A community ecology lab using the R statistical environment – introduction to variation partitioning using the vegan package for R

Eric R. Sokol, Biological Sciences, Virginia Tech (sokole@vt.edu)
Updated 2014 Mar 24

Objectives

- Learn how to use the **vegetarian** package to conduct diversity partitioning
- Learn how to use Principal Coordinates of Neighbor Matrices as "spatial filters"
- Learn how to conduct a variation partitioning analysis on a community data matrix

What you need

- R and RStudio and the **vegetarian** and **vegan** packages installed on your computer. This lab is based on R v3.0.1 and vegan 2.0-7.

Introduction

What is a metacommunity?

A metacommunity is a collection of assemblages of interacting organisms (usually the same trophic status). Assemblages are connected by regional dispersal dynamics (e.g., immigration, emigration). We typically deal with data sets that represent metacommunities, for example, a site by species table of abundances can represent a metacommunity.

How is diversity measured?

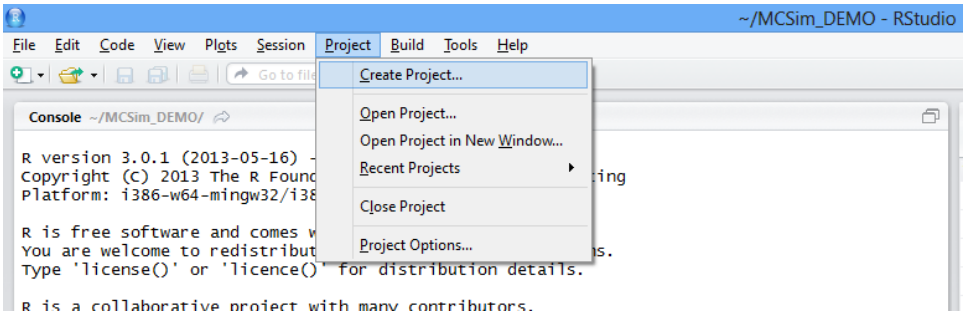
We will use the **vegetarian** package (Charney and Record 2012) to calculate alpha, beta, and gamma diversity using Jost's multiplicative method (Jost 2006, 2007, Jost et al. 2011, Chao et al. 2012) – dividing total metacommunity diversity into alpha, beta, and gamma components is known as diversity partitioning. Alpha diversity is a measure of mean local richness and/or evenness, gamma diversity is a measure of regional richness and/or evenness, and beta diversity is calculated as gamma / alpha and represents the number of "distinct" communities represented in the metacommunity. Beta-diversity can also be thought of as a measure of how much community composition tends to vary among sites in a metacommunity.

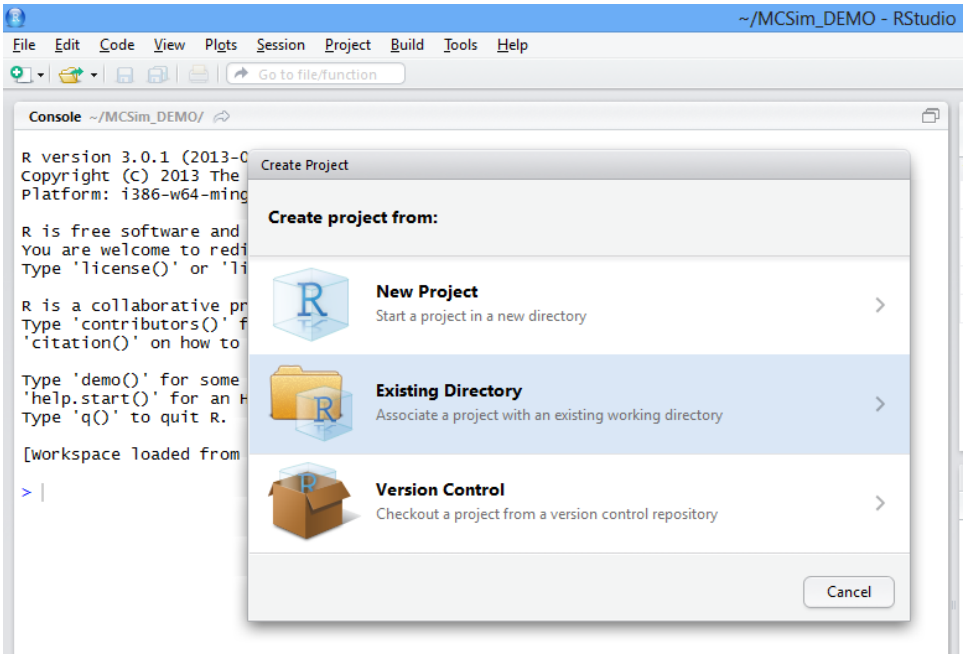
Beta-diversity can be analyzed using ***variation partitioning*** (Borcard et al. 1992, Peres-Neto et al. 2006), which determines the proportion of beta-diversity that is explained by environmental variation [E] and

