Diversity analysis of simplex output

Ken Locey

October, 2015

OVERVEW

This R Markdown document is designed to be opened and ran in the RStudio. The chunks of code below allow for diversity analysis of **simplex** output. This includes analyzing trends in univariate diversity metrics and analyzing distributions of abundance among species. Note that **simplex** tracks and/or calculates the values of several diversity related metrics. These include:

1.) Total abundance (N): Total number of individuals
2.) Absolute dominance ($Nmax$): Number of individuals belonging to the most abundant species
3.) Species richness (S): Total number of species
4.) Simpson's evenness: Simpson's measure of species evenness
5.) Smith and Wilson's evenness: a normalized log-transform of the sample variance
6.) Whittakers species turnover: Unweighted proportional turnover among species between two samples.
7.) Jaccard's pair-wise dissimilarity: Weighted proportional turnover among species between two samples.
See: http://cc.oulu.fi/~jarioksa/softhelp/vegan/html/vegdist.html

SETUP

A. Clear and set the working directory

```
rm(list=ls())
getwd()
setwd("~/GitHub/simplex")
```

B. Import packages; install if needed

```
#install.packages("vegan")
require("vegan")
require("car")
```

C. Import custom modules of R functions

```
source("~/GitHub/simplex/tools/Rbin/metrics.R")
```

D. Import simulated data

```
# A table where each columns corresponds to a state variable or model output.
simplex.dat <- read.csv("~/GitHub/simplex/results/simulated_data/examples/SimData.csv")
simplex.dat$h.tau <- log((simplex.dat$width * simplex.dat$height)/simplex.dat$flow.rate)
```

```
# Replacing 0's with 1's to allow log-transforms below
simplex.dat$total.abundance[simplex.dat$total.abundance<=0] <- 1
simplex.dat$N.max[simplex.dat$N.max<=0] <- 1
simplex.dat$species.richness[simplex.dat$species.richness<=0] <- 1

simplex.dat$total.abundance <- log(simplex.dat$total.abundance)
simplex.dat$N.max <- log(simplex.dat$N.max)
simplex.dat$species.richness <- log(simplex.dat$species.richness)
simplex.dat[is.na(simplex.dat)] <- 0

# Two files, where the ith element of the jth row correspond between files
# and where jth row correspond to the jth row of the above three files
species.list <- get.vectors("~/GitHub/simplex/results/simulated_data/examples/Species.csv", type = 'cha
rads <- get.vectors("~/GitHub/simplex/results/simulated_data/examples/RADs.csv", type = 'number')
```

DIVERSITY ANALYSIS

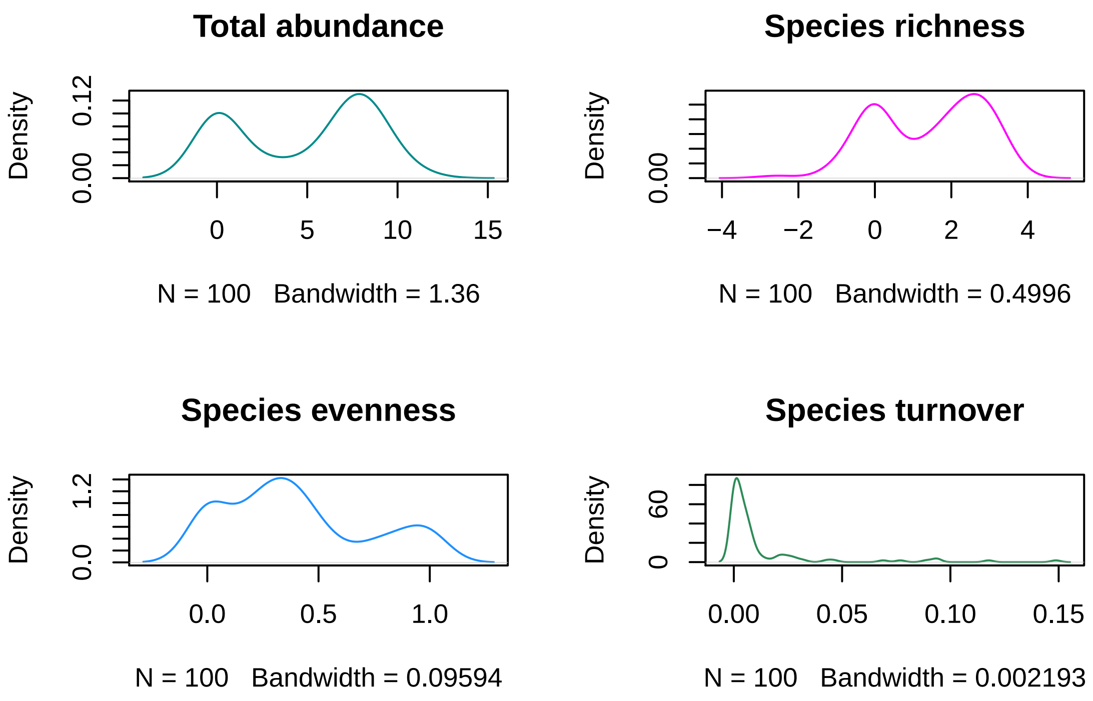
A. Graphical exploration via kernel density estimation

