Assignment: Spatial Diversity

Erik Parker; Z620: Quantitative Biodiversity, Indiana University

06 February, 2017

OVERVIEW

This assignment will emphasize primary concepts and patterns associated with spatial diversity, while using R as a Geographic Information Systems (GIS) environment. Complete the assignment by refering to examples in the handout.

After completing this assignment you will be able to:
1. Begin using R as a geographical information systems (GIS) environment.
2. Identify primary concepts and patterns of spatial diversity.
3. Examine effects of geographic distance on community similarity.
4. Generate simulated spatial data.

Directions:

1. Change "Student Name" on line 3 (above) with your name.
2. Complete as much of the assignment 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 assignment.
4. Be sure to **answer the questions** in this assignment document. 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. When you are done wit the assignment, **Knit** the text and code into an html file.
7. After Knitting, please submit the completed assignment by creating a **pull request** via GitHub. Your pull request should include this file *spatial_assignment.Rmd* and the html output of `Knitr` (*spatial_assignment.html*).

1) R SETUP

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 "*/Week4-Spatial*" folder, and

```
rm(list=ls())
getwd()
setwd("./")
```

2) LOADING R PACKAGES

In the R code chunk below, do the following:

1. Install and/or load the following packages: `vegan`, `sp`, `gstat`, `raster`, `RgoogleMaps`, `maptools`, `rgdal`, `simba`, `gplots`, `rgeos`

Question 1: What are the packages `simba`, `sp`, and `rgdal` used for?

> ***Answer 1***:
> Simba: used for reshaping species lists into matrices, and back again, and also for calculation of similarty measures when working with presence/absence data. Sp: provides functions for importing, manipulating, and exporting spatial data. Rgdal: Access to a geospatial data abstraction library.

3) LOADING DATA

In the R code chunk below, use the example in the handout to do the following:

1. Load the Site-by-Species matrix for the Indiana ponds datasets: BrownCoData/SiteBySpecies.csv
2. Load the Environmental data matrix: BrownCoData/20130801_PondDataMod.csv
3. Assign the operational taxonomic units (OTUs) to a variable 'otu.names'
4. Remove the first column (i.e., site names) from the OTU matrix.

```
Ponds <- read.table("BrownCoData/20130801_PondDataMod.csv", head = TRUE, sep = ",")
OTUs <- read.csv("BrownCoData/SiteBySpecies.csv", head = TRUE, sep = ",")
otu.names <- names(OTUs)
OTUs <- as.data.frame(OTUs[-1])

dim(OTUs)

S.Obs <- function(x = ""){
  rowSums(x > 0) * 1
}
S.Obs(OTUs)
```

Question 2a: How many sites and OTUs are in the SiteBySpecies matrix?

> ***Answer 2a***:
> There are 51 sites and 16383 OTUs in the site by species matrix.

Question 2b: What is the greatest species richness found among sites?

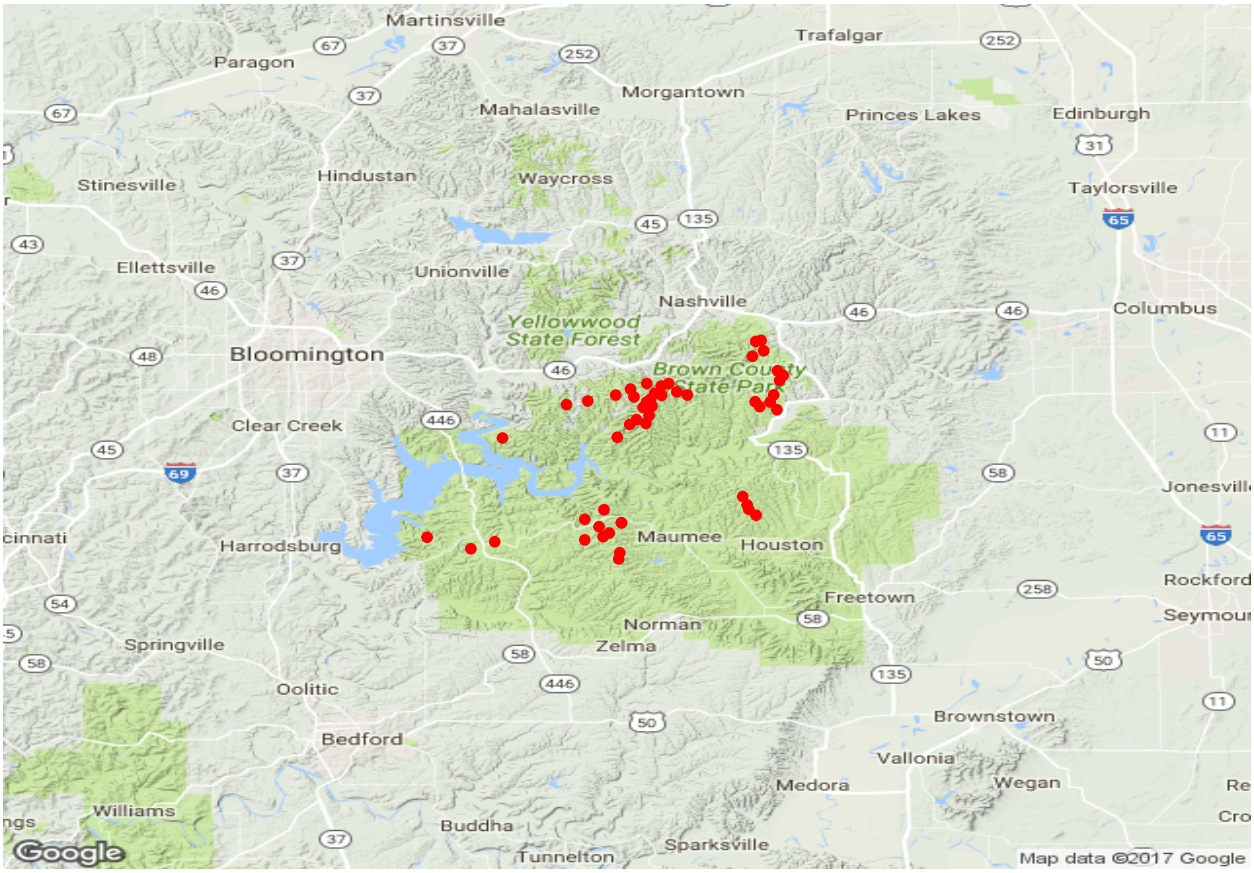
> ***Answer 2b***:
> The greatest species richness found among sites is 3259 for the first site.

