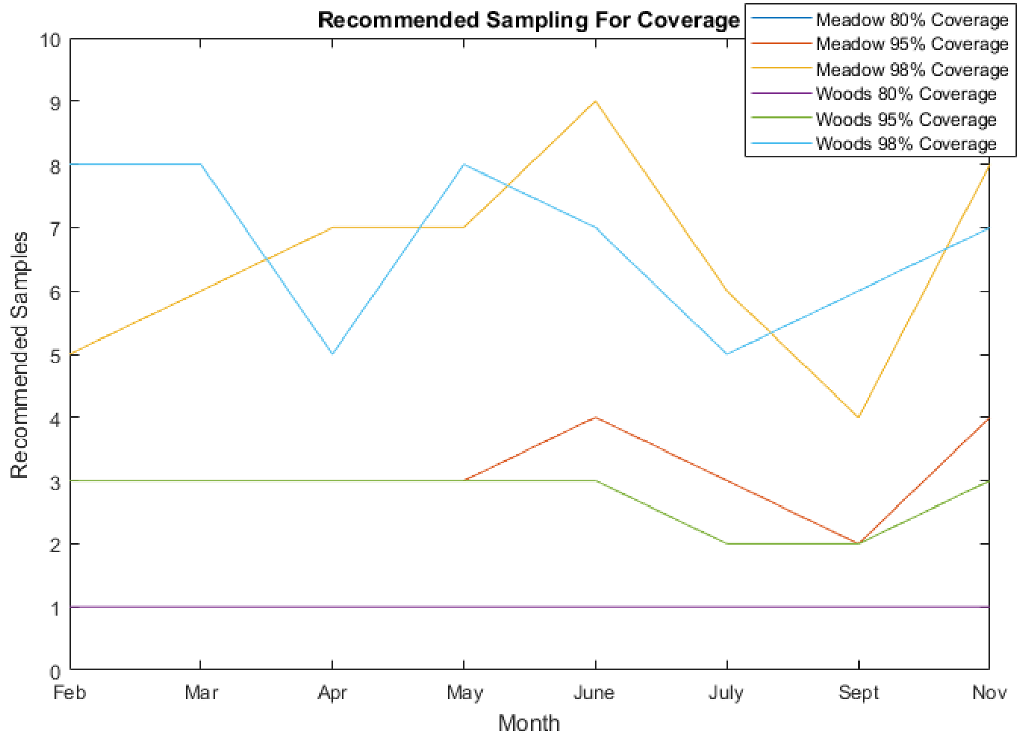
- Ratio of species richness between sites does not vary by coverage level