Each row in our **simplex** output files corresponds to a model with an average total abundance, species richness, species turnover, etc. One way of developing a feel for the values of these variables simulated by **simplex** is through the use of kernel density estimation (kde). In short, kde produces a smoothed histogram where integrating between two points on the x-axis gives the probability that a randomly chosen value will occur within that range. Let's plot kernel density curves for log10-transformed total abundance, log10-transformed species richness, Simpson's evenness metric, and Whittaker's measure of species turnover.

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

Sx <- density(simplex.dat$species.richness)
Nx <- density(simplex.dat$total.abundance)
Ex <- density(simplex.dat$simpson.e)
Tx <- density(simplex.dat$Whittakers.turnover)

plot(Nx, main = "Total abundance", col = 'DarkCyan')
plot(Sx, main = "Species richness", col = 'magenta')
plot(Ex, main = "Species evenness", col = 'dodgerblue')
plot(Tx, main = "Species turnover", col = 'seagreen')
```



As we can see, there are some interesting aspects of modality, variance, and skewness. For instance, both total abundance and species richness are characterized by two modes, one at very low values (near log10 of 0) and one at higher values (near a few hundred or several thousand). Hint: graphing both N and S against residence time, (height*width)/flow rate, would reveal why.

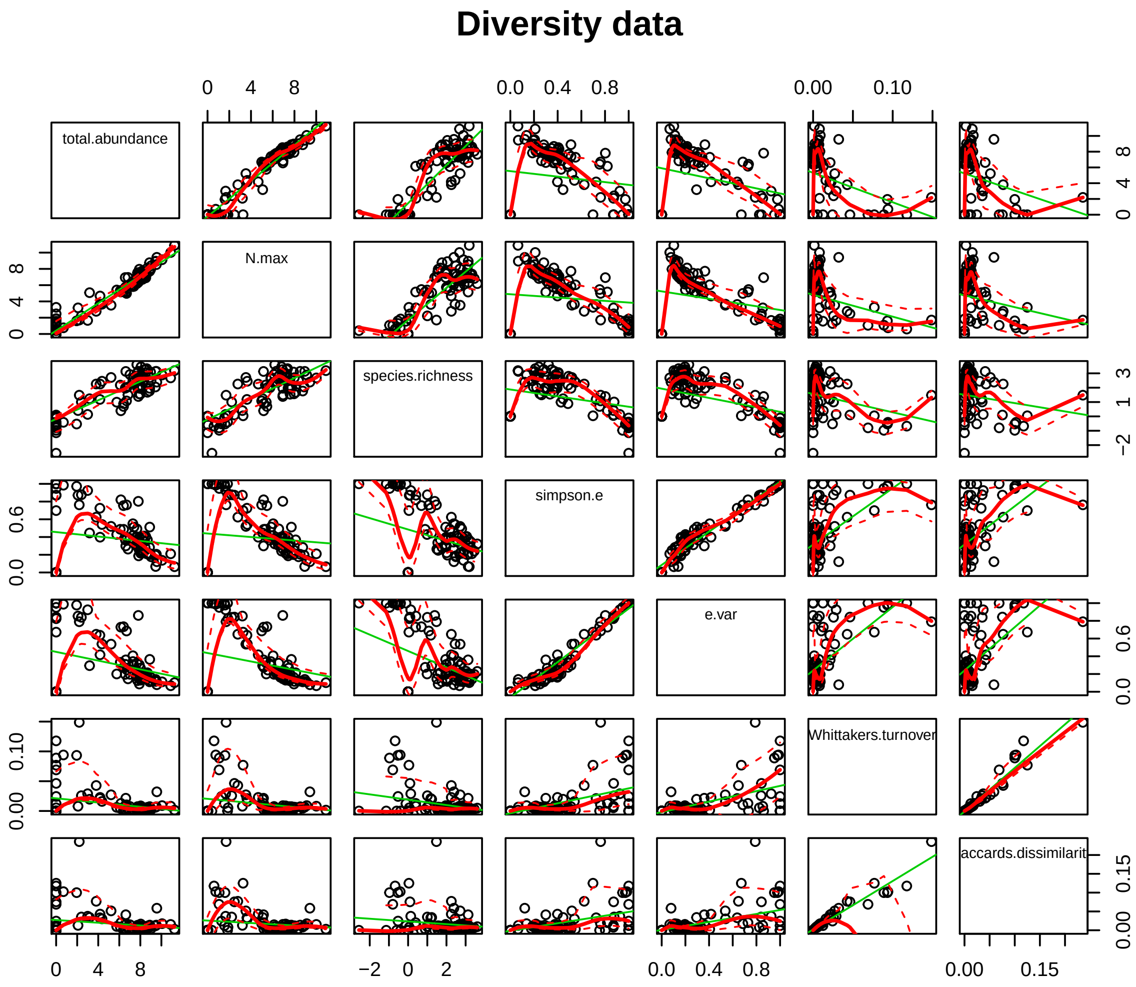
Likewise, we can see that species turnover is largely concentrated at low values but has a long tail extending to higher values (right-skewed). Here, species turnover is the portion of species differing between two samples. We can also see that evenness ranges across its full breadth, from 0.0 (no evenness) to 1.0 (perfect evenness). Notice that because kernel density curves are smoothed and based on probabilities, the tails will extend beyond the range of the data, i.e., unless parameters like the kernel width are adjusted. type 'help(density)' for greater explanation.

Univariate relationships

Let's explore relationships between diversity related variables. As we can see different evenness measure basically reflect each other, as do different measures of species turnover. Likewise, we see a strong postivie relationships between total abundance and species richness and between total abundance and the abundance of the most abundant species; which we should expect.