spatial variation [S]. We will use the pcnm() function in the **vegan** package to create spatial variables that represent different scales of spatial heterogeneity – thus representing broad scale to fine scale spatial filters – to model [S]. Variation partitioning calculates the proportion of beta-diversity that is associated with [a] pure environmental variation, [b] spatially structured environmental variation, [c] pure spatial variation, and [d] unexplained beta-diversity. We will use the varpart() function in **vegan** for variation partitioning analyses.

Getting started

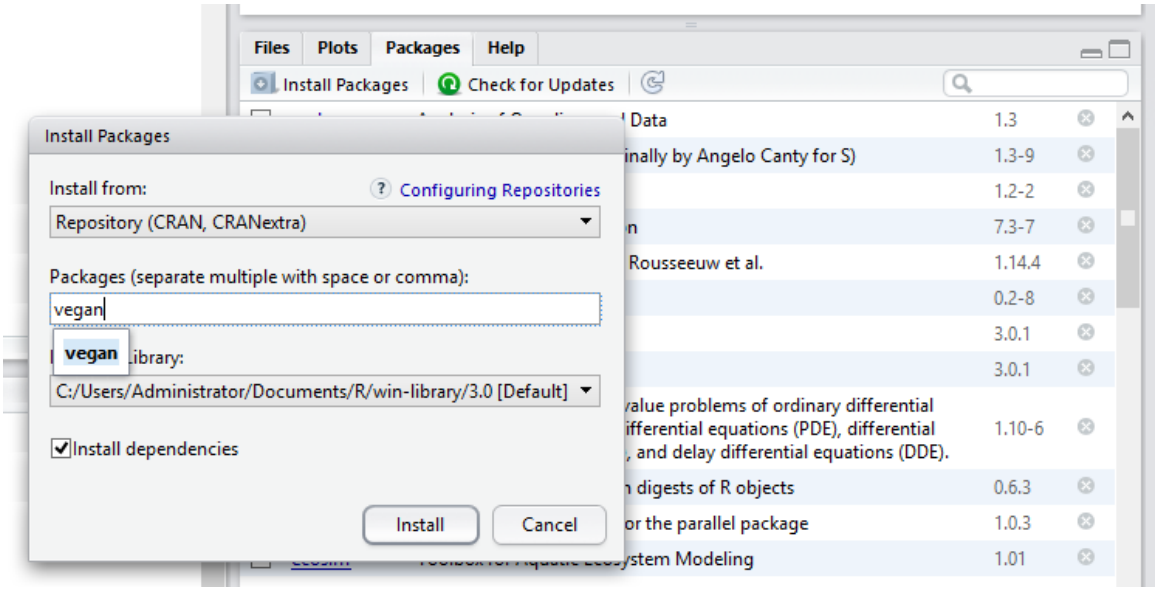
1. Create a directory for your project. I am using a windows machine, and thus I created a working directory called "DEMO" as a subdirectory in my Documents folder. Start your R session and set the working directory using setwd(). Alternatively, if you are using RStudio you can select Project→Create Project… →Existing Directory → Browse… (navigate to the folder you wish to use as a working directory, in my case "C:\Users\Administrator\Documents\DEMO").





2. Make sure you have downloaded and installed the **vegan** and **vegetarian** packages for R. If you are using RStudio, you can do this by selecting the "packages" tab, selecting "install Packages".