Species Richness



Species Richness





Diversity Index: Species Evenness

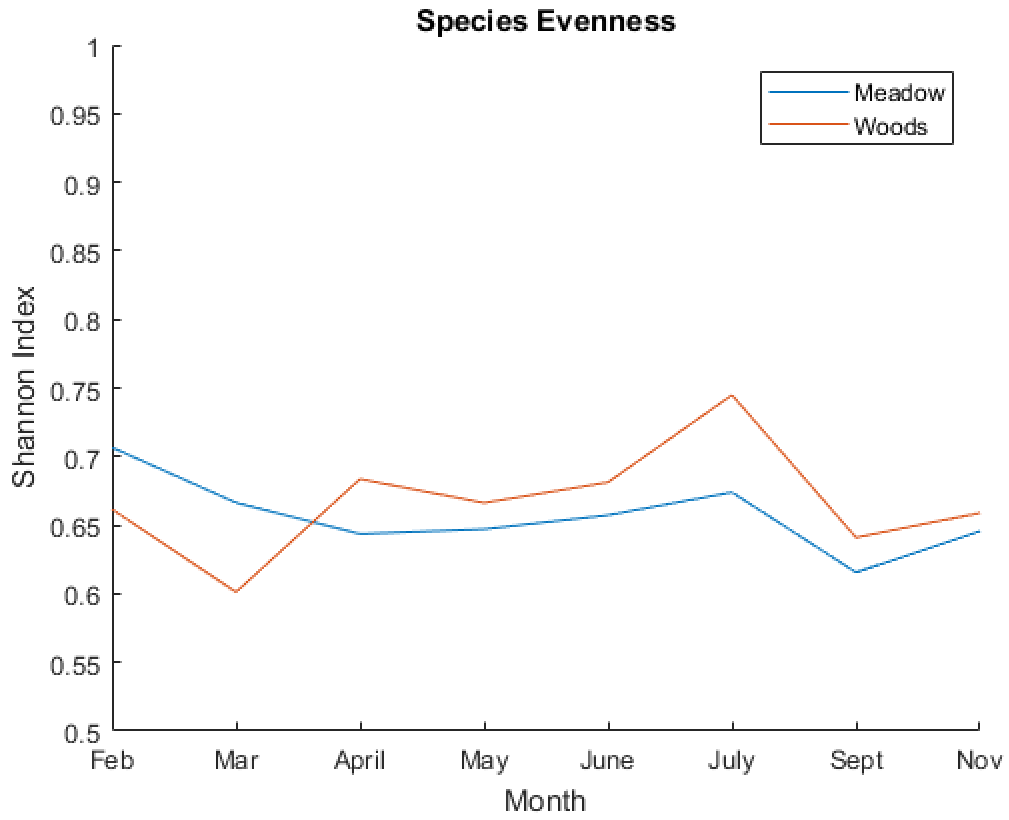
Evenness

- A measure of how equally distributed the individuals are between species

Pielou's Evenness: A measure of species evenness

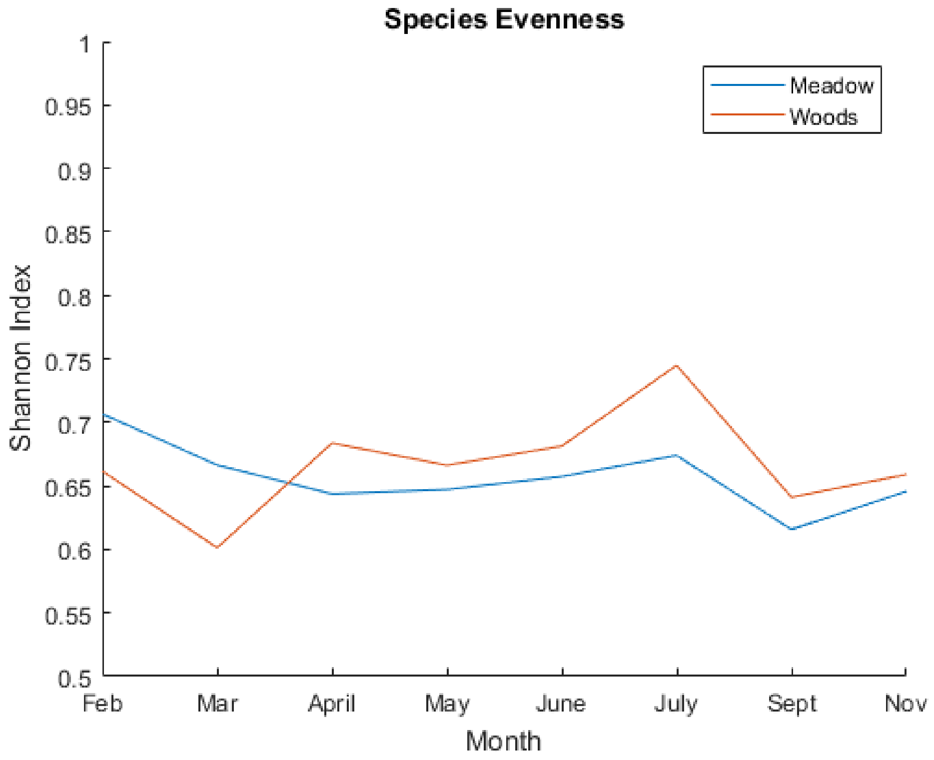
- $p_i = \frac{N_i}{N}$: Relative Species abundance in sample
- $H' = -\sum p_i \ln p_i$
- E = $H'/H'_{\max}$= H'/ln(S)
- Higher values indicate a more even population

