Among Site (Beta) Diversity

Student Name; Z620: Quantitative Biodiversity, Indiana University

06 February, 2015

OVERVIEW

In this excercise, we move beyond the investgation of within-site α-diversity. We will explore β-diversity, which is defined as the diveristy that occurs among sites. This requires that we examine the compositional similarity of assemblages that vary in space or time.

After completing this exercise you will know how to:

1. formally quantify β-diversity
2. visualize β-diversity with heatmaps, cluster analysis, and ordination
3. test hypotheses about β-diversity using multivariate statistics

Directions:

1. Change "Student Name" on line 3 (above) with your name.
2. Complete as much of the exercise as possible during class; what you do not complete in class will need to be done on your own outside of class.
3. Use the handout as a guide; it contains a more complete description of data sets along with the proper scripting needed to carry out the exercise.
4. Be sure to **answer the questions** in this exercise document; they also correspond to the handout. Space for your answer is provided in this document and indicated by the ">" character. If you need a second paragraph be sure to start the first line with ">".
5. Before you leave the classroom, **push** this file to your GitHub repo.
6. For homework, follow the directions at the bottom of this file.
7. When you are done, **Knit** the text and code into a PDF file.
8. After Knitting, please submit the completed exercise by creating a **pull request** via GitHub. Your pull request should include this file *beta_exercise.Rmd* and the PDF output of `Knitr` (*beta_exercise.pdf*).

1) R SETUP

Typically, the first thing you will do in either an R script or an RMarkdown file is setup your environment. This includes things such as setting the working directory and loading any packages that you will need.

In the R code chunk below, provide the code to:
1) clear your R environment,
2) print your current working directory,
3) set your working directory to your "*/Beta*" folder, and
4) load the `vegan` R package (be sure to install if needed).

2) LOADING DATA

A. Description of Data Set

To date, we have analyzed biodiversity data sets for freshwater zooplankton, tropical trees, and soil bacteria. In this exercise, we introduce a new dataset containing information on stream fish assemblages from the

Doubs river, which runs near the France-Switzerland boarder in the Jura Mountains. The data set (`doubs`) includes fish abundances, environmental variables, and spatial cooridinates for 30 sites. The data set has previously been used to demonstrate that fish communities can be good indicators of ecological zones in rivers and streams.

In the R code chunk below, do the following:
1. load the `doubs` data set, and
2. explore the structure of the data set using `str()`, `help()`, and the dollar sign ($) as needed.

Question 1: Describe some of the attributes of the `doubs` dataset.
a. How many objects are in `doubs`?
b. What types of data structures are contained in `doubs`?
c. What are the units of nitrate ("nit") in the stream water?
d. How many fish species are there in the `doubs` data set?

Answer 1a:
Answer 1b:
Answer 1c:
Answer 1d:

D. Visualizing the Doubs River Dataset

There is a wealth of information in the `doubs` dataset that can be used to address various issues related to β-diversity. For example, we might use the environmental or spatial data to develop or test a hypothesis. In the handout, we have generated two plots of the `doubs` fish data. The first plot shows fish richness at each site in the stream. The second plot shows the abundance of a particular fish species, Brown Trout (*Salmo trutta*), at each site in the stream.

Question 2: Answer the following questions based on the spatial patterns of richness (i.e., α-diversity) and Brown Trout (*Salmo trutta*) abundance in the Doubs River.
a. How does fish richness vary along the sampled reach of the Doubs River?
b. How does Brown Trout (*Salmo trutta*) abundance vary along the sampled reach of the Doubs River?
c. What do these patterns say about the limitations of using richness when examining patterns of biodiversity?

Answer 2a:
Answer 2b:
Answer 2c:

3) QUANTIFYING BETA-DIVERSITY BETWEEN TWO SAMPLES

There are various ways to quantify β-diversity. Perhaps one of the simplest metrics is **Whittaker's β-Diversity**, which was developed by Robert Whittaker (1960).

In the R code chunk below, do the following:
1. write a function for calculating Whittaker's β-diversity (i.e., β_w) and
2. use this function to calculate the β diversity between the fish assemblages in site 1 and 2 of the Doubs River.

Question 3: Using the `beta.w` function above, answer the following questions:
a. What is the β-diversity for fish assemblages sampled from site 1 and site 2 of the `doubs` data set?
b. Based on the formula for β_w, what does this value represent?