4) GENERATE MAPS

In the R code chunk below, do the following:

1. Using the example in the handout, visualize the spatial distribution of our samples with a basic map in RStudio using the `GetMap` function in the package `RgoogleMaps`. This map will be centered on Brown County, Indiana (39.1 latitude, -86.3 longitude).

```
lats <- as.numeric(Ponds[,3])
lons <- as.numeric(Ponds[,4])

newmap <- GetMap(center = c(39.1,-86.3), zoom = 10,
                 destfile = "PondsMap.png", maptype = "terrain")
PlotOnStaticMap(newmap, zoom = 10, cex = 2, col = 'blue')
PlotOnStaticMap(newmap, lats, lons, cex = 1, pch = 20, col = 'red', add = TRUE)
```



Question 3: Briefly describe the geographical layout of our sites.

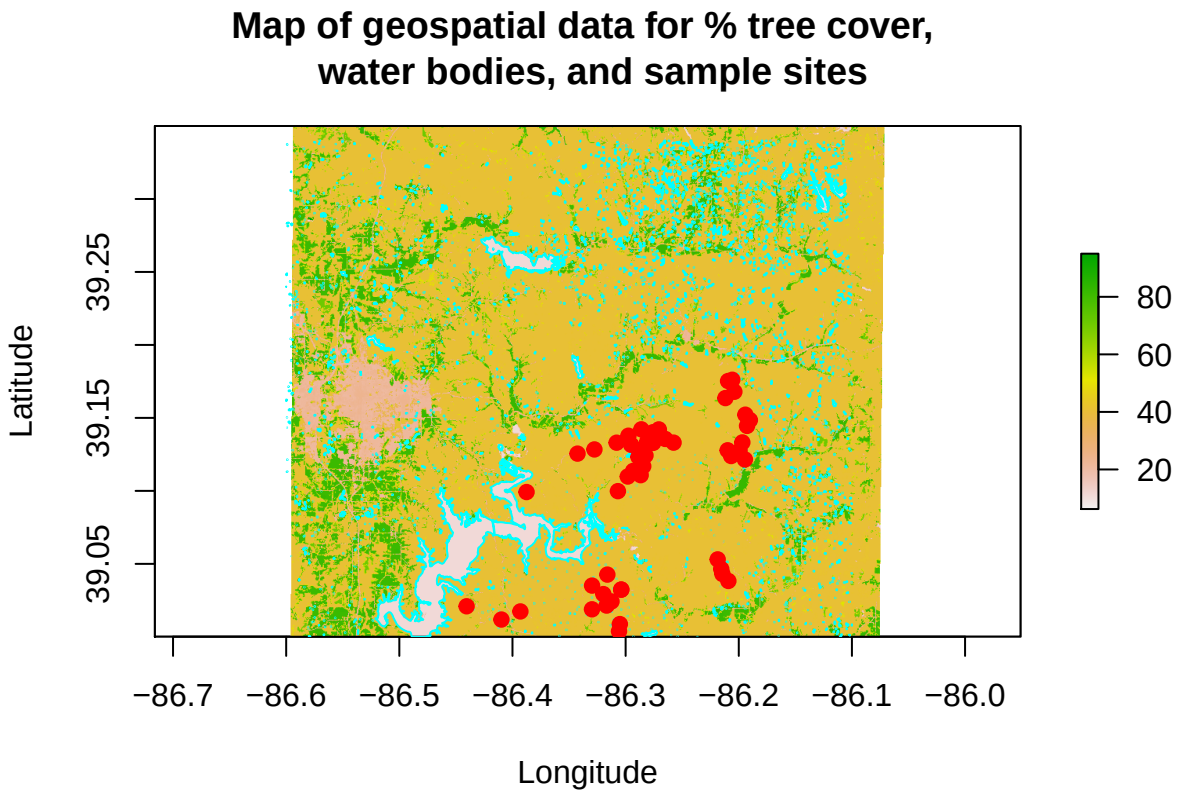
> ***Answer 3***:
> The sites are all located around Brown county state park, with most lying the the north east of lake monroe, and a few to the south and south east of the lake. The sites are also for the most part arranged in a few distinct clusters with little distance between them (within a cluster), and very few sites stand alone not as part of a larger cluster.