Evenness

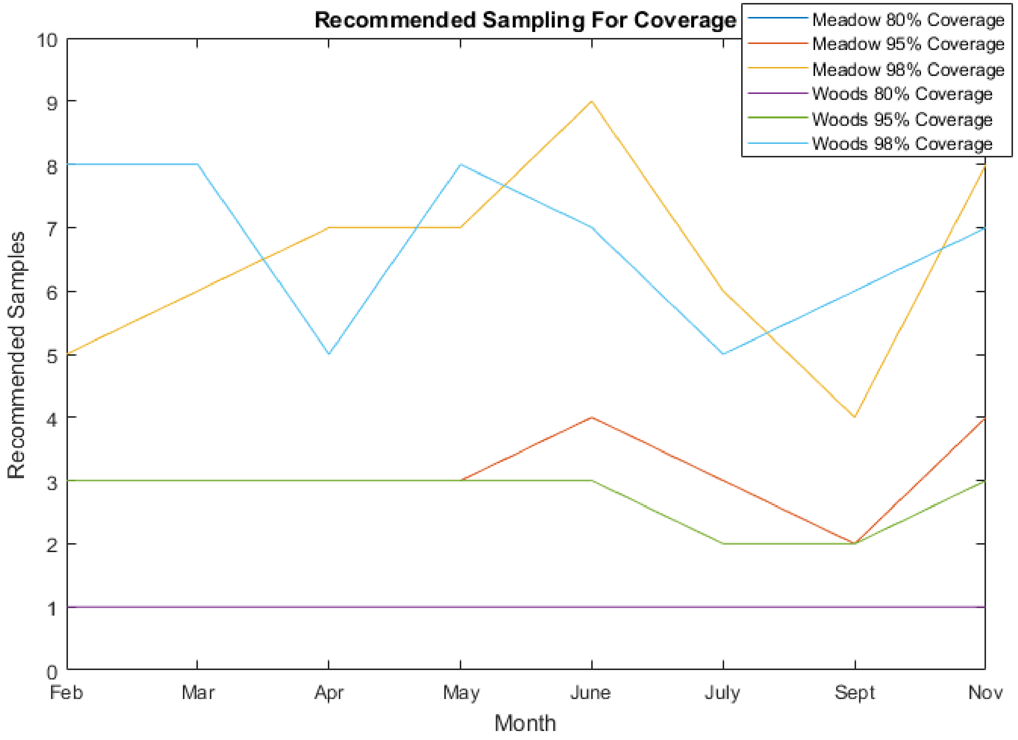


Species Evenness
Meadow
Woods
Shannon Index
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Feb
Mar
April
May
June
July
Sept
Nov
Month

Evenness



Species Evenness
Meadow
Woods
Shannon Index
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Feb
Mar
April
May
June
July
Sept
Nov
Month



Recommended Sampling For Coverage
Meadow 80% Coverage
Meadow 95% Coverage
Meadow 98% Coverage
Woods 80% Coverage
Woods 95% Coverage
Woods 98% Coverage
Recommended Samples
10
9
8
7
6
5
4
3
2
1
0
Feb
Mar
Apr
May
June
July
Sept
Nov
Month

Extensions

Develop unbiased variance and confidence intervals for bootstrap rarefaction

Verify Temperature Data

Identify cause of the difference in sampling

Identify source of error in combinatorics method

References

Allen, Benjamin, et al. “A New Phylogenetic Diversity Measure Generalizing the Shannon Index and Its Application to Phyllostomid Bats.” *The American Naturalist*, vol. 174, no. 2, 2009, pp. 236–243.

Borah, Pallavi, Manoj Singh, and Sadhan Manapatra. "Estimation of degree-days for different climatic zones of North-East India." *Sustainable Cities and Society,* vol. 14, 2015, pp. 70-81.

Chao, Anne, and Lou Jost. “Coverage-Based Rarefaction and Extrapolation: Standardizing Samples by Completeness Rather than Size.” *Ecology*, vol. 93, no. 12, 2012, pp. 2533–2547.

Chao, Anne. “Nonparametric Estimation of the Number of Classes in a Population.” *Scandinavian Journal of Statistics*, vol. 11, no. 4, 1984, pp. 265–270.

Colwell, Robert K., et al. "Models and Estimators Linking Individual-Based and Sample-Based Rarefaction, Extrapolation and Comparison of Assemblages." *Journal of Plant Ecology*, vol. 5, no. 1, Jan. 2012, pp. 3-21.

Gotelli, Nicholas, and Robert Colwell. "Estimating Species Richness." *Biological Diversity: Frontiers in Measurement and Assessment*. 1st ed. N.p.: Oxford U Press, 2011. 39-54.

Gotelli, Nicholas J. and Robert K. Colwell. "Quantifying Biodiversity: Procedures and Pitfalls in the Measurement and Comparison of Species Richness." *Ecology Letters*, vol. 4, no. 4, July 2001

Smith, Woollcott, and J. Frederick Grassle. “Sampling Properties of a Family of Diversity Measures.” *Biometrics*, vol. 33, no. 2, 1977, pp. 283–292.

Acknowledgements

West Chester University

Dr. Kolpas

Matt Wilson and Val Oullet Ph.D

Stroud Water Research Center