Original version (September 2016)

User's Guide for iNEXT Online: Software for Interpolation and Extrapolation of Species Diversity

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

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

Overview

iNEXT (iNterpolation and EXTrapolation) Online is the R-based interactive online version of iNEXT available via the link https://chao.shinyapps.io/iNEXTOnline/ or http://chao.stat.nthu.edu.tw/wordpress/software_download/. Clicking these links, you will be directed to the online interface window. Users do not need to learn/understand R to run iNEXT Online. The interactive web application was built using the Shiny (a web application framework).

iNEXT features two statistical analyses (non-asymptotic and asymptotic) for species diversity based on Hill numbers:

(1) A non-asymptotic approach based on interpolation and extrapolation

iNEXT computes the estimated diversities for standardized samples with a common sample size or sample completeness. This approach aims to compare diversity estimates for equally-large (with a common sample size) or equally-complete (with a common sample coverage) samples; it is based on the seamless rarefaction and extrapolation sampling curves of Hill numbers for $q = 0$, 1 and 2. See Colwell et al. (2012), Chao and Jost (2012) and Chao et al. (2014) for pertinent background and methods.

(2) An asymptotic approach to infer asymptotic diversity

iNEXT computes the estimated asymptotic diversity profiles. It is based on statistical estimation of the true Hill number of any order $q \geq 0$; see Chao and Jost (2015) for the statistical estimation detail.

How to cite

If you publish your work based on results from iNEXT Online, you should make references to the relevant methodology papers mentioned below in the reference list and also use the following reference to cite iNEXT Online:

> Chao, A., Ma, K. H., and Hsieh, T. C. (2016) iNEXT (iNterpolation and EXTrapolation) Online: Software for Interpolation and Extrapolation of Species Diversity. Program and User's Guide published at http://chao.stat.nthu.edu.tw/wordpress/software_download/.

Data

Two types of data

iNEXT Online supports two types of data:

- Individual-based abundance data: data for each assemblage/site include sample species abundances in an empirical sample of individuals (called a reference sample).

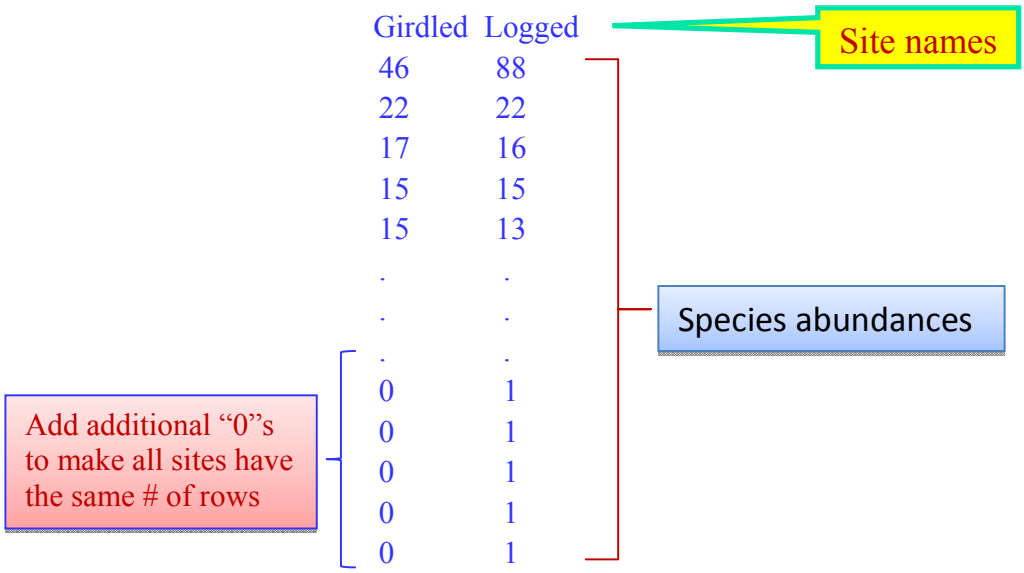
- Sampling-unit-based incidence data: for each assemblage/site, data for a reference sample consist of species presence/absence (or detection/non-detection) in each of multiple sampling units.