Answer 3a:
Answer 3b:

4) QUANTIFYING BETA-DIVERSITY FOR TWO OR MORE SAMPLES

Often, we often want to compare the diversity of more than just a pair of samples. For example, it would be nice to be able to compare fish assemblages for *all* of sites in the Doubs River. In this section we will estimate β-diversity for multiple samples. During this process, you will learn how to generate similarity and dissimilarity matrices for different data sets that will be needed for visualizing and quantifying β-diversity.

A. Introducing the Resemblance Matrix

In order to quantify β-diversity for more than two samples, we need to introduce our second primary ecological data structure: the **Resemblance Matrix**. In the context of biodiversity, a resemblance matrix is a data structure that calculates the pairwise **similarity** or **dissimiliarity** for all samples in a site-by-species matrix. The resemblance matrix can be generated from a site-by-species matrix containing incidence (presence-absence) data or abundance data.

Question 4: Answer the following questions about incidence-based measures of similarity:
a. What are the differences between Jaccard and Sørensen metrics?
b. When might you use one instead of the other?
c. In what situations would these incidence-based metrics of β-diversity fail?

Answer 4a:
Answer 4b:
Answer 4c:

E. Constructing the Resemblance Matrix

In the R code chunk below, do the following:
1. make a new object, "fish", containing the fish abundance data for the Doubs River,
2. remove any sites where no fish were observed,
3. construct a resemblance matrix based on Jaccard Similarity ("fish.dj"), and
4. construct a resemblance matrix based on Bray-Curtis Distance ("fish.db").

Question 5: Does the resemblance matrix (`fish.db`) represent similarity or dissimilarity? What information in the resemblance matrix led you to arrive at your answer?

Answer 5:

5) VISUALIZING BETA-DIVERSITY

A. Heatmaps

One way to visualize β-diversity is to plot the data in our resemblance matrix using a **heatmap**. Heatmaps are a two-dimensional, color representation of a data matrix.

In the R code chunk below, do the following:
1. create a custom color palette (feel free to make your own),
2. define the order of sites in the Doubs River, and
3. use the `levelplot()` function to create a heatmap of fish abundances in the Doubs River.

B. Cluster Analysis

Another common way to visualize β-diversity is through cluster analysis. Cluster analysis is an exploratory technique that assigns objects to groups based on their similarity to one another. In this exercise, we will use hierarchical clustering, specifically **Ward's Clustering**. Ward's Clustering (a.k.a., Ward's minimum variance method) is an agglomerative clustering technique based on the linear model criterion of least squares.

In the R code chunk below, do the following:
1. perform a cluster analysis using Ward's Clustering, and
2. plot your cluster analysis (use either `hclust` or `heatmap.2`).

Question 6: Based on cluster analyses and the introdutory plots that we generated after loading the data, develop a hypothesis for the `doubs` data set?

Answer 6:

C. Ordination

The primary aim of ordination is to represent multiple objects in a reduced number of orthogonal (i.e., independent) axes. The first axis of an ordination explains the most variation in the data set, followed by the second axis, then the third, and so on, where the total number of axes is less than or equal to the number of objects. Ordination plots are particularly useful for visualizing the similarity among objects. For example, in the context of β diversity, sites that are closer in ordination space have species assemblages that are more similar to one another than sites that are further apart in ordination space.

i. Principal Coordinates Analysis (PCoA)

In this exercise, we focus on Principal Coordinates Analysis (PCoA), which is sometimes referred to as metric multidimensional scaling. For a description of PCoA, please refer to the handout.

In the R code chunk below, do the following:
1. perform a Principal Coordinates Analysis,
2. calculate the variation explained by the first three axes in your ordination, and
3. make a plot that compares the eigenvalues of each axis with the expectations of the Kaiser-Guttman criterion and the Broken-Stick model.

Question 7: Based on the three criteria described above, does the PCoA do a good job of explaining variation in the `doubs` data set? Please justify.

Answer 7:

iii. Creating a PCoA Ordination Plot

In the R code chunk below, do the following:
1. plot the PCoA ordination,
2. label the sites as points using the Doubs River site number, and
3. identify influential species and add species corrdinates to PCoA plot.

In the R code chunk below, do the following:
1. identify influential species based on correlations along each PCoA axis (use a cutoff of 0.70), and
2. use a permutation test (999 permutations) to test the correlations of each species along each axis.

Question 9: Address the following questions about the ordination results of the `doubs` data set:
a. Generate a hypothesis about the grouping of sites in the Doubs River based on fish community composition.
b. Generate a hypothesis about which fish species are potential indicators of river quality.
c. Do the different approaches described in the ordination section agree or disagree? Explain.

