Instructions for Alpha, Beta, and Gamma Diversity Analysis

Overview

This guide explains how to use the `diversity_analysis.R` script to calculate alpha (α), beta (β), and gamma (γ) diversity for stream macroinvertebrate communities using multiple methods.

Required Files

1. **masterTaxaGenus.csv** - Taxonomic reference data
2. **stationBenthicsTESTSITE.csv** - Benthic sample data (species counts)
3. **stationInfoBenSampsTESTSITE.csv** - Station and sample metadata

Required R Packages

Install these packages before running the analysis:

```
install.packages(c("vegan", "iNEXT", "betapart", "dplyr", "tidyr"))
```

What the Script Calculates

1. Alpha Diversity (α) - Within-Site Diversity

Alpha diversity is calculated for **each individual site** (each benthic sample). The script computes:

- **q = 0 (Species Richness)**: Total number of species present
  - Treats all species equally
  - Most sensitive to rare species
- **q = 1 (Hill-Shannon Diversity)**: Exponential of Shannon entropy
  - Accounts for both richness and evenness
  - Balanced emphasis on common and rare species
- **q = 2 (Hill-Simpson Diversity)**: Inverse Simpson index
  - Emphasizes dominant/common species
  - Less sensitive to rare species

Interpretation: Higher alpha diversity = more diverse community at that specific site

2. Gamma Diversity (γ) - Regional Diversity

Gamma diversity is calculated **across all sites within each year**:

- **q = 0**: Total number of unique species found across all sites in the year
- **q = 1**: Shannon-based diversity pooling all individuals from the year
- **q = 2**: Simpson-based diversity pooling all individuals from the year

Interpretation: Higher gamma diversity = more species present in the region

3. Beta Diversity (β) - Between-Site Diversity

The script calculates beta diversity using three different approaches:

A. Hill Numbers Method (β = γ/α)

Following the chapter equation: **γ = α × β**, therefore **β = γ/α**

- **Beta (q=0)**: Number of distinct community compositions
- **Beta (q=1)**: Effective number of equally common communities
- **Beta (q=2)**: Effective number of dominant communities

Interpretation: - β ≈ 1: Sites have very similar species composition (low turnover) - β >> 1: High turnover in species composition across sites - The value tells you "how many times" the species composition changes across your region

B. Jaccard Index (Presence/Absence)

- Based only on which species are present or absent
- Ignores abundance differences
- Range: 0 (identical communities) to 1 (completely different)
- Good for understanding compositional turnover

C. Bray-Curtis Dissimilarity (Abundance-Based)

- Accounts for both presence/absence AND abundance
- Sensitive to dominant species
- Range: 0 (identical) to 1 (completely different)
- Better for detecting abundance-driven differences

How to Run the Analysis

Step 1: Prepare Your Workspace

```r
# Set your working directory to where your CSV files are located
setwd("path/to/your/data/files")

# Verify files are present
list.files(pattern = "*.csv")
```

Step 2: Run the Script

```r
# Option A: Run the entire script
source("diversity_analysis.R")
```

```
# Option B: Run section by section (recommended for learning)
# Copy and paste each section into R console
```

Step 3: Review Output

The script will print results to your console and create two output files:

1. **alpha_diversity_results.csv**: Site-level diversity metrics
    - One row per site (BenSampID)
    - Includes q0, q1, q2 diversity values
    - Includes year and season information
2. **diversity_summary_by_year.csv**: Annual summary statistics
    - Gamma diversity (q0, q1, q2)
    - Mean alpha diversity (q0, q1, q2)
    - Beta diversity from Hill numbers (q0, q1, q2)
    - Mean Jaccard dissimilarity
    - Mean Bray-Curtis dissimilarity

Understanding Your Results

Example Interpretation

If for Year 2015 you get: - **gamma_q0 = 45** (45 total species found) - **mean_alpha_q0 = 15** (average of 15 species per site) - **beta_q0 = 3** (45/15 = 3)