How to prepare your data input files

- Abundance Data

  When there are *N* assemblages**,** the observed species abundances should be arranged as a species (in rows) by assemblage (in columns) matrix. The first row (including *N* entries) lists the assemblage labels or site names for the *N* assemblages; see an example below. Beginning with the second row, the entry for the first column in each row denotes a species abundance from Assemblage 1, and the entry for the second column denotes a species abundance from Assemblage 2, and so on. If species names or any identification codes are recorded in your original data, then they must be removed to conform to the iNEXT Online data format.

  For example, the spider data (one of the two sets for demo abundance data) consist of species sample abundances from two canopy manipulation treatments ("Girdled" and "Logged") of hemlock trees; see Chao et al. (2014) for analysis details and data interpretations. The observed species abundances should be arranged in a text data file as follows for two sites: Girdled and Logged.



Missing data, which are empty cells without numbers, cannot be read by iNEXT Online. The number of rows must be the same for the *N* assemblages. For example, there are 26 observed species in the Girdled site and 37 species in the Logged sites, the input data must consist of 37 rows (in addition to the site labels). Therefore, eleven "0" abundances should be added to the Girdled column. Adding "0" abundances will not have any effect on our analysis. Additional rows with all 0 frequencies 0 for unobserved species may be included, but that species will not have any effect on the analysis.

A typical species-check-list abundance data can be used for input data, but species identification names must be removed from your original data file. When a species was not observed or detected in a community, you must enter the frequency 0 for a "non-detection". In this kind of species-check-list, numbers in the same row refer to the same species, but this is not required for iNEXT Online input. For example, in the above spider data, each row can be allowed to correspond to different species because the between-assemblage information is not needed in our comparison of within-assemblage species diversity.

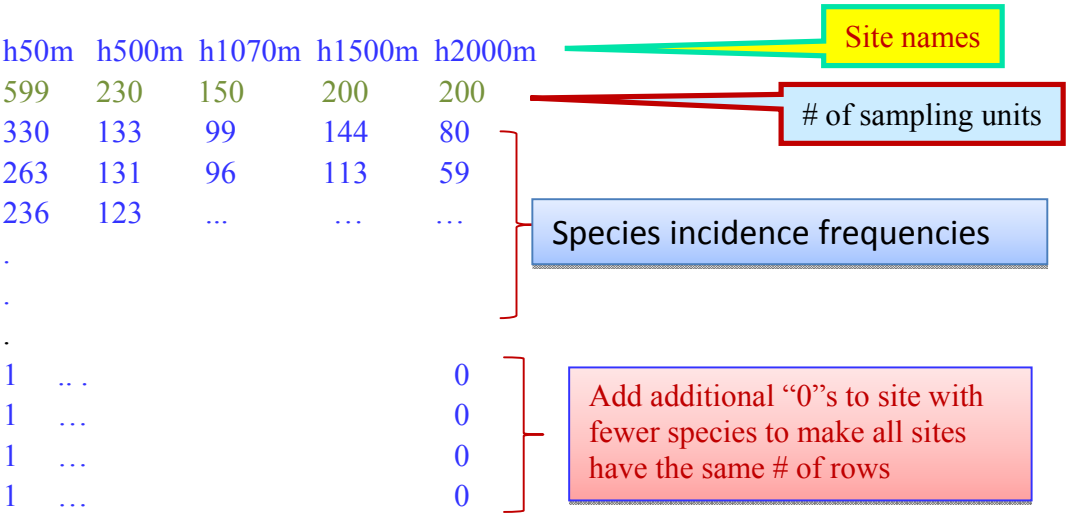
In the special case of $N = 1$ (i.e., there is only one assemblage), all data should be read in one column. That is, the first entry denotes the assemblage label or site name, and the observed species abundances are entered in different rows.