You will need an internet connection and you may need to choose a CRAN mirror (the server from which you download the packages from the CRAN).



3. For this tutorial, we will use data from a study of mite biodiversity that is included in the vegan package.

```
> require(vegan)
Loading required package: vegan
This is vegan 2.0-7
> data(mite)
> head(mite)
  Brachy PHTH HPAV RARD SSTR Protopl MEGR MPRO TVIE HMIN HMIN2 NPRA TVEL ONOV SUCT LCIL
1     17    5    5    3    2       1    4    2    2    1     4    1   17    4    9   50
2      2    7   16    0    6       0    4    2    0    0     1    3   21   27   12  138
3      4    3    1    1    2       0    3    0    0    0     6    3   20   17   10   89
4     23    7   10    2    2       0    4    0    1    2    10    0   18   47   17  108
5      5    8   13    9    0      13    0    0    0    3    14    3   32   43   27    5
6     19    7    5    9    3       2    3    0    0   20    16    2   13   38   39    3
  Oribatl1 Ceratoz1 PWIL Galumna1 Stgncrs2 HRUF Trhypch1 PPEL NCOR SLAT FSET Lepidzts
1        3        1    1        8        0    0        0    0    0    0    0        0
2        6        0    1        3        9    1        1    1    2    2    2        1
3        3        0    2        1        8    0        3    0    2    0    8        0
4       10        1    0        1        2    1        2    1    3    2   12        0
5        1        0    5        2        1    0        1    0    0    0   12        2
6        5        0    1        1        8    0        4    0    1    0   10        0
  Eupelops Miniglmn LRUG PLAG2 Ceratoz3 Oppiminu Trimalc2
1        0        0    0     0        0        0        0
2        0        0    0     0        0        0        0
3        0        0    0     0        0        0        0
4        0        0    0     0        0        0        0
5        0        0    0     0        0        0        0
6        0        0    0     0        0        0        0
```

The mite data frame included in the **vegan** package is site by species table of abundances of 35 mite species collected from 70 sites. The package also includes data frames with site xy coordinates and a table of environmental variables describing the habitat at each site.

```
> data(mite.xy)
```

```
> head(mite.xy)
    x   y
1 0.2 0.1
2 1.0 0.1
3 1.2 0.3
4 1.4 0.5
5 2.4 0.7
6 1.8 0.9
```

```
> data(mite.env)
> head(mite.env)
  SubsDens WatrCont Substrate Shrub    Topo
1    39.18   350.15   Sphagn1   Few Hummock
2    54.99   434.81    Litter   Few Hummock
3    46.07   371.72 Interface   Few Hummock
4    48.19   360.50   Sphagn1   Few Hummock
5    23.55   204.13   Sphagn1   Few Hummock
6    57.32   311.55   Sphagn1   Few Hummock
```

Diversity partitioning

4. First, we'll use the d() function in the **vegetarian** package (Charney and Record 2012) to conduct a diversity partitioning analysis to determine the magnitude of alpha, beta, and gamma diversity. Check out the help documentation. The d() function will calculate alpha, beta, or gamma diversity using multiplicative diversity partitioning based on Hill numbers (AKA "species equivalents") (Jost 2007, Chao et al. 2012). The parameter "q" represents the order of the Hill number. Order q = 0 Hill numbers will give you diversity measures based off of presence/absence data. So order q = 0 alpha diversity is simply the number of species at a site (richness). Order q = 1 Hill numbers are derived from the Shannon index, and order q = 2 are related to the Simpson index. Order q > 0 diversity indices reflect both richness and evenness.

```
> require(vegetarian)
Loading required package: vegetarian
> help(d)
```

Notice that the species count is equal to q = 0 gamma diversity:

```
> sum( colSums( mite ) > 0)
[1] 35
> d(mite, lev = "gamma", q = 0)
[1] 35
```

And q = 0 alpha diversity is the same as averaging the species count at each site:

```
> d(mite, lev = "alpha", q = 0)
[1] 15.11429
> mean( rowSums( mite > 0 ) )
[1] 15.11429
```

Calculate the q order 0, 1, and 2 beta diversity.