In the R code chunk below, do the following:

1. Using the example in the handout, build a map by combining lat-long data from our ponds with land cover data and data on the locations and shapes of surrounding water bodies.

```
# 1. Import TreeCover.tif as a raster file.

Tree.Cover <- raster("TreeCover/TreeCover.tif")

# 2. Plot the % tree cover data

plot(Tree.Cover, xlab = 'Longitude', ylab = 'Latitude',
     main = 'Map of geospatial data for % tree cover, \n water bodies, and sample sites')

# 3. Import water bodies as a shapefile.

Water.Bodies <- readShapeSpatial("water/water.shp")

# 4. Plot the water bodies around our study area, i.e., Monroe County.
```

```r
plot(Water.Bodies, border = 'cyan', axes = TRUE, add = TRUE)

# 5. Convert lat-long data for ponds to georeferenced points.

Refuge.Ponds <- SpatialPoints(cbind(lons, lats))

# 6. Plot the refuge pond locations

plot(Refuge.Ponds, line = 'r', col = 'red', pch = 20, cex = 1.5, add = TRUE)
```



Question 4a: What are datums and projections?

> ***Answer 4a***:
> A datum in this case is a particular model for Earth's shape, while a projection is the way in which coordinates on a sphere are projected onto a 2-D surface.

5) UNDERSTANDING SPATIAL AUTOCORRELATION

Question 5: In your own words, explain the concept of spatial autocorrelation.

> ***Answer 5***: Sites that are close together are going to look more similar than sites further from one another in general because closer sites experience more similar environmental conditions than do distant sites, and it is more likely that species will disperse over shorter distances to other closer sites. In general, close sites are more similar than distant sites.

6) EXAMINING DISTANCE-DECAY

Question 6: In your own words, explain what a distance decay pattern is and what it reveals.

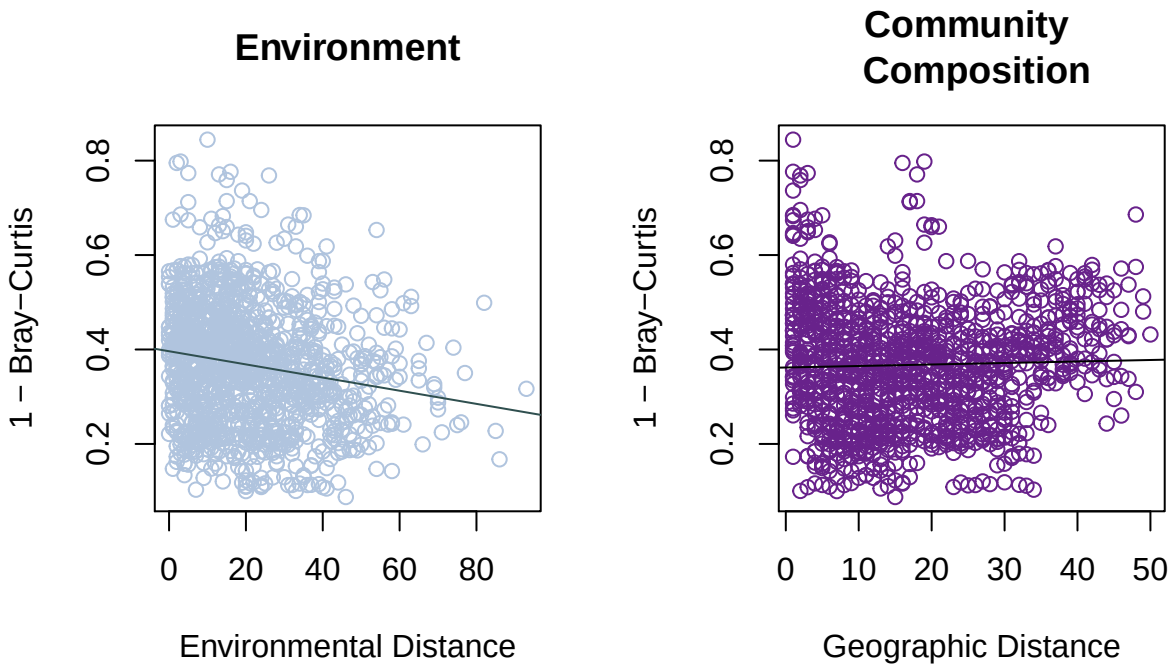
Answer 6: A distance decay pattern is a way to represent the idea of spatial autocorrelation by showing that as distance between sites increases, similarity decreases.

In the R code chunk below, do the following:

1. Generate the distance decay relationship for bacterial communities of our refuge ponds and for some of the environmental variables that were measured. Note: You will need to use some of the data transformations within the *semivariogram* section of the handout.

```
# 1) Calculate Bray-Curtis similarity between plots using the `vegdist()` function

comm.dist <- 1 - vegdist(OTUs)

# 2) Assign UTM lattitude and longitude data to 'lats' and 'lons' variables

xy <- data.frame(env = Ponds$TDS, pond.name = Ponds$Sample_ID, lats = Ponds$lat, lons = Ponds$long)

coordinates(xy) <- ~lats+lons

proj4string(xy) <- CRS("+proj=longlat +datum=NAD83")

UTM <- spTransform(xy, CRS("+proj=utm +zone=51 +ellps=WGS84"))
UTM <- as.data.frame(UTM)

xy$lats_utm <- UTM[,2]
xy$lons_utm <- UTM[,3]

lats <- as.numeric(xy$lats_utm)
lons <- as.numeric(xy$lons_utm)

# 3) Calculate geographic distance between plots and assign to the variable 'coord.dist'

coord.dist <- dist(as.matrix(lats,lons))

# 4) Transform environmental data to numeric type, and assign to variable 'x1'

x1 <- as.numeric(Ponds$"SpC")

# 5) Using the `vegdist()` function in `simba`, calculate the Euclidean distance between the plots for

env.dist <- vegdist(x1, "euclidian")

# 6) Transform all distance matrices into database format using the `liste()` function in `simba`:

comm.dist.ls <- liste(comm.dist, entry = "comm")
env.dist.ls <- liste(env.dist, entry = "env")
coord.dist.ls <- liste(coord.dist, entry = "dist")

# 7) Create a data frame containing similarity of the environment and similarity of community.

df <- data.frame(coord.dist.ls, env.dist.ls[,3], comm.dist.ls[,3])

# 8) Attach the columns labels 'env' and 'struc' to the dataframe you just made.

names(df)[4:5] <- c("env", "struc")
```

```r
attach(df)

# 9) After setting the plot parameters, plot the distance-decay relationships, with regression lines in

par(mfrow = c(1,2), pty = "s")
plot(env, struc, xlab = "Environmental Distance", ylab = "1 - Bray-Curtis", main = "Environment", col =
OLS <- lm(struc ~ env)
abline(OLS, col = 'darkslategray' )

plot(dist, struc, xlab = "Geographic Distance", ylab = "1 - Bray-Curtis",
     main = "Community \n Composition", col = 'darkorchid4')
OLS <- lm(struc ~ dist)
abline(OLS, col = 'black')
```




```r
# 10) Use `simba` to calculates the difference in slope or intercept of two regression lines

diffslope(env, struc, dist, struc)
```

