Methods in Ecology and Evolution

BRITISH ECOLOGICAL SOCIETY

Methods in Ecology and Evolution 2016, **7,** 1451–1456 doi: 10.1111/2041-210X.12613

APPLICATION

iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers)

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

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

Summary

1. Hill numbers (or the effective number of species) have been increasingly used to quantify the species/taxonomic diversity of an assemblage. The sample-size- and coverage-based integrations of rarefaction (interpolation) and extrapolation (prediction) of Hill numbers represent a unified standardization method for quantifying and comparing species diversity across multiple assemblages.

2. We briefly review the conceptual background of Hill numbers along with two approaches to standardization. We present an R package iNEXT (iNterpolation/EXTrapolation) which provides simple functions to compute and plot the seamless rarefaction and extrapolation sampling curves for the three most widely used members of the Hill number family (species richness, Shannon diversity and Simpson diversity). Two types of biodiversity data are allowed: individual-based abundance data and sampling-unit-based incidence data.

3. Several applications of the iNEXT packages are reviewed: (i) Non-asymptotic analysis: comparison of diversity estimates for equally large or equally complete samples. (ii) Asymptotic analysis: comparison of estimated asymptotic or true diversities. (iii) Assessment of sample completeness (sample coverage) across multiple samples. (iv) Comparison of estimated point diversities for a specified sample size or a specified level of sample coverage.

4. Two examples are demonstrated, using the data (one for abundance data and the other for incidence data) included in the package, to illustrate all R functions and graphical displays.

Key-words: abundance data, incidence data, sample coverage, Shannon diversity, Simpson diversity, species richness

Introduction

Hill numbers (or the effective number of species) have been increasingly used to quantify the species/taxonomic diversity of an assemblage because they represent an intuitive and statistically rigorous alternative to other diversity indices (see Chao *et al.* 2014, for a recent review). Hill numbers are parameterized by a diversity order q, which determines the measures' sensitivity to species relative abundances. Hill numbers include the three most widely used species diversity measures as special cases: species richness ($q = 0$), Shannon diversity ($q = 1$) and Simpson diversity ($q = 2$).

For a given diversity measure, the goal in many diversity analyses is to make fair comparison and assessment of diversities across multiple assemblages that may vary in the size of their species pools or in the way in which they are sampled. For species richness, it is well known that the empirical species richness in a sample is highly dependent on sample size or sampling efforts. The traditional approach to compare species richnesses of different assemblages is to use rarefaction to down-sample the larger samples until they contain the same number of observed individuals or observations as the smallest sample. Ecologists then compare the richnesses of these equally large samples, but this necessitates that some data in larger samples are thrown away.

To avoid discarding data, Colwell *et al.* (2012) proposed using a sample-size-based rarefaction and extrapolation (R/E) sampling curve for species richness that can be rarefied to smaller sample sizes or extrapolated to a larger sample size, guided by an estimated asymptotic species richness. Chao & Jost (2012) showed that R/E to a given degree of sample completeness was better able to judge the magnitude of the differences in richness among communities, and ranked communities more efficiently, compared to traditional R/E to equal sample sizes. The sample completeness is measured by sample coverage (the proportion of the total number of individuals that belong to the species detected in the sample), a concept originally developed by Alan Turing and I. J. Good in their cryptographic analysis during World War II.

Hill number of any order (including species richness) is also dependent on sample size and inventory completeness. Chao *et al.* (2014) extended Colwell *et al.* (2012) and Chao & Jost (2012) to Hill numbers and proposed the sample-size- and coverage-based R/E of Hill numbers as a unified framework for estimating species diversity, and for making statistical comparisons based on these estimates.