Answer 9a:
Answer 9b:
Answer 9c:

6) HYPOTHESIS TESTING

The visualization tools that we just learned about (i.e., heatmaps, cluser analysis, and ordination) are powerful for exploratory analysis and for *generating* hypotheses. In this section we introduce some methods that are better suited for *testing* hypotheses and predictions related to β-diversity.

A. Multivariate Procedures for Categorical Designs

PERMANOVA stands for permutational multivariate analysis of variance (Anderson 2001). It is a multivariate analog to univariate ANOVA and has less restrictions than parametric multivariate analysis of variance (MANOVA). As the name suggests, it evaluates differences according to a specified model by randomly permuting the data. PERMANOVA can easily handle simple designs, but also accomodates nested and higher-order studies. In addition, it can deal with missing values and unbalanced designs. The PERMANOVA output is similar to the output of classic ANOVA; it includes (pseudo) F-tests, p-values, and R^2 values.

Earlier work done in the Doubs River suggested that the river can be broken into four distinct regions based on fish habitat quality. The first region (sites 1-14) has been rated as being "high quality". The second (sites 15 - 19) and fourth (sites 26 - 30) regions have been rated as being "moderate quality". And the third region (sites 20 - 25) has been rated has being "low quality".

In the R code chunk below, do the following:
1. create a vector the identifies fish habitat quality along the Doubs River, and
2. using that vector, conduct a PERMANOVA to test the hypothesis that habitat quality influences fish community composition.

Question 10: Based on the PERMANOVA results, evaluate the prediction that river quality influences fish community composition.

Answer 10:

B. Mutivariate Procedures for Continuous Designs

i. Mantel Test

A Mantel test is essentially a multivariate correlation analysis. It produces an r value that is analagous to the Pearson's correlation coefficient. In addition, it produces a p-value that is derived from the deviation of observed correlation to that of correlations derived from randomizations of the two matrices.

In the R code chunk below, do the following:
1. create distance matrices for both fish communities and environmental factors, and
2. use a Mantel test to determine if these matrices are correlated, and test the hypothesis that fish assemblages are correlated with stream environmental variables.

Question 11: What do the results from our Mantel test suggest about fish diversity and stream environmental conditions? How might this relate to your hypothesis about stream quality influencing fish communities?

Answer 11:

ii. Constrained Ordination

Another way we can test hypotheses with continuous data is to use **constrained ordination**, which is sometimes referred to as canonical ordination. Constrained ordination explores the relationships between two matrices: an **explanatory matrix** and a **response matrix**. Canonical correspondence analysis (CCA) and redundancy analysis (RDA) are two types of constrained ordination. These techniques are based on the linear model framework and thus can be used to formally test hypotheses.

In the R code chunk below, do the following:
1. create an environmental matrix consisting of the the water chemistry data included in the `doubs` data set,
2. conduct a Canonical Correspondance Analysis using the environmental matix as the explanatory variables,
3. use a permutation test to determine the significance of the constrained analysis,
4. use a permutaiton test to determine the correlation of each environmental factor on the constrained axes,
5. calculate the explained variation on the first and second constrained axes,
6. plot the constrained ordination results including labeled points for each site, and
7. add vectors that demonstrate the influence of each environmental factor the the constrained ordination (scale appropriately).

Question 12: Based on the CCA, what are the environmental variables that seem to be contributing to stream water quality for fish assemblages?

Answer 12:

7) HOMEWORK

1. We are going to revsit the soil bacteria data set that we introduced during the the α-diversity exercise. You may recall that 16S rRNA sequences were generated from replicate sites from four different land-use treatments (T1 = agriculture, T7 = grassland, DF = deciduous forest, and CF = coniferous forest). On top of this, we characterized bacteria composition from the different soils under experimentally manipulated dry vs. wet condtions. Thus, we have a 2 x 2 full-factorial design (with the exception of a missing sample from the DF land-use treatment). More background on this study can be found in Aanderud et al. 2015 (http://goo.gl/TRgISq)

Two files are available to you in the "~/QB2015_[username]/Beta/data"" directory:

a. Site-by-species matrix ("soilbacfull.txt")
b. A factor file that describes the treatments ("soil.factors.txt")

Use a combination of visualization and hypothesis-testing techniques described above to interpret the results from the exepriment in a β-diversity framework. Peform this analysis using an incidence-based distance metric and an abundance-based distance metric. Compare and contrast the outcomes.

2. Use Knitr to create a pdf of your completed alpha_exercise.Rmd document, push it to GitHub, and create a pull request. The due date for this assignment will be announced in class and/or canvas.