- Incidence Data

The data input format is generally similar to that for abundance data, except that the total number of sampling unit in each assemblage must be supplied. That is, the first row lists the assemblage labels or site names, and the second row lists the number of

sampling units in the *N* assemblages, as shown below. Beginning with the third row, the entry in the first column in each row denotes an incidence frequency (total number of detections of a species in all sampling units) from Assemblage 1, and the entry for the second column denotes an incidence frequency from Assemblage 2, and so on.

For example, in the demo ant data, species incidence frequencies were collected at five elevations (50m, 500m, 1070m, 1500m, and 2000m). See Chao et al. (2014) for analysis details and data interpretations. The numbers of trapping sampling units in the five elevations are respectively 599, 230, 150, 200 and 200. The data should be read in a text file as follows.



As with the abundance data, the number of rows for each assemblage must be the same, so additional "0" abundances should be added to those assemblages with fewer species.

How to key in data in the screen window

You can also directly key in site names and species abundances on the screen panel. For example, the key-in data for the spiders from the two sites Girdled and Logged are shown as follows.

Girdled 46 22 17 15 15 9 8 6 6 4 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1
Logged 88 22 16 15 13 10 8 8 7 7 7 5 4 4 4 3 3 3 3 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

The data for each assemblage starts with its assemblage label/name, followed by species abundances. Each number needs to be separated by at least one blank space, and with a "return" at the end of data entry for each assemblage. Here the numbers of entry for multiple assemblages are allowed to differ; no need to add "0" abundances. Note that adding additional "0" abundances will not affect all analyses.

For incidence data, the additional information about the number of sampling units must be provided in the second entry. For example, the key-in data for the ants from h1500m (with 200 sampling units) and h2000m (with also 200 sampling units) are shown as follows.

h1500m 200 144 113 79 76 74 73 53 50 43 33 32 30 29 25 25 25 24 23 23 19 18 17 17 11 11 9 9 9 9 6 6 5 5 5 5 4 4 3 3 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1
h2000m 200 80 59 34 23 19 15 13 8 8 4 3 2 2 1

Running procedures for "Interpolation/Extrapolation" analysis

Species identification names are irrelevant in species diversity assessment so they must be removed from your original data to conform to iNEXT Online format. The following Steps 1, 2 and 4 are required procedures (selections are placed in blocks); Step 3 is optional.

Step 1. Select an analysis part (Interpolation/Extrapolation or Asymptotic Analysis) from the top menu of iNEXT Online window.

Step 2. "Data Setting" (on the left hand side of the window screen)

(2a) Select Demo data to load demo data, Upload data to load your own data, or Key in data) by typing in your data in the input window.

(2b) Select data type: Abundance data or Incidence data; see above for data input formats.

(2c) For abundance data, there are two demo data sets. Select one Spiders or Beetles.

(2d) Then the names/labels for the uploaded assemblages/sites will be automatically shown in the window below; you can select one set or multiple sets for comparison.

Step 3. "General Setting" for statistical procedures

(3a) Optional: select a diversity order of q, q = 0 for species richness; q = 1 for Shannon diversity; q = 2 for Simpson diversity (default is q = 0).

(3b) Optional: specify the number of bootstrap replications to compute s.e. and confidence intervals for each estimator. (Default bootstrap replications = 50). You can type in "0" to skip all bootstrapping to save running time.

(3c) Optional: specify the level for confidence interval (default level = 0.95).

(3d) Optional: you can either specify the endpoint of extrapolation range and/or the number of knots or specify the samples sizes for which diversity estimates will be

calculated. (Default endpoint = double of the minimum reference sample size, and the default number of knots = 40). If you choose to specify the sample sizes, then type in those sample sizes in the window space.

Step 4. Press the Run! button to get the output.

Note 1: We use a bootstrap resampling method to compute s.e. and confidence interval of any estimator involved in the analysis. The default number of bootstrap replications is 50 to save running time. You may specify a larger number (say, 100) to obtain more accurate results for publication purposes, but it will take a longer time to get the output.