Question 7: What can you conclude about community similarity with regars to environmental distance and geographic distance?

> ***Answer 7***: From the first plot above we can see that as environmental distance increases, 1 - the bray-curtis dissimilarity decreases (so, similarity decreases and dissimilarity increases!). This is contrasted by the second plot where it is seen that as geographic distance increases, similarity actually increases slightly. This geographic distance plot initially seems like an odd result, but from the map we generated earlier, perhaps the similarity increases with geographic distance because the majority of the sites seem to be in roughly similar habitats (the forest), and maybe there are a few sites that had very different environmental measures from the rest of the sample, that were close to a lot of other points. If this was the case, I can imagine that perhaps it skewed the data a bit, and made it so that a lot of sites had some of their closest comparison sites as the radiacally different ones, while the majority of sites (including those more geographically distant) were quite similar to eachother. This is a total guess, but seems to be supported by the data seen in the environmental distance plot (decreasing relationship, and many points clustered around

the left side of low difference).

7) EXAMINING SPECIES SPATIAL ABUNDANCE DISTRIBUTIONS

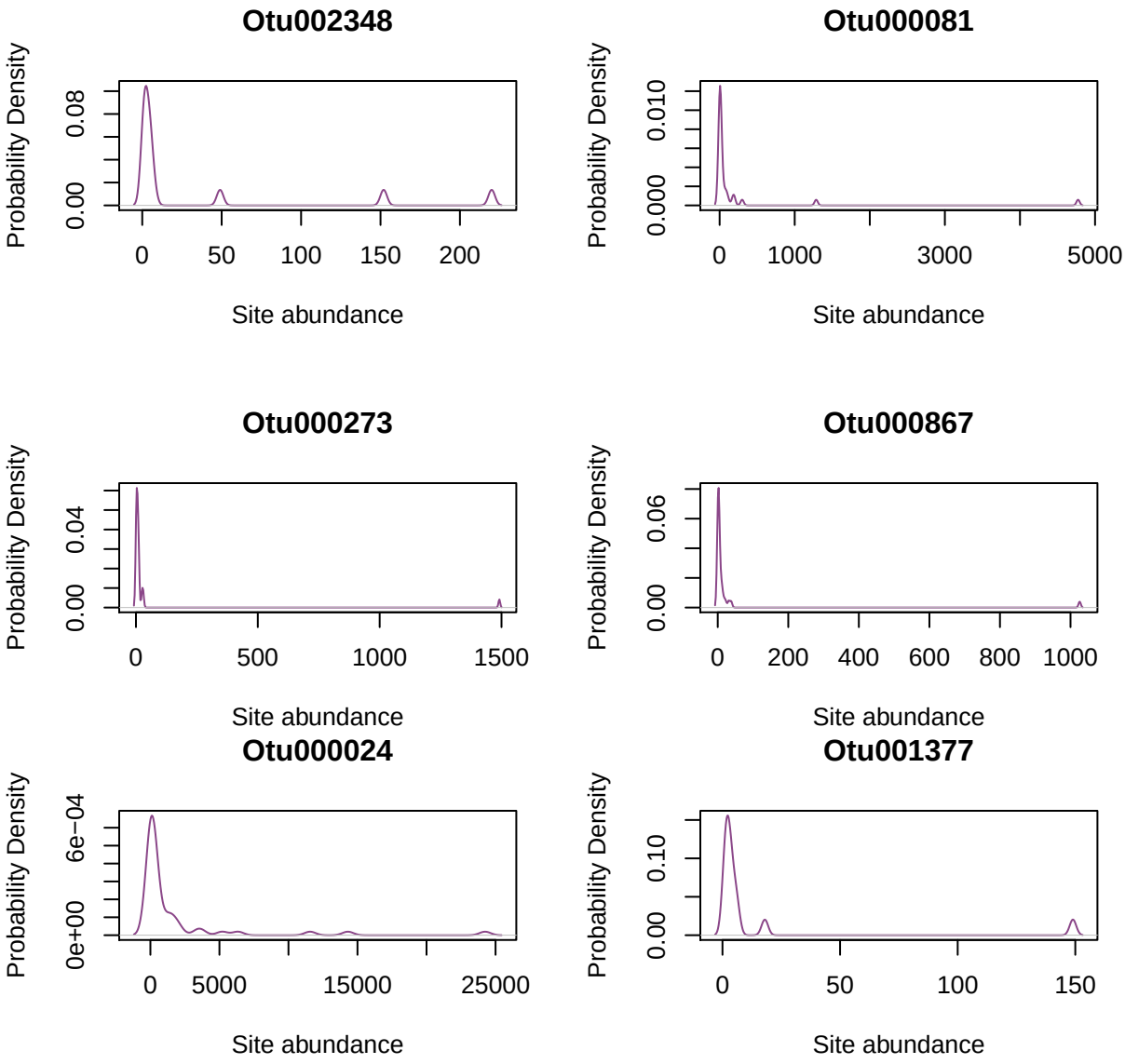
Question 8: In your own words, explain the species spatial abundance distribution and what it reveals.

