Latest version: Sept. 26, 2016

A Brief Introduction to SpadeR (R package): Species-Richness Prediction and Diversity Estimation

Anne Chao, K. H. Ma, T. C. Hsieh and C. H. Chiu

Institute of Statistics, National Tsing Hua University, Hsin-Chu, Taiwan 30043

Overview

`SpadeR` (Species-Richness Prediction and Diversity Estimation with R) is an updated R package from the original version of SPADE. `SpadeR` provides simple R functions to compute various biodiversity indices and related (dis)similarity measures based on individual-based (abundance) data or sampling-unit-based (incidence) data taken from one or multiple communities/assemblages. The `SpadeR` package is available in CRAN. We have been updating `SpadeR` and you can download the latest version from Github (see below) or from Anne Chao's website.

- You need to acquire basic knowledge about R to use functions supplied by `SpadeR`.
- For readers without R background, please try our `SpadeR Online`, an R-based online version via the link Anne Chao's website or https://chao.shinyapps.io/SpadeR/. Users do not need to learn/understand R to run `SpadeR Online`.

Both `SpadeR` (R package) and `SpadeR Online` include nearly all of the important features from the original program SPADE while also having the advantages of expanded output displays and simplified data input formats. See SpadeR Manual for all details of the functions supplied in the package. For numerical examples with proper interpretations, see the detailed Online SpadeR User's Guide.

This package contains six main functions:

1. ChaoSpecies (estimating species richness for one community).
2. Diversity (estimating a continuous diversity profile and various diversity indices in one community including species richness, Shannon diversity and Simpson diversity). This function also features plots of empirical and estimated continuous diversity profiles.
3. ChaoShared (estimating the number of shared species between two communities).
4. SimilartyPair (estimating various similarity indices between two assemblages). Both richness and abundance-based two-community similarity indices are included.
5. SimilarityMult (estimating various similarity indices among *N* communities). Both richness and abundance-based *N*-community similarity indices are included.

6. Genetics (estimating allelic dissimilarity/differentiation among sub-populations based on multiple subpopulation genetics data).

Except for the Genetics function, there are at least three types of data are supported for each function.

Data Types

It is very important to prepare your data in correct format. Data are generally classified as abundance data and incidence data and there are five types of data input formats options (datatype="abundance", "abundance_freq_count", "incidence_freq", "incidence_freq_count", "incidence_raw").

- Individual-based abundance data when a sample of individuals is taken from each community.

  Type (1) abundance data (datatype = "abundance"): Input data consist of species (in rows) by community (in columns) matrix. The entries of each row are the observed abundances of a species in N communities.

  Type (1A) abundance-frequency counts data only for a single community (datatype = "abundance_freq_count"): input data are arranged as (1 f_1 2 f_2 ... r f_r)(each number needs to be separated by at least one blank space or separated by rows), where r denotes the maximum frequency and f_k denotes the number of species represented by exactly k individuals/times in the sample. Here the data (f_1, f_2, ..., f_r) are referred to as "abundance-frequency counts".

- Sampling-unit-based incidence data when a number of sampling units are randomly taken from each community. Only the incidence (detection/non-detection) of species is recorded in each sampling unit. There are three data formats options.

  Type (2) incidence-frequency data (datatype="incidence_freq"): The first row of the input data must be the number of sampling units in each community. Beginning with the second row, input data consist of species (in rows) by community (in columns) matrix. The entries of each row are the observed incidence frequencies (the number of detections or the number of sampling units in which a species are detected) of a species in N communities.

  Type (2A) incidence-frequency counts data only for a single community (datatype="incidence_ freq_count"): input data are arranged as (T 1 Q_1 2 Q_2 ... r Q_r) (each number needs to be separated by at least one blank space or separated by rows), where Q_k denotes the number of species that were detected in exactly k sampling units, while r denotes the number of sampling units in which the most frequent species were found. The first entry must be the total number of sampling units, T. The data (Q_1, Q_2, ..., Q_r) are referred to as "incidence frequency counts".