5. The mite.env data frame has a column of factors that identifies the dominant substrate type at each site. We can use the by() function along with the d() function to group sites by “Substrate” and calculate beta diversity for each group.

```
> by(
+   data = mite,
+   INDICES = list(mite.env$Substrate),
+   FUN = d,
+   lev = "beta",
+   q = 0
+   )
: Sphagn1
[1] 2.220812
------------------------------------------------------------------------
: Sphagn2
[1] 2.214286
------------------------------------------------------------------------
: Sphagn3
[1] 1
------------------------------------------------------------------------
: Sphagn4
[1] 1.212121
------------------------------------------------------------------------
: Litter
[1] 1.282051
------------------------------------------------------------------------
: Barepeat
[1] 1.4
------------------------------------------------------------------------
: Interface
[1] 2.344913
```

Calculate alpha and beta diversity for each habitat type. How might observation number affect the outcomes?

Variation partitioning

The objective of variation partitioning is to use constrained ordination, such as redundancy analysis (RDA), to identify how much beta diversity is explained by environmental variables [E] and how much is explained by spatial variables [S]. We then use the RDA models to calculate how much biodiversity is explained by pure environment [a], spatially structured environment [b], and pure spatial variation [c], and how much is unexplained [d]. The figure below is from Borcard (1992).

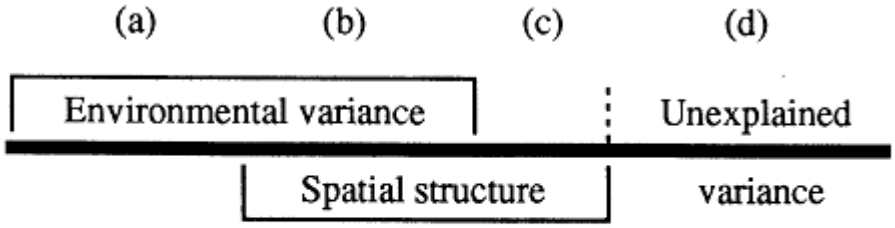


FIG. 2. Variation partitioning of a species data table, showing that fraction (b) is the intersection of the environmental and spatial components of the species variation.

Spatial variables

6. We will use the pcnm() function in vegan to calculate a Principal Coordinate of Neighbor Matrices analysis on the mite.xy data frame. We first calculate a distance matrix from the xy-coordinates of the sites from which mites were sampled:

```
> mite.dist <- dist( mite.xy )
```

(Note that the xy coordinates are on an arbitrary grid that is measured in meters. If you have Lat-Long data, you will need to convert it to a coordinate system that makes sense. I use the distm() function in the **geosphere** package to convert Lat-Long data to a distance matrix in meters.)

Then calculate a pcnm model from the distance matrix and extract the eigenvector scores for the positive eigenvectors. PCNM eigenvector 1 is a variable that represents the broadest scale spatial filter, PCNM 2 is the next finest scale spatial filter, PCNM 3 represents finer scale variation, and so on.

```
> mod.pcnm <- pcnm( dist(mite.dist) )
> vectors.pcnm <- data.frame(mod.pcnm$vectors)
```

To illustrate what we mean by "spatial filters" we will plot sites on a xy coordinate system with "bubble size" representing the sites' scores for the PCNM1 spatial filter.