> ***Answer 8***: The SSAD is a visualization of the probability of finding a certain species at a given abundance throughout its range. This function reveals general trends in a given species' (taxa's?) abundance patterns - how often are they rare vs. common in the environment / at about what abundance level can you most expect to find them when sampling?

In the R code chunk below, do the following:

1. Define a function that will generate the SSAD for a given OTU.
2. Draw six OTUs at random from the IN ponds dataset and and plot their SSADs as kernel density curves. Use **while loops** and **if** statements to accomplish this.

```
# 1. Define an SSAD function

ssad <- function(x){
  ad <- c(2,2)
  ad <- OTUs[, otu]
  ad = as.vector(t(x = ad))
  ad = ad[ad > 0]
}

# 2. Set plot parameters

par(mfrow = c(2, 2))

# 3. Declare a counter variable

ct <- 0

# 4. Write a while loop to plot the SSADs of six species chosen at random

while (ct < 6){
  otu <- sample(1:length(OTUs), 1)
  ad <- ssad(otu)
  if (length(ad) > 10 & sum(ad > 100)){
    ct <- ct+1
    plot(density(ad), col = 'orchid4', xlab = 'Site abundance',
         ylab = 'Probability Density', main = otu.names[otu])
  }
}
```



8) UNDERSTANDING SPATIAL SCALE

Many patterns of biodiversity relate to spatial scale.

Question 9: List, describe, and give examples of the two main aspects of spatial scale

> ***Answer 9***: The two main aspects of spatial scale, extent and grain, are the greatest distance considered in a given observation or study, and the smallest unit by which the extent is measured, respectively. Examples of these, again respectively, would be the IU campus, and 10X10 meter plots if we were quantifying arthropod diversity across the entire campus by one 10mˆ2 plot at a time.

9) CONSTRUCTING THE SPECIES-AREA RELATIONSHIP

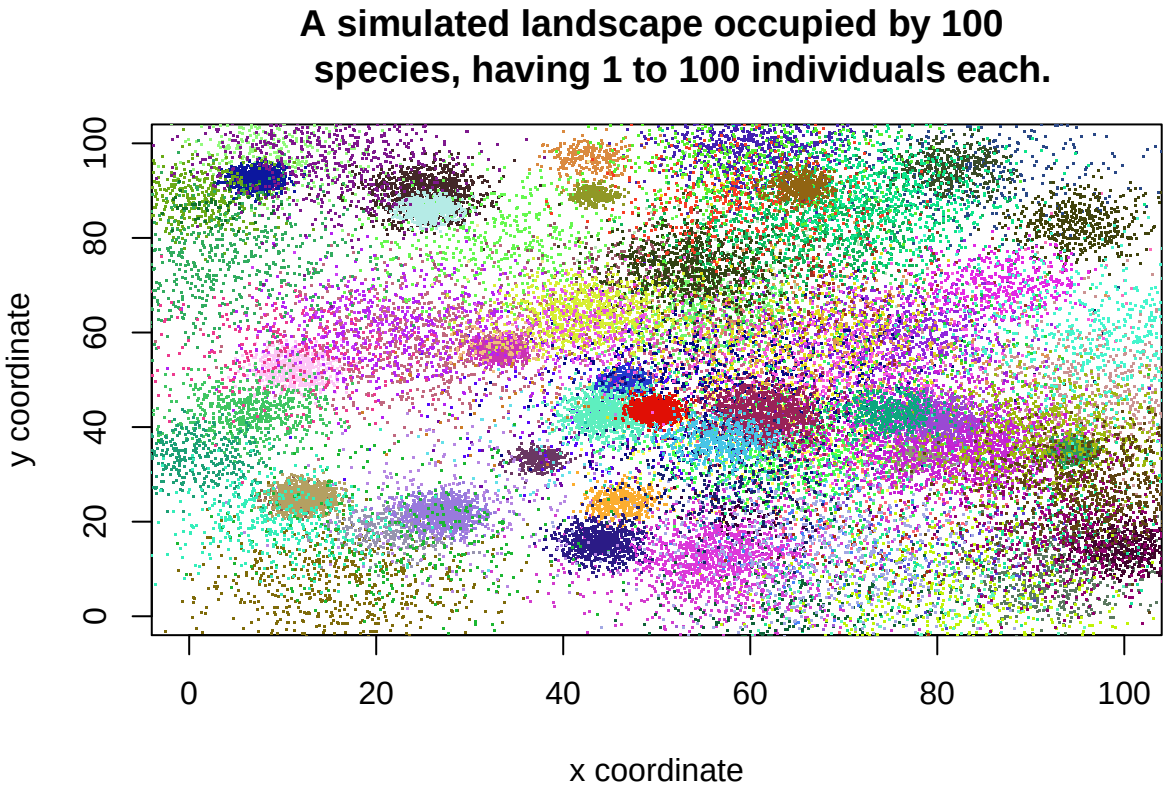
Question 10: In your own words, describe the species-area relationship.