Note 2: If you just want to take a glance at the pattern of rarefaction/extrapolation sampling curves and diversity profiles without requiring confidence intervals, then type in "0" for the number of bootstrap replications; in this case, the computation of s.e. and confidence intervals will be skipped so that the output can be promptly shown.

Note 3: The bootstrap resampling procedures vary with trial, meaning that two different runs for the same data may result in different s.e. estimates and different confidence intervals.

Running procedures for "Asymptotic Analysis"

When you select the "Asymptotic Analysis" from the top menu, all the procedures are the same as those for the "Interpolation/Extrapolation" part except for Step (3c):

Step (3d) Optional: you either use the default diversity orders (q = 0 to 3 in increments of 0.25) or you type in those diversity orders for which asymptotic estimates will be computed.

Output for Interpolation/Extrapolation

Along the second row (output) menu, there are five output selection tabs:

- In the "**Data Summary**" tab panel, basic data information (sample size, observed species richness and estimated sample coverage) and the first 10 abundance (or incidence) frequency counts are shown for the reference sample.

- In the "**Rarefaction and Extrapolation**" tab panel, various estimates for interpolated or extrapolated samples and their confidence intervals are supplied. These results can be downloaded by clicking "download as CSV file" at the bottom of the displayed figure.

- In the "**Figure Plots**" tab panel, three types of sampling curves are shown: sample-size-based rarefaction and extrapolation sampling curve, sample completeness curve and Coverage-based rarefaction and extrapolation sampling curve. These figures can be downloaded by clicking "download as PNG file" at the bottom of the displayed figures.

- In the "**Introduction**" panel, users can view a brief introduction to iNEXT Online and a summary of the running procedures.

- In the "**User Guide**" panel, a link will direct users to this user guide.

Example 1: run the demo abundance data (spiders) for interpolation/extrapolation analysis of species richness (see Chao et al. 2014 for interpretation)

Step 1. Select Interpolation/Extrapolation from the top menu.

Step 2. "Data Setting" (on the left hand side of the window)

(2a) Check the Demo data radio button to load demo data.

(2b) Select data type Abundance data.

(2c) There are two demo data sets for abundance data; select Spiders.

(2d) There are two sites (Girdled and Logged); you can select one set or multiple sets for comparison. Here we select both.

Step 3. "General Setting": use all default settings for species richness; no any selection is needed.

Step 4. Press the Run! button to get the output.

OUTPUT

In the "Data Summary" panel, the following statistics are supplied:

	Girdled	Logged
n	168	252
S.obs	26	37
C.hat	0.9289	0.9446
f1	12	14

f2	4	4
f3	0	4
f4	1	3
f5	0	1
f6	2	0
f7	0	3
f8	1	2
f9	1	0
f10	0	1

Notes

n = number of observed individuals in the reference sample (sample size).
S.obs = number of observed species in the reference sample.
C.hat = estimator of the sample coverage of the reference sample
f1-f10 = the first ten species abundance frequency counts in the sample.

In the "Interpolation/Extrapolation" tab panel, all relevant diversity estimates (see Notes below) of 40 sample sizes for the Girdled site are shown:

	m	method	order	qD	qD.LCL	qD.UCL	SC	SC.LCL	SC.UCL	site
1	1	interpolated	0	1.00	1.00	1.00	0.12	0.10	0.14	Girdled
2	10	interpolated	0	6.48	6.15	6.80	0.60	0.56	0.64	Girdled
3	19	interpolated	0	9.45	8.84	10.06	0.74	0.70	0.78	Girdled
4	28	interpolated	0	11.51	10.64	12.39	0.80	0.77	0.84	Girdled
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
38	318	extrapolated	0	34.05	25.26	42.84	0.96	0.93	1.00	Girdled
39	327	extrapolated	0	34.40	25.32	43.47	0.96	0.93	1.00	Girdled
40	336	extrapolated	0	34.73	25.37	44.09	0.96	0.93	1.00	Girdled