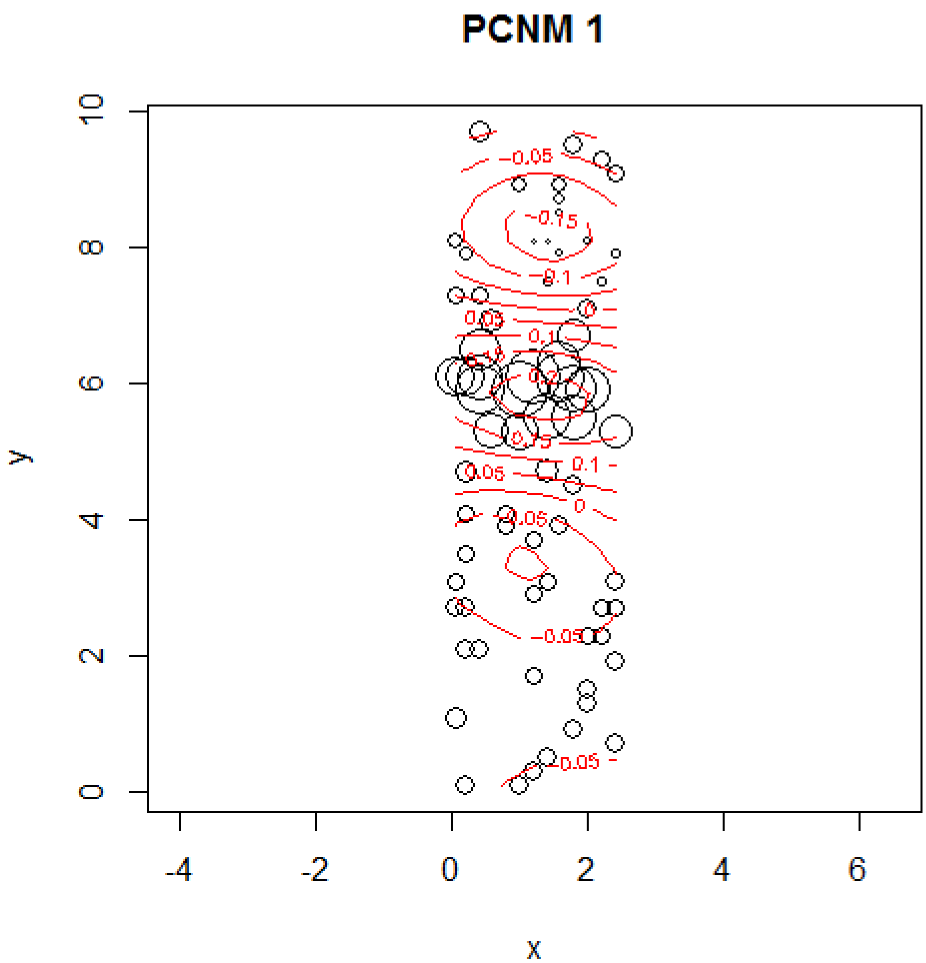
```
> ordisurf(mite.xy, scores(mod.pcnm, choi=1), bubble = 4, main = "PCNM 1")

Family: gaussian
Link function: identity
```

```
Formula:
y ~ s(x1, x2, k = knots)
<environment: 0x119c042c>

Estimated degrees of freedom:
8.93  total = 9.93

GCV score: 0.001557368
```