Answer 10: In general the species-area relationship is the observation that in many cases larger areas hold larger numbers of species, and that as area increases so does the number of species following a particular log / log relationship. This relationship has been long represented as a power law where rates of increase in richness are linear with respect to space.

In the R code chunk below, provide the code to:

1. Simulate the spatial distribution of a community with 100 species, letting each species have between 1 and 1,000 individuals.

```
# 1. Declare variables to hold simulated community and species information

community <- c()
species <- c()

# 2. Populate the simulated landscape

plot(0,0,col='white',xlim=c(0,100),ylim=c(0,100),
     xlab = 'x coordinate', ylab = 'y coordinate',
     main = 'A simulated landscape occupied by 100
     species, having 1 to 100 individuals each.')

while (length(community) < 100){
  std <- runif(1,1,10)
  ab <- sample(1000,1)
  x <- rnorm(ab, mean = runif(1,0,100), sd = std)
  y <- rnorm(ab, mean = runif(1,0,100), sd = std)
  color <- c(rgb(runif(1), runif(1), runif(1)))
  points(x,y,pch = ".", col = color)
  species <- list(x,y, color)
  community[[length(community)+1]] <- species
}
```



While consult the handout for assistance, in the R chunk below, provide the code to:

1. Use a nested design to examine the SAR of our simulated community.
2. Plot the SAR and regression line.

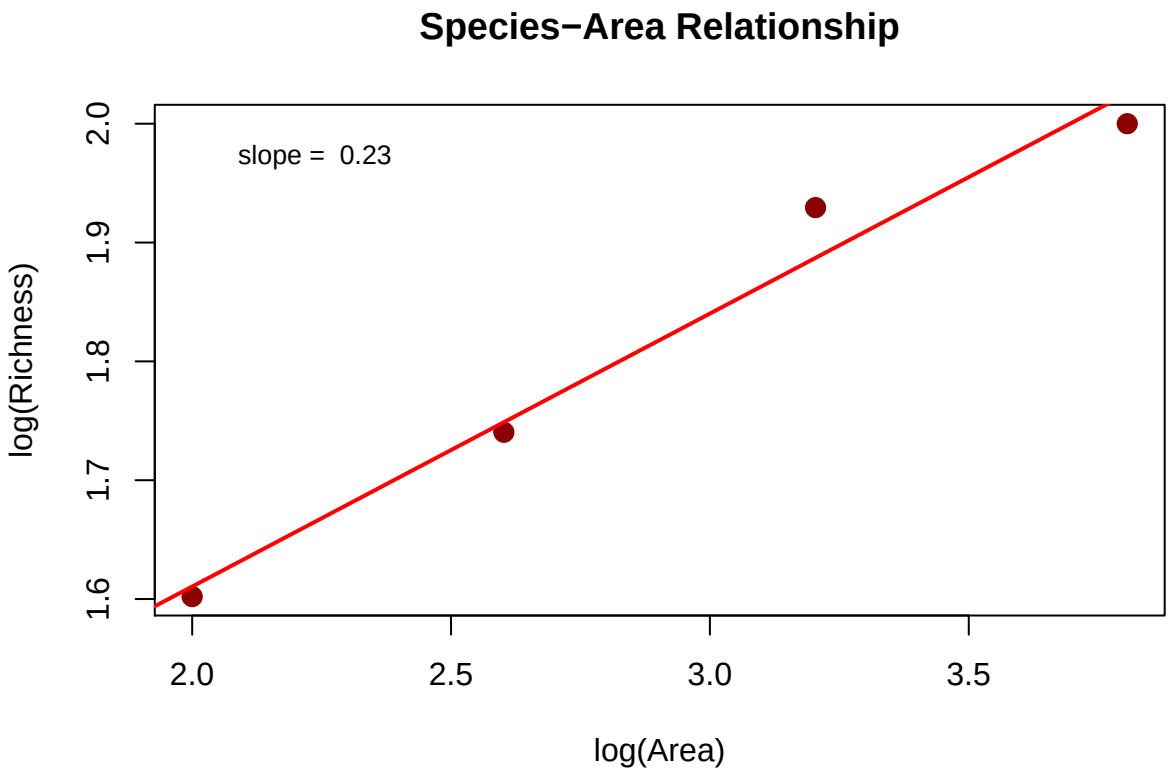
```
# 1. Declare the spatial extent and lists to hold species richness and area data

lim <- 10
S.list <- c()
A.list <- c()

# 2. Construct a 'while' loop and 'for' loop combination to quantify the numbers of species for progres

while (lim <= 100){
  S <- 0
  for (sp in community){
    xs <- sp[[1]]
    ys <- sp[[2]]
    sp.name <- sp[[3]]
    xy.coords <- cbind(xs, ys)
    for (xy in xy.coords){
      if (max(xy) <= lim){
        S <- S + 1
        break
      }
    }
  }
  S.list <- c(S.list, log10(S))
  A.list <- c(A.list, log10(lim^2))
  lim <- lim * 2
```

```
}

results <- lm(S.list ~ A.list)

# 3. Be sure to log10-transform the richness and area data
```

In the R code chunk below, provide the code to:

1. Plot the richness and area data as a scatter plot.
2. Calculate and plot the regression line
3. Add a legend for the z-value (i.e., slope of the SAR)

```
results <- lm(S.list ~ A.list)
plot(A.list, S.list, col = 'dark red', pch = 20, cex = 2,
     main = "Species-Area Relationship",
     xlab = 'log(Area)', ylab = 'log(Richness)')

abline(results, col = 'red', lwd = 2)

int <- round(results[[1]][[1]],2)

z <- round(results[[1]][[2]],2)
legend(x = 2, y = 2, paste(c('slope = ', z), collapse = " "), cex = 0.8,
       box.lty = 0)
```



Question 10a: Describe how richness relates to area in our simulated data by interpreting the slope of the SAR.