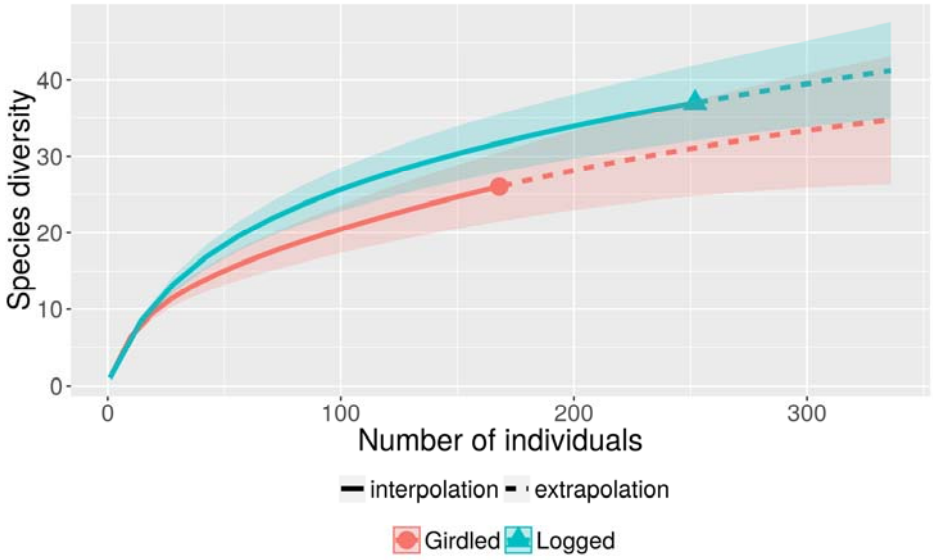
Notes

- m = sample size for which diversity estimates of order q are computed; by default setting (in the left hand side of the screen), m represents the sample size for each of the 40 knots between 1 and the default endpoint (double the reference sample size). On the "General Setting", you can also either specify the endpoint and knots or specify the samples sizes for which you like to calculate diversity estimates.
- method = interpolated, observed, or extrapolated, depending on whether the size m is less than, equal to, or greater than the reference sample size.
- order = the diversity order of q you selected in the "General Setting" on the left hand side of the screen.

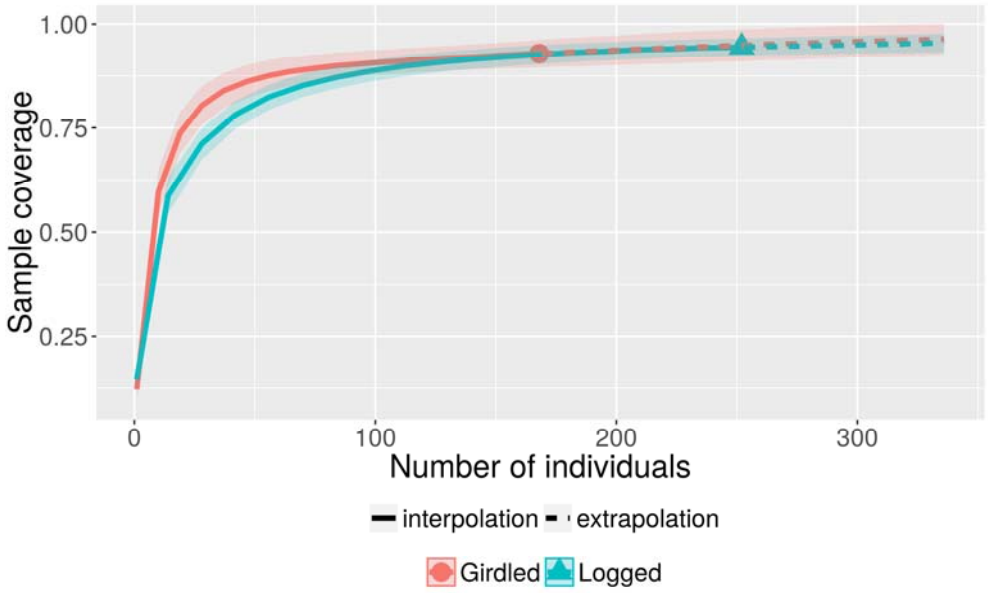
- qD = the estimated diversity of order q for a sample of size m.
- SC = the estimated sample coverage for a sample of size m.
- qD.LCL, qD.UCL = the bootstrap lower and upper confidence limits for the diversity of order q at the specified level in the setting (with a default value of 0.95).
- SC.LCL, SC.UCL = the bootstrap lower and upper confidence limits for the expected sample coverage at the specified level in the setting (with a default value of 0.95).