To compare, here’s PCNM13, which represents a “finer scale” spatial filter…

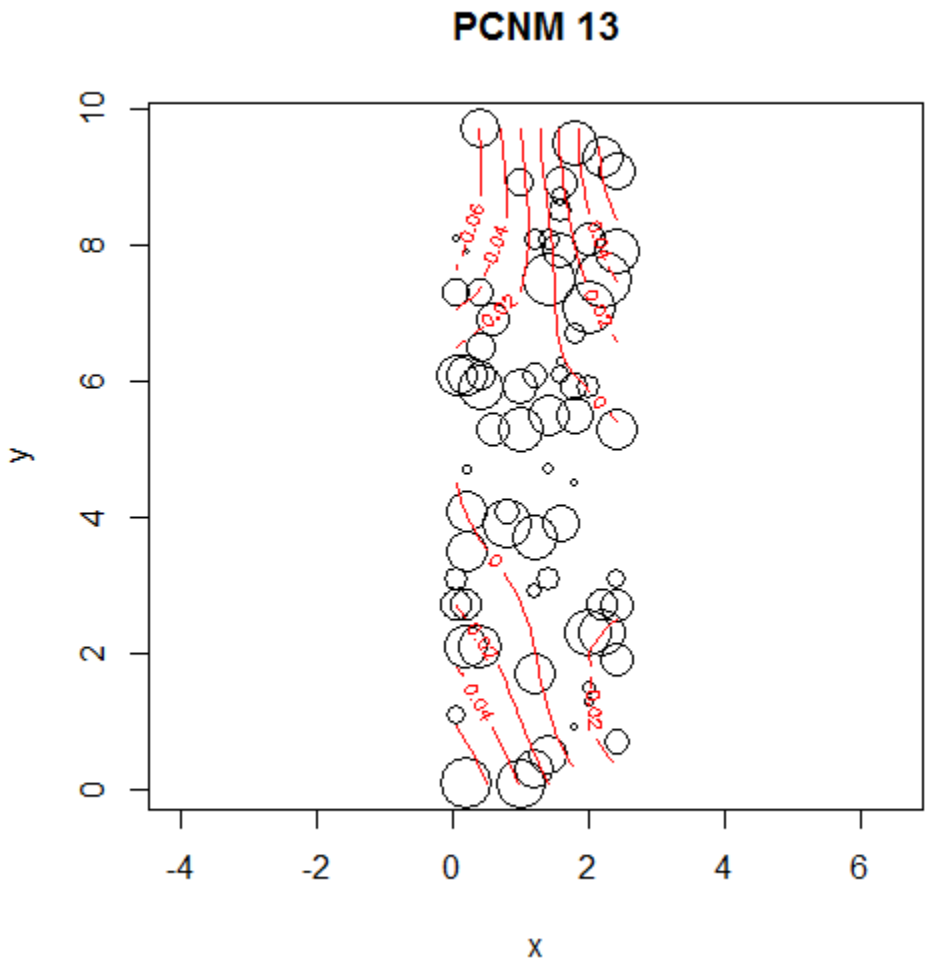
```
> ordisurf(mite.xy, scores(mod.pcnm, choi=13), bubble = 4, main = "PCNM 13")

Family: gaussian
Link function: identity

Formula:
y ~ s(x1, x2, k = knots)
<environment: 0x115254e4>
```

```
Estimated degrees of freedom:
5.82  total = 6.82

GCV score: 0.01529402
```



What is your interpretation of the spatial filters represented by PCNM1 and PCNM13?

7. The next step is to determine which spatial filters correspond with variation in mite community composition. For this analysis, we need to transform the mite count data using a Hellinger transformation.

```
> mite.hel <- decostand( mite, "hel")
```

8. Use stepwise model selection to determine which spatial variables correspond with variation in mite community composition. Note that we include the raw "x" and "y" coordinates in this analysis because they represent broad scale, linear spatial filters.

```
> d.space <- data.frame( mite.xy, vectors.pcnm)
> #combine all spatial variables, x, y, and PCNMs
>
> d.space.scaled <- data.frame( scale(d.space) )
> #center spatial variables on 0, and standardize


> # null model with intercept
> mod.0 <- rda( mite.hel ~ 1, data = d.space.scaled)
>
> # model with all spatial variables included
> mod.1 <- rda( mite.hel ~ ., data = d.space.scaled)
>
> #stepwise selection of the best model
> mod.best <- ordiR2step(mod.0, scope = mod.1 )


> summary(mod.best)

Call:
rda(formula = mite.hel ~ y + PCNM5 + PCNM1 + PCNM3 + PCNM4 +      PCNM12, dat
a = d.space.scaled)

Partitioning of variance:
              Inertia Proportion
Total          0.3943     1.0000
Constrained    0.1839     0.4664
Unconstrained  0.2104     0.5336
```

Here's how to extract the names of the significant spatial variables:

```
> S.keepers <- names( mod.best$terminfo$ordered )
> S.keepers
[1] "y"      "PCNM5"  "PCNM1"  "PCNM3"  "PCNM4"  "PCNM12"
```

Spatial variables

9. Next we look at the environmental variables in mite.env. “SubsDens” and “WatrCont” are the only two that columns that are numeric, so we’ll use those two. First check to see if they need to be log transformed using the Shapiro test, then use model selection to determine which environmental variables are significantly related to variation in mite community composition.

```
> shapiro.test( mite.env$SubsDens )
        Shapiro-Wilk normality test
data:  mite.env$SubsDens
W = 0.9369, p-value = 0.001581

> shapiro.test( log( ite.env$SubsDens ) )
        Shapiro-Wilk normality test
data:  log(mite.env$SubsDens)
W = 0.9842, p-value = 0.5219
```

Looks like “SubsDens” should be log transformed

```
> shapiro.test( mite.env$WatrCont)
        Shapiro-Wilk normality test
data:  mite.env$WatrCont
W = 0.9873, p-value = 0.701

> shapiro.test( log( mite.env$WatrCont ) )
        Shapiro-Wilk normality test
data:  log(mite.env$WatrCont)
W = 0.9616, p-value = 0.03075
```

Looks like “WatrCont” is good as is.