```
diversity.dat <- as.matrix(subset(simplex.dat,
                      select = c(total.abundance,
                      N.max,
                      species.richness,
                      simpson.e,
                      e.var,
                      Whittakers.turnover,
                      Jaccards.dissimilarity)))

scatterplotMatrix(diversity.dat, main="Diversity data", diagonal = "none")
```



Rank-abundance curves (RACs)

This section pull heavily from the Quantative Biodiversity course designed and taught by Jay Lennon, myself (Ken Locey), and Mario Muscarella. https://github.com/QuantitativeBiodiversity/Assignments/tree/master/Alpha

While the diversity metrics can be useful in compressing one or more aspects of diversity into a single number, information is invariably lost in the process. More specifically, each metric quantifies an aspect of the vector of species abundances, whether that be the number of species (richness) or similarity in abundance among species (evenness).

Consequently, we can also analyze the ranked vector of species abundances resulting from **simplex** models. This vector is simply known as the rank-abundance curve (RAC) and its form is predicted by no less than 20 theories or models of ecology/biodiversity! In fact, the uneven shape of the RAC is one of the most intensively studied patterns in ecology, and underpins all or most ecological theories of biodiversity. Predicting the form of the RAC is the first test that any biodiversity theory must pass and there are there are dozens of models that have attempted to explain the uneven form of the RAC across ecological systems. These models attempt to predict the form of the RAC according to mechanisms and processes that are believed to be important to the assembly and structure of ecological systems.

Let's use the `radfit()` function in `vegan` to fit the predictions of various species abundance models to the RAC of a chosen **simplex** model output (i.e. a row in the RADs.csv file). Descriptions of several species abundance models (as used in vegan) are provided below.

First, choose a simplex model with a relatively large number of species. You can use the `sapply` function to find the length of each rad.

```
sapply(rads, length)
```

Now, lets use the `radfit` function to fit several species abundance models.