*Correspondence author. E-mail: chao@stat.nthu.edu.tw

Here we first introduce two types of biodiversity data, briefly review the conceptual background of Hill numbers and present two approaches to standardization. We introduce iNEXT (iNterpolation/EXTrapolation), an R package that provides simple functions to compute and plot sample-size and coverage-based R/E sampling curves, along with confidence bands. We focus on the three most widely used members of the family of Hill numbers (q = 0, 1 and 2) based on two types of biodiversity data: abundance data and incidence data. The estimated asymptote along with a confidence interval for each diversity measure is also computed. iNEXT offers three graphic displays. In addition to plots of the sample-size- and coverage-based R/E sampling curves, iNEXT also plots the sample completeness curve: this curve plots the sample completeness (as measured by sample coverage) with respect to sample size. Several applications of iNEXT packages are reviewed. Two examples (one for abundance data and the other one for incidence data) are used to illustrate the use of iNEXT. A quick introduction to iNEXT via examples is provided in the Appendix S1 (Supporting Information) which is now included as a new vignette document for iNEXT in R and can be viewed using the command vignette ("Introduction", package="iNEXT"). This document and detailed information about iNEXT functions are also available at http://chao.stat.nthu.edu.tw/wordpress/software_download/. An online version of iNEXT (https://chao.shinyapps.io/iNEXT/) is also available for users without an R background.

Two types of data

Assume that there are S species in the focal assemblage, where S is unknown. Let $\{p_1, p_2, \ldots, p_S\}$ denote the true, unknown species relative abundances. In most biological surveys, data can be generally classified into two types: individual-based abundance data and sampling-unit-based incidence data, as described below.

ABUNDANCE DATA

For abundance data, the sampling unit is an individual. We assume a *reference sample* of n individuals is selected from the assemblage. Let X_i denote the sample abundance or frequency of the i-th species in the reference sample, i = 1, 2, …, S. Only those species with abundance $X \geq 1$ are detected in the sample. The input data for iNEXT for a single assemblage are the sample abundance vector (X_1, X_2, …, X_S). When there are N assemblages, input data consist of an S by N abundance matrix or N `lists` of species abundances. In iNEXT, this type of abundance data (from 1 to N assemblages) is specified by an argument `datatype="abundance"`.

INCIDENCE DATA

For incidence data, the sampling unit is usually a trap, net, quadrat, plot, or timed survey. A *reference sample* consists of a species-by-sampling-unit incidence matrix with S rows and T columns, where T denotes the number of sampling units; the (i, j) element is 1 if species i is detected in sampling unit j, and 0 otherwise. Let Y_i be the row sum, the number of sampling units in which species i is detected; here Y_i is referred to as the sample *species incidence frequency* and is analogous to X_i in the abundance data. There are two kinds of incidence input data for iNEXT: (i) incidence-raw data: for each assemblage, input data consist of a species-by-sampling-unit matrix; when there are N assemblages, input data consist of N matrices via a `list` object, with each matrix being a species-by-sampling-unit matrix. In iNEXT, this type of data is specified by `datatype="incidence_raw"`. (ii) Incidence-frequency data: input data for each assemblage consist of the number of sampling units (T) followed by the observed incidence frequencies (Y_1, Y_2, …, Y_S). When there are N assemblages, input data consist of an S + 1 by N matrix or N `lists` of species incidence frequencies. The first entry of each column/list must be the total number of sampling units, followed by the species incidence frequencies. In iNEXT, this type of data is specified by `datatype="incidence_freq"`.

Conceptual background

HILL NUMBERS

Here, we briefly review the concept of Hill numbers for abundance data (see Chao *et al.* 2014, for a similar conceptual background based on incidence data). Complete agreement was reached in an *Ecology* forum (Ellison 2010) that Hill numbers should be the diversity measure of choice. Hill (1973) integrated species richness and species relative abundances into a class of diversity measures later called Hill numbers, or effective numbers of species, defined as

$$ {}^qD = \left(\sum_{i=1}^{S} p_i^q\right)^{1/(1-q)}. \qquad \text{eqn 1}$$

When q = 0, ${}^0D$ is simply species richness, which counts *species* equally without regard to their relative abundances. For q = 1, eqn 1 is undefined, but its limit as q tends to 1 is the exponential of the familiar Shannon index, referred to as Shannon diversity (Chao *et al.* 2014):

$${}^1D = \lim_{q\to 1} {}^qD = \exp\left(-\sum_{i=1}^{S} p_i \log p_i\right).$$

The measure for q = 1 counts *individuals* equally and thus counts species in proportion to their abundances; the measure ${}^1D$ can be interpreted as the effective number of common species in the assemblage. The measure for q = 2, referred to as Simpson diversity, discounts all but the dominant species and can be interpreted as the effective number of dominant species in the assemblage.