Now make a data frame with the two environmental variables, log transform “SubsDens”, and center and scale the variables.

```
> d.env <- mite.env[, c("SubsDens", "WatrCont")]
> d.env$SubsDens <- log( d.env$SubsDens )
> d.env.scaled <- data.frame( scale(d.env) )
```

Here's the model selection to determine which environmental variables are significantly correlated with variation in mite community composition:

```
> # null model with intercept
> mod.0 <- rda(mite.hel ~ 1, data = d.env.scaled)
>
> # model with all spatial variables included
> mod.1 <- rda(mite.hel ~ ., data = d.env.scaled)
>
> #stepwise selection of the best model
> mod.best <- ordiR2step(mod.0, scope = mod.1 )
> E.keepers <- names(mod.best$terminfo$ordered)
> E.keepers
[1] "WatrCont" "SubsDens"
```

…And it looks like they're both significantly related to variation in mite community composition

10. Now we put it all together. Our "response variable" is the matrix of Hellinger transformed mite abundances (mite.hel), our "spatial variable" [S] is a matrix of significant spatial variables, and our "environmental variable" [E] is a matrix of significant environmental variables. We will use the varpart() function from vegan to determine how much variation in mite community composition is due to pure environmental variation [a], spatially structured environmental variation [b], pure spatial variation [S], and how much is unexplained by the variables we have in this analysis [d].

```
> d.E <- d.env.scaled[,E.keepers]
> d.S <- d.space.scaled[,S.keepers]
> mod.varpart <- varpart(mite.hel, d.E, d.S)
> mod.varpart

Partition of variation in RDA

Call: varpart(Y = mite.hel, X = d.E, d.S)

Explanatory tables:
X1:  d.E
X2:  d.S

No. of explanatory tables: 2
Total variation (SS): 27.205
            Variance: 0.39428
No. of observations: 70

Partition table:
                     Df R.squared Adj.R.squared Testable
[a+b] = X1            2   0.32707       0.30699     TRUE
[b+c] = X2            6   0.46636       0.41553     TRUE
[a+b+c] = X1+X2       8   0.52680       0.46475     TRUE
Individual fractions
[a] = X1|X2           2                 0.04921     TRUE
[b]                   0                 0.25777    FALSE
[c] = X2|X1           6                 0.15776     TRUE
[d] = Residuals                         0.53525    FALSE
---
Use function 'rda' to test significance of fractions of interest
```

How would you interpret this output?

Can you figure out how to test if each partition is significant? (Hint: you can calculate partial rda models and then use anova() to calculate the p-value)

Literature Cited

Borcard, D., P. Legendre, and P. Drapeau. 1992. Partialling out the spatial component of ecological variation. Ecology 73:1045–1055.

Chao, A., C.-H. Chiu, and T. C. Hsieh. 2012. Proposing a resolution to debates on diversity partitioning. Ecology 93:2037–2051.

Charney, N., and S. Record. 2012. vegetarian: Jost Diversity Measures for Community Data.

Jost, L. 2006. Entropy and diversity. Oikos 113:363–375.

Jost, L. 2007. Partitioning diversity into independent alpha and beta components. Ecology 88:2427–2439.

Jost, L., A. Chao, and R. L. Chazdon. 2011. Compositional similarity and beta diversity. Pages 68–84 *in* A. E. Magurran and B. J. McGill, editors. Biological Diversity: Frontiers in Measurement and Assessment. Oxford University Press, Oxford, UK.

Peres-Neto, P. R., P. Legendre, S. Dray, and D. Borcard. 2006. Variation partitioning of species data matrices - estimation and comparison of fractions. Ecology 87:2614–2625.