Type (2B) incidence-raw data (datatype="incidence_raw"): Data consist of a species-by-sampling-unit incidence (detection/non-detection) matrix; typically "1" means a detection and "0" means a non-detection. Each row refers to the detection/non-detection record of a species in T sampling units. Users must specify the number of sampling units in the function argument "units". The first T_1 columns of the input matrix denote species detection/non-detection data based on the T_1 sampling units from Community 1, and the next T_2 columns denote the detection/non-detection data based on the T_2 sampling units from Community 2, and so on, and the last T_N columns denote the detection/non-detection data based on T_N sampling units from Community N, $T_1 + T_2 + \cdots + T_N = T$.

Software needed

- Required: R
- Suggested: RStudio IDE

How to install

The `SpadeR` package is available on CRAN and can be downloaded with a standard installation procedure by using the commands in the R Console:

```
## install SpadeR package from CRAN
install.packages("SpadeR")
library(SpadeR)
```

The latest version can be downloaded from the github.

```
## install the latest SpadeR from github
install.packages('devtools')
library(devtools)
install_github('AnneChao/SpadeR')
library(SpadeR)
```

Remark: In order to install devtools package, you should update R to the last version. Also, to get install_github to work, the httr package should be installed.

Run SpadeR by examples

In the package, we have included many demo datasets for illustration. To gain familiarity with the program, we suggest that users first run the demo data sets included in SpadeR package and check the output with that given in the SpadeR User's Guide. Part of the output for each example is also interpreted in the guide to help users understand the statistical results. The formulas for estimators featured in SpadeR with relevant references are also provided in the SpadeR User's Guide.

1. ChaoSpecies (estimating species richness for one community)

```
# Data for Function ChaoSpecies(data, datatype, k = 10, conf = 0.95)
data(ChaoSpeciesData)

# Type (1) abundance data
ChaoSpecies(ChaoSpeciesData$Abu,"abundance", k=10, conf = 0.95)

# Type (1A) abundance-frequency counts data
ChaoSpecies(ChaoSpeciesData$Abu_count, "abundance_freq_count", k=10, conf = 0.95)

# Type (2) incidence-frequency data
ChaoSpecies(ChaoSpeciesData$Inci, "incidence_freq", k=10, conf = 0.95)

# Type (2A) incidence-frequency counts data
ChaoSpecies(ChaoSpeciesData$Inci_count, "incidence_freq_count", k=10, conf = 0.95)

# Type (2B) incidence-raw data
ChaoSpecies(ChaoSpeciesData$Inci_raw, "incidence_raw", k=10, conf = 0.95)
```

2. Diversity (estimating a continuous diversity profile and various diversity indices in one community including species richness, Shannon diversity and Simpson diversity). This function also features plots of empirical and estimated continuous diversity profiles.

```
# Data for Function Diversity(data, datatype, q = NULL)
data(DiversityData)

# Type (1) abundance data
Diversity(DiversityData$Abu,"abundance", q=c(0,0.5,1,1.5,2))

# Type (1A) abundance-frequency counts data
Diversity(DiversityData$Abu_count,"abundance_freq_count", q = seq(0, 3, by = 0.5))

# Type (2) incidence-frequency data
Diversity(DiversityData$Inci,"incidence_freq", q = NULL)

# Type (2A) incidence-frequency counts data
Diversity(DiversityData$Inci_freq_count, "incidence_freq_count" , q = NULL)

# Type (2B) incidence-raw data
Diversity(DiversityData$Inci_raw, "incidence_raw", q = NULL)
```

3. ChaoShared (estimating the number of shared species between two communities).

```
# Data for Function ChaoShared(data, datatype, units, se = TRUE, nboot = 200, conf = 0.95)
data(ChaoSharedData)

# Type (1) abundance data
ChaoShared(ChaoSharedData$Abu,"abundance", se = TRUE, nboot = 200, conf = 0.95)

# Type (2) incidence-frequency data
ChaoShared(ChaoSharedData$Inci,"incidence_freq", se = TRUE, nboot = 200, conf = 0.95)

# Type (2B) incidence-raw data
ChaoShared(ChaoSharedData$Inci_raw,"incidence_raw", units = c(16,17), se = TRUE, nboot = 200, conf = 0.95)
```