STANDARDIZATION: SAMPLE-SIZE-BASED R/E

Like species richness, the expected diversity in a sample of size m, denoted as ${}^qD(m)$, m = 1, 2,. . ., is a non-decreasing function of size m (see Chao *et al.* 2014, for the theoretical formula and

its derivation). A diversity accumulation curve depicts ${}^{q}D(m)$ as a function of m. In the special case of $q = 0$, the curve reduces to the familiar species accumulation curve. Based on a reference sample of size n, Chao *et al.* (2014) derived an interpolated diversity estimator ${}^{q}\hat{D}(m)$ for any rarefied sample of size $m < n$ and an extrapolated diversity estimator ${}^{q}\hat{D}(n + m^*)$ for any enlarged sample of size $n + m^*$; see tables 1 and 2 of their paper. The sample-size-based R/E curve includes a rarefaction part (which plots ${}^{q}\hat{D}(m)$ as a function of $m < n$), and an extrapolation part (which plots ${}^{q}\hat{D}(n + m^*)$ as a function of $n + m^*$); both join smoothly at the reference point (n, S_{obs}), where S_{obs} denotes the observed species richness in the reference sample. The confidence intervals based on the bootstrap method developed by Chao *et al.* (2014) also join smoothly.

For species richness, the size in the R/E curve can be extrapolated to at most double or triple the minimum observed sample size, guided by an estimated asymptote. For Shannon diversity and Simpson diversity, if the data are not too sparse, the extrapolation can be reliably extended to infinity to attain the estimated asymptote provided in Chao *et al.* (2014).

STANDARDIZATION: COVERAGE-BASED R/E

The expected sample coverage for a hypothetical sample of size m, $C(m)$, is also a function of m. Chao & Jost (2012) derived an interpolated coverage estimator $\hat{C}(m)$ for any rarefied sample of size $m < n$ and an extrapolated coverage estimator $\hat{C}(n + m^*)$ for any enlarged sample of size $n + m^*$. The coverage-based sampling curve includes a rarefaction part [which plots ${}^{q}\hat{D}(m)$ as a function of $\hat{C}(m)$] and an extrapolation part [which plots ${}^{q}\hat{D}(n + m^*)$ as a function of $\hat{C}(n + m^*)$]; both join smoothly at the reference sample point ($\hat{C}(n)$, S_{obs}), where $\hat{C}(n)$ denotes the estimated sample coverage for the reference sample. The confidence intervals based on the bootstrap method (Chao & Jost 2012) also join smoothly. The curve can be extended to the coverage of the maximum size used in the sample-size-based sampling curve.

Package description

The iNEXT package provides simple functions for computing and plotting seamless R/E sampling curves for Hill numbers. The iNEXT package is available on CRAN (https://cran.r-project.org/web/packages/iNEXT/index.html) and can be downloaded with a standard installation procedure. For first-time installations, an additional visualization extension package (ggplot2) *must be* loaded. The list of the functions in iNEXT and their description are shown in Table 1; we demonstrate the use of these functions in *Examples*.

Applications

The functions in the iNEXT package have been applied to various types of data and have potential to be useful in many research fields. These applications can be classified into four categories as summarized below. In each category, we only provide one representative reference due to space restriction.

1 A non-asymptotic analysis: comparison of estimated diversities of standardized samples with a common sample size or sample completeness. This analysis aims to compare diversity estimates for equally large or equally complete samples; it is based on the seamless rarefaction and extrapolation sampling curves of Hill numbers via our main `iNEXT()` function (see Eren *et al.* 2016, for archaeological data).

2 An asymptotic analysis: comparison of the estimated asymptotic diversities. It is based on statistical estimation of Hill numbers via three functions: `ChaoRichness()`, `ChaoShannon()` and `ChaoSimpson()`; these functions return, respectively, the estimated asymptote for Hill numbers of order $q = 0$, 1 and 2 (see Kendrick *et al.* 2015, for ant data).

3 Assessment of sample completeness of multiple samples via the `iNEXT()` function (see Uchida & Ushimaru 2015, for grassland plants and herbivorous insects data).