> ***Answer 10a***: From the slope of our SAR calculated above, we can see that as area increases, from about 100 to 1000, species richness increases from about 45 to about 70. So, richness definitely increases as area does, and they do so with a linear relationship when placed on a log

scale.

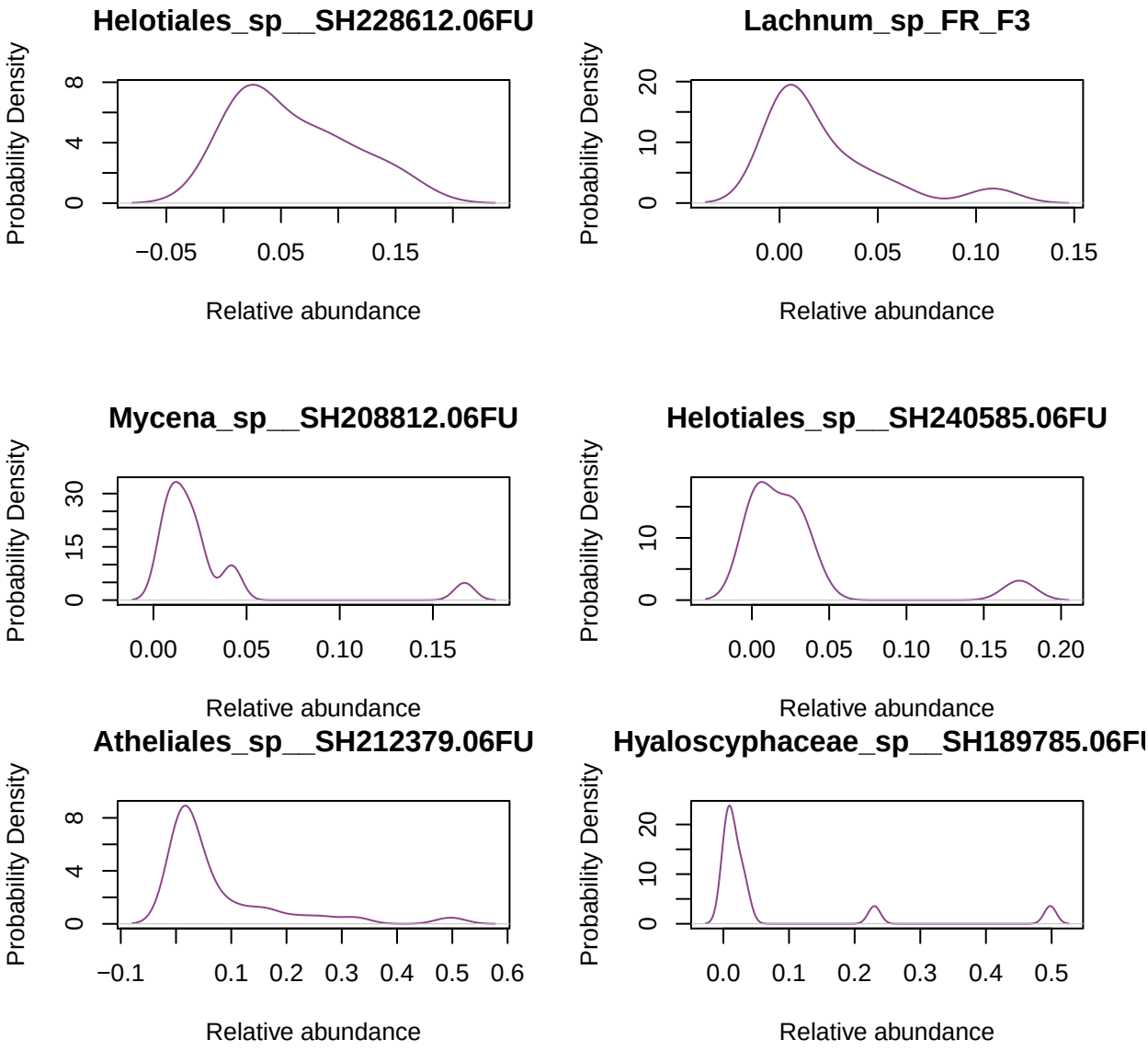
Question 10b: What does the y-intercept of the SAR represent?

> ***Answer 10b***: The y-intercept of the SAR represents the lowest level of species richness we would expect to find in the smallest area habitat possible, given our model parameters.