In the “Figure Plots” tab panel, three types of plots are shown:

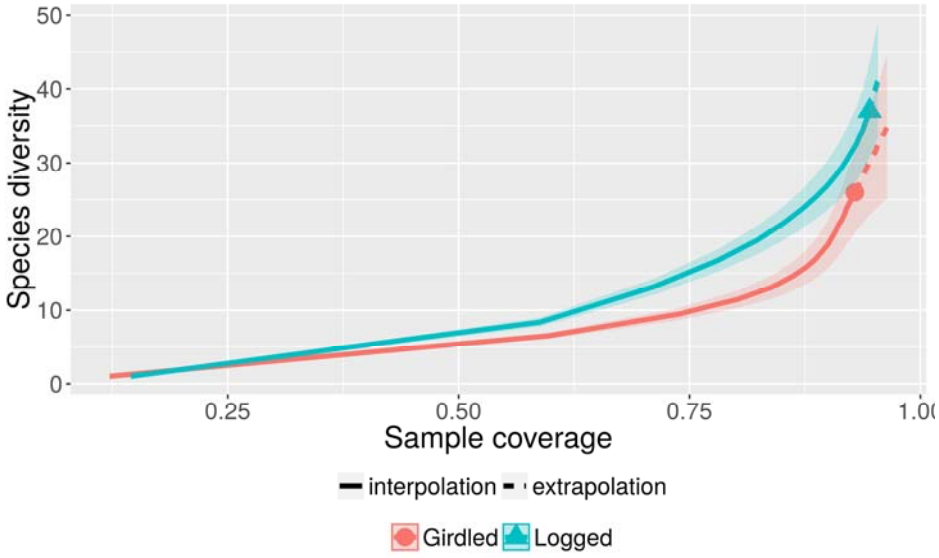
(1) Sample-size-based rarefaction and extrapolation sampling curve:



(2) Sample completeness curve



(3) Coverage-based rarefaction and extrapolation sampling curve



Output for Asymptotic Analysis

Along the second row (output) menu, there are four output selection tabs:

- In the "**Diversity Profile (estimated)**" tab panel, the estimated diversity profiles for diversity order between 0 and 3 are shown based on the statistical method proposed in

Chao and Jost (2015). The figure can be downloaded by clicking "download as PNG file" at the bottom of the displayed figure. Below the profile, all numerical values of diversity (and Tsallis entropy) estimates are tabulated along with s.e. and confidence intervals. These results can be downloaded by clicking "download as CSV file" below the table.

- In the "**Diversity Profile (empirical)"** tab panel, the observed diversity profiles for diversity order between 0 and 3 are shown. Below the profile, all observed values of diversity (and Tsallis entropy) are tabulated along with s.e. and confidence intervals. The figure and table can be downloaded as described earlier for estimated part.

- In the "**Introduction**" panel, users can view a brief introduction to iNEXT Online and a summary of the running procedures.

- In the "**User Guide**" panel, a link will direct users to this user guide.

Example 2: run the demo abundance data (beetles) for Asymptotic Analysis of species richness (See Chao and Jost 2012, 2015 for interpretation)

Step 1. Select Asymptotic Analysis from the top menu.

Step 2. "Data Setting" (on the left hand side of the window)

(2a) Check the Demo data radio button to load demo data.

(2b) Select data type Abundance data.