4 Comparison of point diversities for a specified sample size or a specified level of sample coverage via the `estimateD()` function (see Mateo-Tomás *et al.* 2015, for vertebrate scavenger data).

Examples

Several data sets are included in the package for demonstration. Here we illustrate all graphical displays using two data sets (`spider` for abundance data and `ant` for incidence data). These data are presented in the supporting information (see

Table 1. List of the functions in the iNEXT package and their description

Type	Function	Description
Main function	`iNEXT()`	Interpolation and extrapolation of Hill numbers
Asymptotic diversity estimation function	`ChaoRichness()`	Estimation of species richness
	`ChaoShannon()`	Estimation of Shannon diversity
	`ChaoSimpson()`	Estimation of Simpson diversity
Point estimation function	`estimateD()`	Estimation of species diversity with a particular sample size/coverage
Graphic displays function	`ggiNEXT()`	`ggplot2` extension for `iNEXT` object to plot rarefaction/extrapolation curves
Others	`DataInfo()`	Summary of basic data information

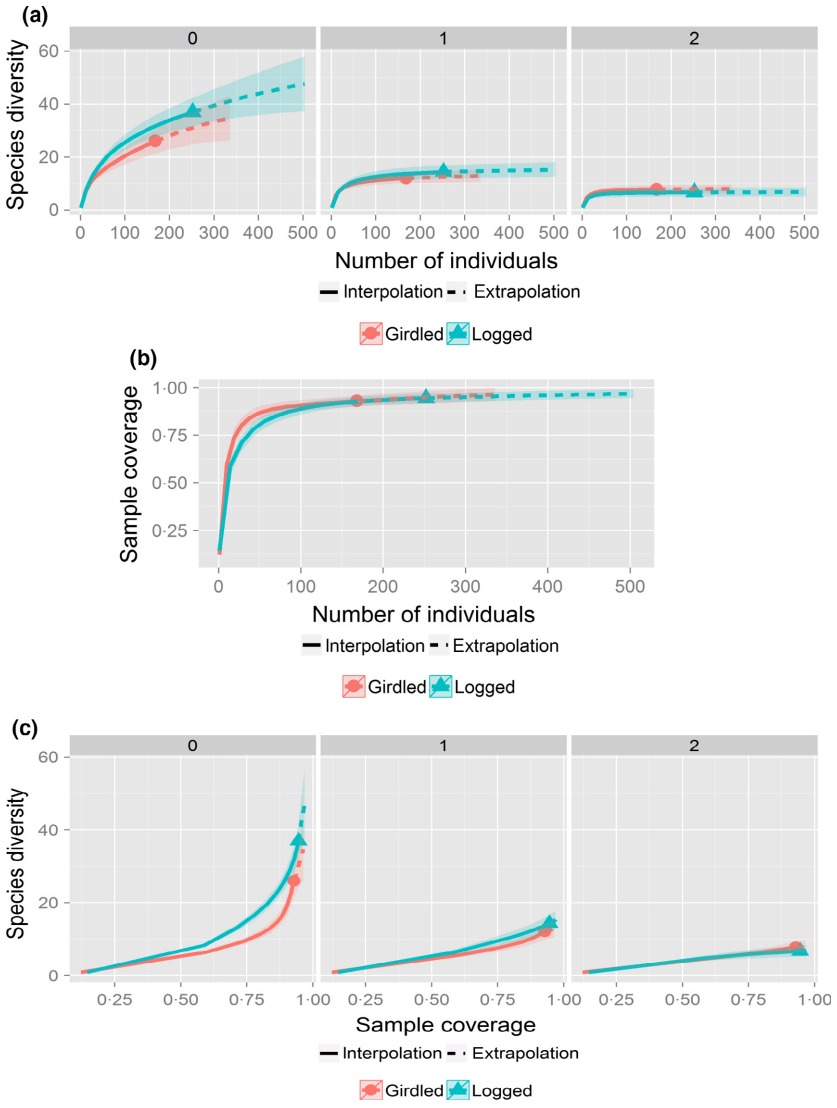

Fig. 1. (a) Sample-size-based and (c) coverage-based rarefaction (solid line segment) and extrapolation (dotted line segments) sampling curves with 95% confidence intervals (shaded areas) for the spider data of two treatments, separately by diversity order: $q = 0$ (species richness, left panel), $q = 1$ (Shannon diversity, middle panel) and $q = 2$ (Simpson diversity, right panel). The solid dots/triangles represent the reference samples. (b) Sample completeness curves linking curves in (a) and (c).