This means: - The region contains 45 unique species - On average, each site contains 15 species - Beta diversity of 3 indicates that species composition changes about 3 times across your sites - This suggests moderate to high turnover in community composition

Comparing Methods

When to focus on each beta diversity metric:

1. **Hill Numbers (β = γ/α)**:
    - Best for understanding how many distinct communities you have
    - Easy to interpret: “species composition changes X times”
    - Use q=0 for rare species, q=1 for balanced view, q=2 for common species
2. **Jaccard Index**:
    - When you only care about which species are present
    - Good for presence/absence data or when abundances are unreliable
    - Classic ecological metric with lots of literature support
3. **Bray-Curtis**:
    - When abundance matters (e.g., detecting pollution effects on dominant taxa)

- More sensitive to changes in common species
- Widely used in stream ecology

Customizing the Analysis

Analyze by Season Instead of Year

Replace Year with Season in the grouping:

```
# In section 4 (Gamma diversity), change:
gamma_by_season <- data_merged %>%
  group_by(Season) %>%  # Changed from Year
  summarise(
    gamma_q0 = length(unique(FinalID)),
    .groups = "drop"
  )
```

Filter to Specific Stations

```
# Before creating the species matrix, add:
data_merged <- data_merged %>%
  filter(StationID %in% c("2-JKS023.61", "another-station-id"))
```

Use Different Taxonomic Levels

The script currently uses `FinalID` from the benthics data. To use Family level:

```
# Merge with master taxa first
data_with_family <- data_merged %>%
  left_join(master_taxa %>% select(FinalID, Final.VA.Family.ID),
            by = "FinalID")

# Then use Final.VA.Family.ID instead of FinalID in the pivot_wider step
```

Troubleshooting

Error: “Package not found”

```
# Install the package
install.packages("package_name")
```

Error: “Object not found”

- Make sure you’ve run all previous sections in order
- Check that your CSV files are in the working directory

Warning: “NaN produced”

- This can occur when a site has zero individuals
- The script handles this, but you may want to filter out empty samples

Results seem wrong

1. Check your date format in the station info file
2. Verify that BenSampID matches between files
3. Ensure Individuals column contains numeric counts

References

The methods in this script are based on:

- **Hill Numbers**: Chao et al. (2014) Ecological Monographs
- **Beta Diversity Theory**: Jost (2007) Ecology; Chao et al. (2012) Ecology
- **Jaccard & Bray-Curtis**: Anderson et al. (2011) Ecology Letters
- **R Implementation**: vegan (Dixon 2003), iNEXT (Hsieh et al. 2016)

Next Steps

After running this analysis, you might want to:

1. **Visualize results**: Create plots of diversity over time
2. **Statistical testing**: Test if diversity differs between years/seasons
3. **Environmental correlations**: Relate diversity to habitat variables
4. **NMDS ordination**: Visualize community composition patterns
5. **Indicator species**: Identify which taxa drive beta diversity patterns

Questions?

Common questions about diversity analysis:

Q: Which beta diversity metric should I report? A: Report multiple! They each capture different aspects. The Hill numbers approach ($\beta = \gamma/\alpha$) is most intuitive for explaining to non-ecologists.

Q: Why are my q=1 and q=2 values so different from q=0? A: This is normal! q=0 counts all species equally, while q=1 and q=2 give more weight to common species. Large differences suggest your communities have many rare species.

Q: Can I compare diversity across sites with different sampling effort? A: The basic script doesn't correct for sampling effort. Use the iNEXT section (section 10) for rarefaction/extrapolation if sampling effort varies substantially.

Q: What's a "high" vs "low" beta diversity? A: It depends on your system! Compare across years or against reference sites. Generally, $\beta > 2$ indicates substantial turnover, while $\beta < 1.5$ suggests relatively homogeneous communities.