4. SimilartyPair (estimating various similarity indices between two assemblages). Both richness and abundance-based two-community similarity indices are included.

```
# Data for Function SimilarityPair(data, datatype, units, nboot = 200)
data(SimilarityPairData)

# Type (1) abundance data
SimilarityPair(SimilarityPairData$Abu,"abundance", nboot = 200)

# Type (2) incidence-frequency data
SimilarityPair(SimilarityPairData$Inci,"incidence_freq", nboot = 200)

# Type (2B) incidence-raw data
SimilarityPair(SimilarityPairData$Inci_raw,"incidence_raw", units = c(19,17), nboot = 200)
```

5. SimilarityMult (estimating various similarity indices among *N* communities). Both richness and abundance-based *N*-community similarity indices are included.

```
# Data for Function SimilarityMult(data, datatype, units, q, nboot = 200, goal)
data(SimilarityMultData)

# Type (1) abundance data
SimilarityMult(SimilarityMultData$Abu,"abundance", q = 2, nboot = 200, "relative")

# Type (2) incidence-frequency data
SimilarityMult(SimilarityMultData$Inci,"incidence_freq", q = 2, nboot = 200, "relative")

# Type (2B) incidence-raw data
SimilarityMult(SimilarityMultData$Inci_raw,"incidence_raw", units = c(19,17,15), q = 2, nboot = 200, "relative")
```

6. Genetics (estimating allelic dissimilarity/differentiation among sub-populations based on multiple subpopulation genetics data).

```
# Data for Function Genetics(X, q, nboot = 200)
data(GeneticsDataAbu)

# Type (1) abundance data
Genetics(GeneticsDataAbu, q=2, nboot = 200)
```

How to cite

If you publish your work based on results from SpadeR, please make references to our relevant methodology papers mentioned below and also use the following reference for citing SpadeR:

Chao, A., Ma, K. H., Hsieh, T. C. and Chiu, C. H. (2016) SpadeR (Species-richness Prediction And Diversity Estimation in R): an R package in CRAN. Program and User's Guide also published at http://chao.stat.nthu.edu.tw/wordpress/software_download/.

We recommend the following recent papers for pertinent background on biodiversity measures and statistical analyses. These papers can be directly downloaded from Anne Chao's website.

Chao, A., and Chiu, C. H. (2012). Estimation of species richness and shared species richness. In N. Balakrishnan (ed). *Methods and Applications of Statistics in the Atmospheric and Earth Sciences*. p.76–111, Wiley, New York. (Background on species richness and shared species richness estimation)

Chao, A., and Chiu, C. H. (2016). Nonparametric estimation and comparison of species richness. *Wiley Online Reference in the Life Science*. In: eLS. John Wiley & Sons, Ltd: Chichester. (Background on comparing species richness across communities)

Chao, A., and Chiu, C. H. (2016). Bridging the variance and diversity decomposition approaches to beta diversity via similarity and differentiation measures. *Methods in Ecology and Evolution*, 7, 919–928. (A unified theoretical framework on similarity/differentiation measures)

Chao, A., Chiu, C. H. and Jost, L. (2014). Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity/differentiation measures through Hill numbers. *Annual Reviews of Ecology, Evolution, and Systematics*, 45, 297–324. (A unified theoretical framework on diversity measures)

Chao, A., Gotelli, N. J., Hsieh, T. C., Sander, E. L., Ma, K. H., Colwell, R. K. and Ellison, A. M. (2014). Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs*, 84, 45–67. (Background on comparing diversity measures across communities)

Chao, A. and Jost, L. (2015). Estimating diversity and entropy profiles via discovery rates of new species. *Methods in Ecology and Evolution*, 6, 873–882. (A unified approach to estimating diversity in a community based on incomplete samples)

Chao, A., Wang, Y. T. and Jost, L. (2013). Entropy and the species accumulation curve: a novel entropy estimator via discovery rates of new species. *Methods in Ecology and Evolution*, 4, 1091–1100. (A nearly optimal estimator of Shannon entropy/diversity based on incomplete samples)