Chao *et al.* 2014, for analysis details and data sources/interpretations). Here we first use the spider data to explain the arguments of the main function `iNEXT()`; the data consist of abundance data from two canopy manipulation treatments ('Girdled' and 'Logged') of hemlock trees.

MAIN FUNCTION: `iNEXT`

The main function `iNEXT()` returns the `"iNEXT"` object, including three data frames: `$DataInfo`, `$iNextEst` and `$AsyEst`, as explained for the `spider` example below.

For the `spider` data, the following code can be run to obtain output after the iNEXT package is installed and the package ggplot2 is loaded (see the Appendix S1 for more details):

```
iNEXT(spider, q=c(0,1,2), datatype="abundance").
```

The first list, `$DataInfo`, summarizes the data set: in the Girdled site, there were 26 species among 168 individuals; in the Logged site, there were 37 species among 252 individuals. In the Girdled treatment site, by default, 40 equally spaced knots (samples sizes) between 1 and 336 (=2 × 168, double the reference sample size) are selected. Diversity estimates and related statistics are computed for these 40 knots (corresponding to sample sizes `m` = 1, 10, 19, …, 336), which locates the reference sample at the mid-point of the selected knots. The list `$iNextEst` (as shown in the Appendix S1) includes diversity estimates along with related statistics for rarefied samples of sizes `m` = 1, 10, …, 159, and also for extrapolated samples of sizes `m` = 177, 186, …, 336. In the list, the sample coverage estimate along with the 95% lower and upper confidence limits is also shown; these estimates and confidence limits are used for plotting the sample completeness curve and coverage-based R/E curves. The list `$AsyEst` shows the asymptotic diversity estimates along with related statistics for Hill numbers of order $q = 0$, 1 and 2. These estimated asymptotes are calculated via the functions `ChaoRichness()`, `ChaoShannon()` and `ChaoSimpson()`.

GRAPHIC DISPLAYS: `ggiNEXT()`

The `ggiNEXT()` function, a wrapper around the ggplot2 package, serves to create a R/E curve. The resulting object can be manipulated using the ggplot2 tools. The following command returns the sample-size-based R/E curve:

```
ggiNEXT(x, type=1, se=TRUE, grey=FALSE, ...)
```

Here `x` is an `"iNEXT"` object. Three types of curves are allowed: (i) sample-size-based R/E curve (`type=1`) with confidence intervals (if `se=TRUE`); see Figs 1a (for $q = 0$, 1 and 2) and 2a (for $q = 0$ only). (ii) Sample completeness curve (`type=2`) with confidence intervals (if `se=TRUE`); see Figs 1b and 2b. This curve plots the sample coverage with respect to sample size. (iii) Coverage-based R/E curve (`type=3`) with confidence intervals (if `se=TRUE`); see Figs 1c (for $q = 0$, 1 and 2) and 2c (for $q = 0$ only). The user may also use the argument `grey=TRUE` to plot black/white figures. Note that ggiNEXT allows `ggplot2` functions such as `xlim()`, `ylim()`, `theme()` and `theme_bw()` to be used to modify the display settings (see the Appendix S1 for examples).

