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 $\frac{1}{4}$ of the pooled set kept

8 Months: February – November
- Skips August and October

White Clay Creek

2 Locations: Meadow and Woods

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

- \( \tilde{S}_{\text{ind}}(m) = S - \sum_{i=1}^{S}(1 - p_i)^m \)
- \( \tilde{S}_{\text{ind}}(m) = S_{\text{obs}} - \sum_{x_i>0} \left[ \binom{n-x_i}{m} \right] \)
- \( \tilde{S}_{\text{ind}}(m) = S_{\text{obs}} - \sum_{k=1}^{n} \left[ \binom{n-k}{m} \right] * f_k \)

Minimum Variance Unbiased Estimator model

- Hypergeometric model and multinomial model
- Assumes sampling without replacement

Estimator for Species Richness

Chao1

- Classic Form: \( S_{\text{chao1}} = S_{\text{obs}} + \frac{f_1^2}{2f_2} \)
- Bias Corrected: \( S_{\text{chao1}} = S_{\text{obs}} + \frac{f_1(f_1-1)}{2(f_2+1)} \)

Variance

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

Bootstrap Method

- Articially 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

\[ AE = \frac{\hat{S}_{ind}(m)}{IC + m} \]

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

Sample Rarefaction: June

Unique Species Accumulated

Coverage

Individuals Sampled
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

\[
\begin{align*}
DD_{\text{day}} &= \left[ \sum_{i=1}^{24} (T - T_{\text{base}})^+ \right] / 24 \\
DD_{\text{month}} &= \sum_{j=1}^{N} DD_{\text{day},j}
\end{align*}
\]
Degree-Days, Meadow vs Woods

Monthly Degree Days

- Woods
- Meadow

- February
- March
- April
- May
- June
- July
- August
- September
- October
- November
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: 95% Coverage](image)
Species Richness

Species Richness: 95% Coverage

Recommended Sampling For Coverage
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
- $H' = -\sum p_i \ln p_i$
- $E = H'/H'_{\text{max}} = H'/\ln(S)$
- Higher values indicate a more even population
Evenness

Species Evenness

Shannon Index

Month

Feb Mar April May June July Sept Nov

Meadow
Woods
Evenness

**Species Evenness**

- Shannon Index

**Recommended Sampling For Coverage**

- Meadow 80% Coverage
- Meadow 95% Coverage
- Woods 80% Coverage
- Woods 95% Coverage

*Month: Feb, Mar, Apr, May, June, July, Sept, Nov*
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


Acknowledgements

West Chester University
Dr. Kolpas
Matt Wilson and Val Oullet Ph.D
Stroud Water Research Center