(2c) There are two demo data sets for abundance data; select Beetles.

(2d) There are two sites (Second Growth and Old Growth); you can select one set or multiple sets for comparison. Here we select both.

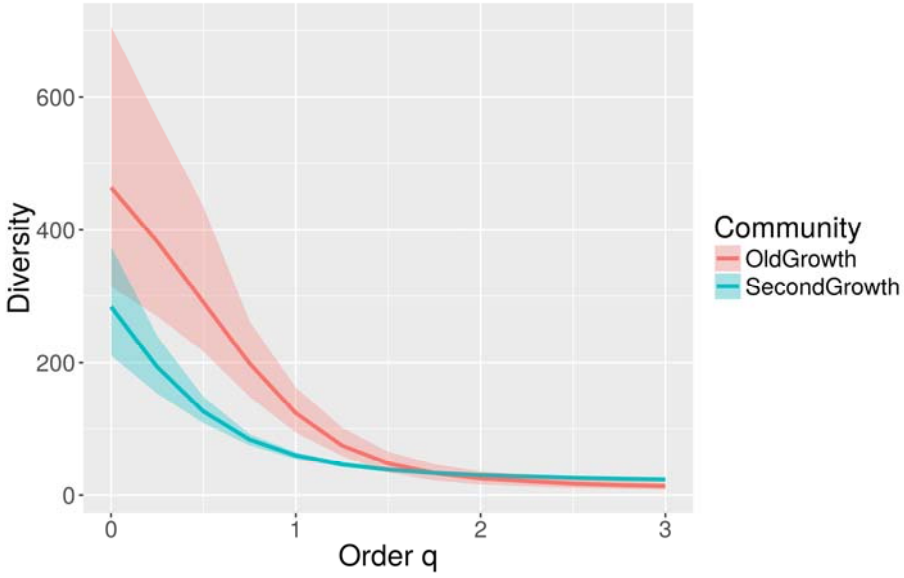
Step 3. "General Setting": use all default settings for species richness; no any selection is needed.

Step 4. Press the Run! button to get the output.

OUTPUT

In the "**Diversity Profile (estimated)**" tab panel, the estimated diversity profiles for diversity order between 0 and 3 are shown based on the statistical method proposed in Chao and Jost (2015), followed by all numerical values for diversity and entropy for each order.

Estimated Diversity Profiles



Estimated Diversity and Entropy for Each Order (Here we only show the estimated diversities)