SYNTHESIS

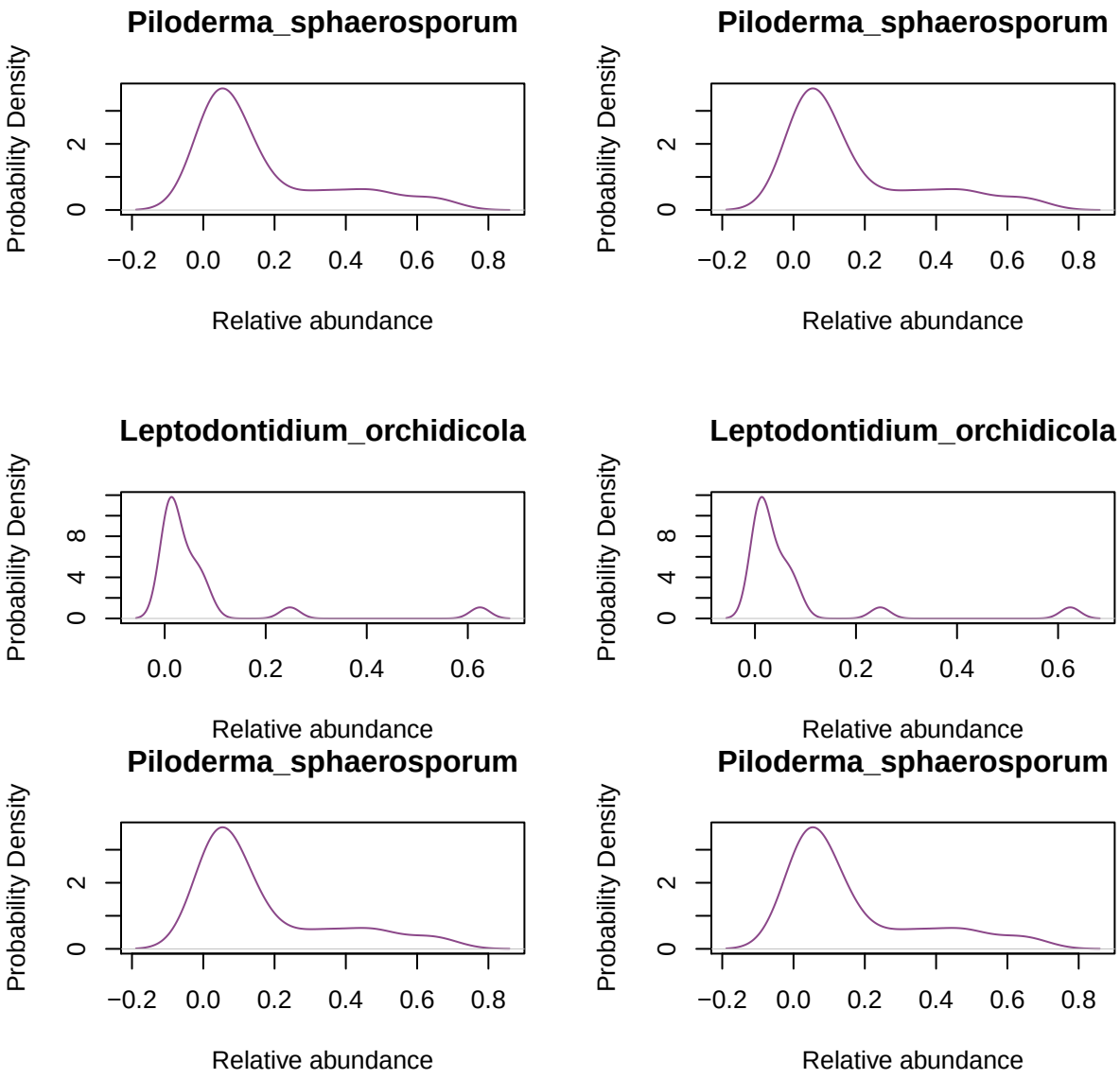
Load the dataset you are using for your project. Plot and discuss either the geogrpahic Distance-Decay relationship, the SSADs for at least four species, or any variant of the SAR (e.g., random accumulation of plots or areas, accumulation of contiguous plots or areas, nested design).

```
fungi <- read.csv("/media/removable/USB Drive/Z620.jay.non.GitHub/Endophyte communities and trees for as
fungi <- as.data.frame(fungi[-1])
fungi.names <- names(fungi)



ssad.fungi <- function(x){
  ad <- c(2,2)
  ad <- fungi[, otu]
  ad = as.vector(t(x = ad))
  ad = ad[ad > 0]
}

par(mfrow = c(2, 2))

ct <- 0

while (ct < 6){
  otu <- sample(1:length(fungi), 1)
  ad <- ssad.fungi(otu)
  if (length(ad) > 10 & sum(ad > 0.1)){
    ct <- ct+1
    plot(density(ad), col = 'orchid4', xlab = 'Relative abundance',
         ylab = 'Probability Density', main = fungi.names[otu])
  }
}
```



Below to identify the really dominant, over-represented fungal species

```
par(mfrow = c(2, 2))

ct <- 0

while (ct < 6){
  otu <- sample(1:length(fungi), 1)
  ad <- ssad.fungi(otu)
  if (length(ad) > 10 & sum(ad > 0.6)){
    ct <- ct+1
    plot(density(ad), col = 'orchid4', xlab = 'Relative abundance',
         ylab = 'Probability Density', main = fungi.names[otu])
  }
}
```



Unfortunately we currently do not have any geographic information for our data set, so we can only look at the species spatial abundance distributions for right now. But! I recently reached out to the author of the paper we are working with, and he promised to send us lat/long coordinates for every site he collected at, in the next week or so. I am quite excited to obtain this information as last week, in the beta diversity unit, we saw that there were some pretty big differences between a few specific sites - grouped by geographic region (USA P. contorta vs European P. contorta). I think that having the actual coordinates of all of these sites would allow us to ask and answer some intersting questions about how fungal communities are changing across geographic space (are there certain species that persist no matter the distance? Are microbial communities most similar between shared host species in different regions, or between different host species in the same region?) With what information we currently do have though, we were able to plot the SSADs for a number of fungal species in our dataset, using a cutoff of presence in 10 sites at a relative abundance > 0.1 to eliminate really rare species. From this we learned that the majority of our fungal species show up at relatively low relative abundances within sites, probably meaning that most of our sites show a good amount of evenness - there isn't any one species that completely dominates a site repeatedly. In fact, to get a really dominant SSAD to show up, I needed to change the cutoff for relative abundance to > 0.6 for 10 sites, anything higher than that and the while loop would not run as there weren't any species which fit the criteria. Again, I think this

just means that our sites in general are pretty even with no single species dominating.