```
RACresults <- radfit(rads[[42]])
RACresults
```

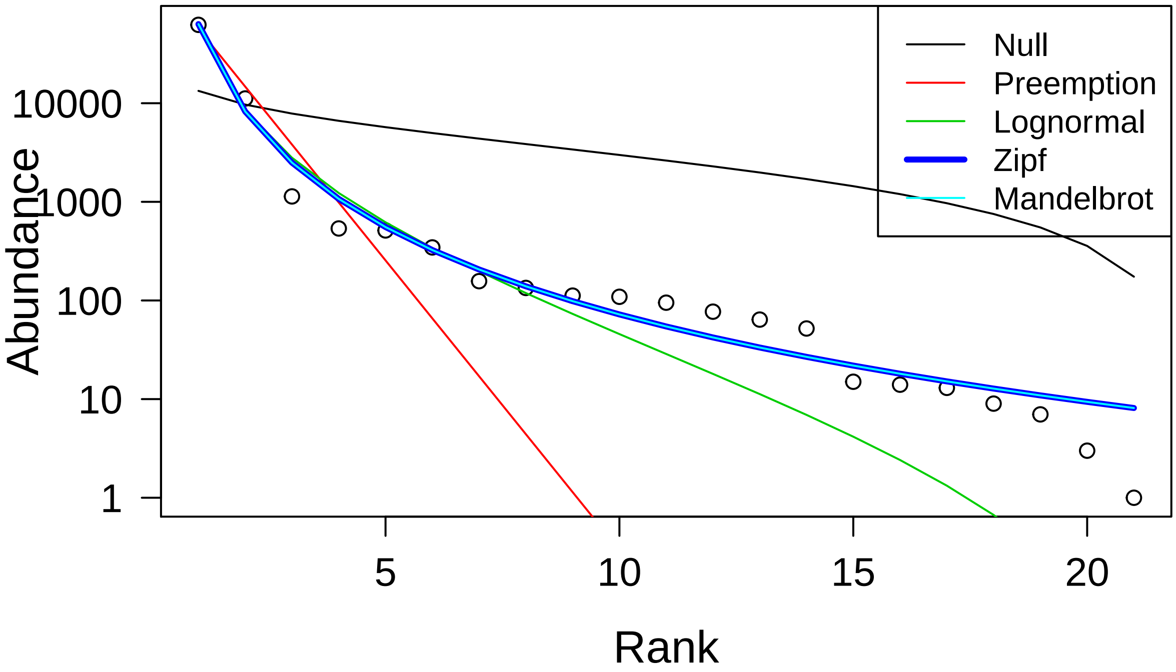
We see that `vegan` fits five models to our RAC: *Null*, *Preemption*, *Lognormal*, *Zipf*, and *Mandelbrot*. Before explaining what these models represent, let's run through the `vegan` output:

1. Next to "RAD models", we see "family poisson", which tells us that by default, `vegan` assumes Poisson distributed error terms.
2. Below this, we see that `vegan` returns the number of species (S) and the number of individuals (N) for the empirical RAC.
3. Next, we see a table of information, the first columns of which are par1, par2, and par3. These columns pertain to model parameters and reveal that the different models use different numbers of parameters; the null model uses none.
4. Next, we see a column for Deviance, which is a quality of fit statistic based on the idea of residual sums of squares.
5. After Deviance, we see columns for AIC and BIC, which are the estimated **Akaike Information Criterion** and the **Bayesian Information Criterion**, respectively.

Notes on AIC and BIC AIC and BIC are commonly used for model selection. In other words, they help us identify a model that is best supported by our data. Obviously, the more parameters a model has, the better it will fit a data set. However, it's not necessarily desirable to have an over-parameterized model. So, AIC and BIC asssign penalties that correspond with the number of parameters that a model uses. In the end, the "best" model has the lowest AIC or BIC value.

Now, let's visualize our results by plotting the empirical RAC and the predicted RAC for each model:

```
plot(RACresults, las=1, cex.lab = 1.4, cex.axis = 1.25)
```



Interpreting the RAC models in vegan:

Null: A **broken stick model** (MacArthur 1960, Pielou 1975) where the expected abundance of a species at rank r is $a_r = \frac{N}{S} \cdot \sum_{x=r}^{S} \frac{1}{x}$. N is the total number of individuals and S is the total number of species. This gives a constraint-based null model where the N individuals are randomly distributed among S species, and there are no fitted parameters. Null models often reveal that realistic patterns can be expected from random sampling, and have been extremely useful in ecology and evolution.

Preemption: The **niche preemption model** (a.k.a., geometric series or Motomura model): Envision an environment occupied by a single species. Now, envision that a second species colonizes the environment and takes some portion of resources equal to α. Then, envision that a third species colonizes the environment and takes a portion of resources equal to α away from the second species. Imagine this process continuing until N is zero. The only estimated parameter is the preemption coefficient α, which gives the decay rate of abundance per rank. The expected abundance (a) of species at rank r is $a_r = N \cdot \alpha \cdot (1 - \alpha)^{(r-1)}$.

Lognormal: Many statistical models assume that the distribution of values are normally distributed. When applied to species abundances, this means that they conform to the shape of a symmetrical bell curve, more precisely known as the Gaussian distribution. In contrast, the log-Normal model assumes that the logarithmic abundances are normally distributed. The expected abundance of a species at rank r is then: $a_r = e^{log(\mu)+log(\sigma) \cdot Z}$, where Z is the Normal deviate. A Normal deviate is simply the number of standard deviations a score is from the mean of its population. The log-normal model was introduced into ecology by Frank Preston in 1948 and is one of the most widely successful species abundance models.

Zipf: The Zipf model is based on Zipf's Law, an well-known observation that many types of ranked data are fit by a simple scaling law (a.k.a., power law). In short, the abundance of a species in the RAC is inversely proportional to its rank. The expected abundance (a) of species at rank r is: $a_r = N \cdot p_1 \cdot r^{\gamma}$, where p_1 is the fitted proportion of the most abundant species, and γ is a decay coefficient.

Mandelbrot: Shortened name for the Zipf–Mandelbrot model, a generalization of the Zipf model made by the mathematician and father of fractal geometry, Benoit Mandelbrot. This model adds one parameter (β) to the Zipf model. The expected abundance of a species (a) at rank r is $a_r = N \cdot c \cdot (r + \beta)^{\gamma}$. Here, the p_1 parameter of the Zipf model changes into a meaningless scaling constant c.

Let's go even further and ask which model, on average, best fits the RACs of **simplex** models. Let's begin by making a data frame of the results.

```r
RAC.dat.frame <- data.frame(BS=double(),GS=double(),LN=double(),
                            Zp=double(),ZM=double(),stringsAsFactors=F)

i = 1
for(rad in rads){
  # Not analyzing any model resulting in less than 6 species
  if(length(rad) > 5){

    RACresults <- radfit(rad)
    # get the BIC scores for each model

    # Broken-stick (vegan's null)
    BS <- as.numeric(RACresults[[3]][[1]][6])
    # Geometric series
    GS <- as.numeric(RACresults[[3]][[2]][6])
    # Preston's lognormal
    LN <- as.numeric(RACresults[[3]][[3]][6])
    # Zipf
    Zp <- as.numeric(RACresults[[3]][[4]][6])
    # Zipf-Mandelbrot
    ZM <- as.numeric(RACresults[[3]][[5]][6])

    irow <- c(i, BS, GS, LN, Zp, ZM)
    RAC.dat.frame <- rbind(RAC.dat.frame, irow)
    i = i+1
  }
}

RAC.dat.frame <- setNames(RAC.dat.frame,
          c("row", "BS","GS","LN","Zp","ZM"))
```

Remember, the "best" fit model has the lowest AIC or BIC value.

```r
par(mfrow=c(1, 1))

print(c("Broken-stick", mean(RAC.dat.frame$BS), median(RAC.dat.frame$BS)))
```

```
## [1] "Broken-stick"     "9064.8676419776"  "806.252684802876"
```

```
print(c("Geometric-series", mean(RAC.dat.frame$GS), median(RAC.dat.frame$GS)))
```

```
## [1] "Geometric-series" "888.78247873453"  "225.581429674949"
```

```
print(c("Lognormal",mean(RAC.dat.frame$LN),median(RAC.dat.frame$LN)))
```

```
## [1] "Lognormal"        "700.997653109312" "253.979698379513"
```

```
print(c("Zipf",mean(RAC.dat.frame$Zp),median(RAC.dat.frame$Zp)))
```

```
## [1] "Zipf"             "901.205383876966" "405.963159537436"
```

```
# Appears to be an "NA" in the ZM data
ZM <- RAC.dat.frame$ZM[!is.na(RAC.dat.frame$ZM)]
print(c("Zipf-Mandelbrot", mean(ZM), median(ZM)))
```

```
## [1] "Zipf-Mandelbrot"  "292.953555243951" "205.249675108776"
```

It appears that the Zipf-Mandelbrot distribution is the winner with respect to the mean and median BIC values.