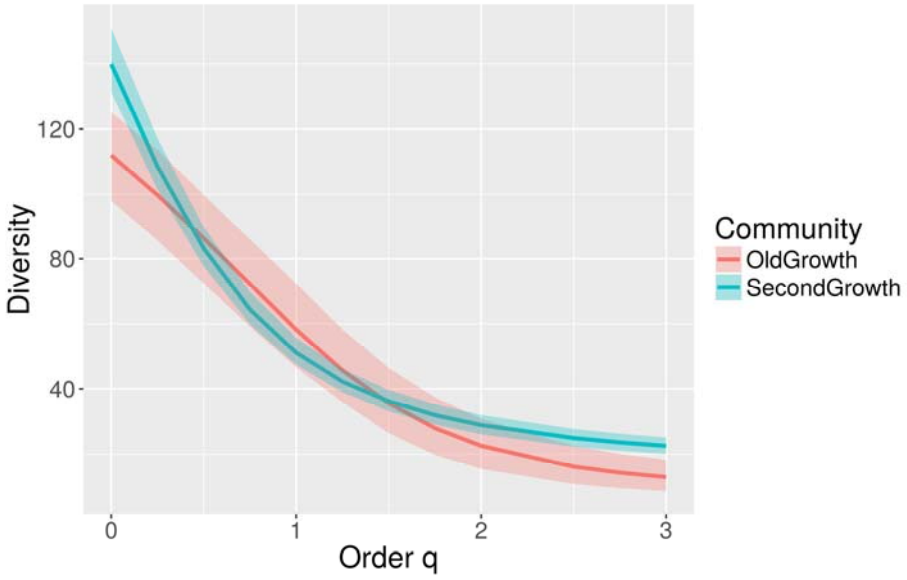
	Order.q	Target	Estimate	s.e.	LCL	UCL	Community
1	0.00	Diversity	283.97	43.60	210.80	375.03	SecondGrowth
2	0.25	Diversity	193.52	23.33	153.15	240.18	SecondGrowth
3	0.50	Diversity	125.98	10.62	108.36	148.41	SecondGrowth
4	0.75	Diversity	83.31	4.79	74.58	91.56	SecondGrowth
5	1.00	Diversity	59.20	2.75	53.55	64.24	SecondGrowth
6	1.25	Diversity	45.82	2.03	41.83	49.11	SecondGrowth
7	1.50	Diversity	38.00	1.71	34.54	40.67	SecondGrowth
8	1.75	Diversity	33.06	1.51	29.98	35.42	SecondGrowth
9	2.00	Diversity	29.73	1.38	26.91	31.74	SecondGrowth
10	2.25	Diversity	27.34	1.54	25.00	30.80	SecondGrowth
11	2.50	Diversity	25.56	1.47	23.36	28.87	SecondGrowth
12	2.75	Diversity	24.19	1.40	22.11	27.37	SecondGrowth
13	3.00	Diversity	23.12	1.35	21.13	26.17	SecondGrowth
14	0.00	Diversity	463.31	138.60	309.25	866.65	OldGrowth
15	0.25	Diversity	382.29	107.26	256.04	696.32	OldGrowth
16	0.50	Diversity	290.54	72.56	195.05	501.77	OldGrowth
17	0.75	Diversity	198.78	40.76	133.68	311.88	OldGrowth
18	1.00	Diversity	123.47	20.07	85.48	173.33	OldGrowth
19	1.25	Diversity	74.34	11.26	52.31	101.32	OldGrowth
20	1.50	Diversity	47.17	7.82	32.57	66.12	OldGrowth
21	1.75	Diversity	32.78	6.00	22.40	45.68	OldGrowth
22	2.00	Diversity	24.81	4.88	16.91	34.48	OldGrowth
23	2.25	Diversity	20.06	4.13	13.68	28.51	OldGrowth
24	2.50	Diversity	17.01	3.61	11.65	24.60	OldGrowth
25	2.75	Diversity	14.95	3.22	10.29	21.86	OldGrowth
26	3.00	Diversity	13.48	2.92	9.33	19.85	OldGrowth

NOTES

- order = the diversity order of q between 0 and 3 in increments of 0.25 (or those orders you type in the input window).
- Target = diversity or entropy
- Estimate = diversity or entropy estimate by the Chao and Jost (2015) method
- LCL, UCL = the bootstrap lower and upper confidence limits for the diversity or entropy of order q at a specified level (default level = 0.95).

In the "**Diversity Profile (empirical)**" tab panel, the observed diversity profiles for diversity order between 0 and 3 are shown as below, followed by all numerical values for diversity and entropy for each order (omitted). The reader is advised to always use the Chao and Jost estimated diversity profile because the empirical values typically underestimate the true diversity and entropy, as this example demonstrated.

Empirical Diversity Profiles



For more sophisticated plots, please use iNEXT R packages available from CRAN; see Hsieh et al. (2016).

References

The following papers for pertinent background on rarefaction/extrapolation and related statistical analyses. These papers can be directly downloaded from Anne Chao's website.

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.

Chao, A. & Jost, L. (2012) Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology*, **93**, 2533–2547.

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.

Colwell, R.K., Chao, A., Gotelli, N.J., Lin, S.-Y., Mao, C.X., Chazdon, R.L. & Longino, J.T. (2012) Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology*, **5**, 3–21.

Hsieh, T.C., Ma, K.H. & Chao, A. (2016) iNEXT: An R package for interpolation and extrapolation of species diversity (Hill numbers). To appear in *Methods in Ecology and Evolution*.