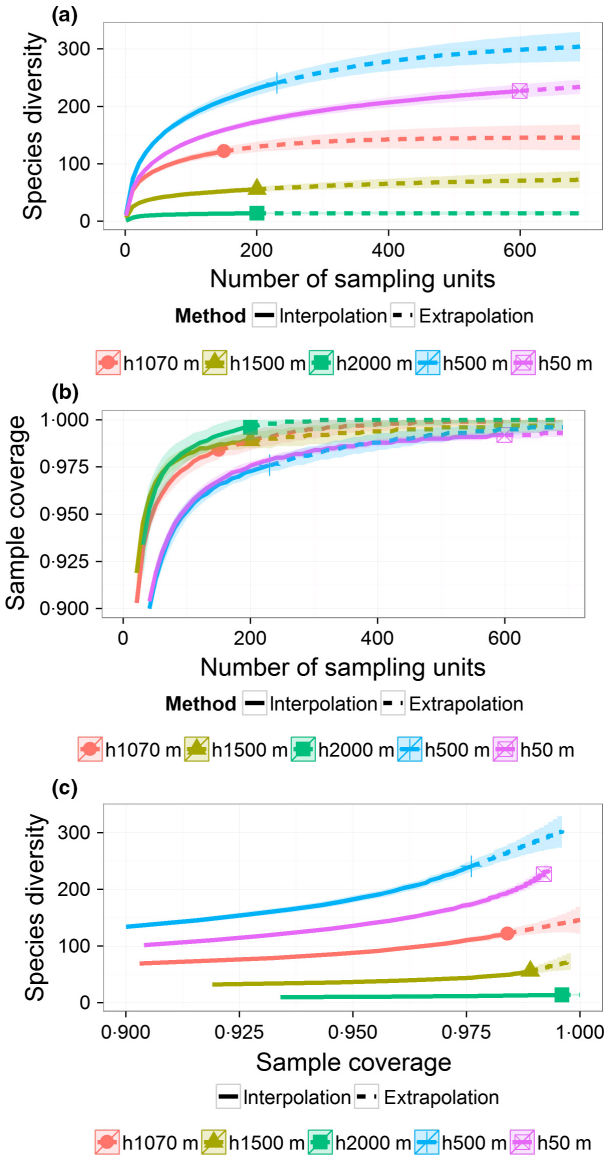

Fig. 2. (a) Sample-size-based and (c) coverage-based rarefaction (solid line segment) and extrapolation (dotted line segments) sampling curves for species richness ($q = 0$) with 95% confidence intervals (shaded areas) for the tropical ant data at five elevations. The solid dots and the other four symbols represent the reference samples. (b) Sample completeness curves linking curves in (a) and (c). `iNEXT` offers a customized graphic theme to change grey background to black and white (see the Appendix S1 for details).

INCIDENCE DATA

We use the tropical ant data (in the file `ant`) at five elevations (50, 500, 1070, 1500 and 2000 m) collected by Longino & Colwell (2011) in Costa Rica for illustration. The first entry of each list must be the total number of sampling units. Figure 2 shows the three types of sampling curves for species richness without grey backgrounds. Details are omitted here due to space restrictions.

POINT ESTIMATION FUNCTION: `estimateD()`

We also supply the function `estimateD()` to compute diversity estimates with $q = 0$, 1, 2 (all three levels of q are reported) for any particular level of sample size or any specified level of sample coverage for either abundance data or incidence data. For example, the following command returns the species diversity with a specified level of sample coverage of 98·5% for the ant data, along with 95% confidence intervals.

```
estimateD(ant, datatype="incidence_freq",
base="coverage", level= 0.985, conf=0.95)
```

See Appendix S1 for details.

Alternative software

There is alternative software and R functions that provide similar tools for rarefaction and extrapolation curves.

1 The freeware EstimateS (Colwell 2013) with a full graphical user interface obtains R/E sampling curves with confidence intervals for both abundance and incidence data. All these tools in EstimateS are designed for species richness. iNEXT is more comprehensive because iNEXT also provides the corresponding output for Shannon diversity and Simpson diversity. EstimateS is a GUI interface, which makes it hard to do reproducible science with it, whereas iNEXT R package does do that. An online version of iNEXT is also available for users without an R background (see the Introduction).

2 The function `"rarefy"`, available in the R package `vegan` (Oksanen *et al.* 2015), provides rarefaction curves for species richness, but this function does not include extrapolation.

Conclusion

We have reviewed the standardization methods for Hill numbers, presented the iNEXT package and illustrated the use of iNEXT in constructing two types (sample-size- and coverage-based) of rarefaction and extrapolation curves with Hill numbers, along with a sample completeness curve that links the

two types of curves. For each type of curve, the sampling curves with confidence intervals for species richness, Shannon diversity and Simpson diversity are suggested to quantify and compare species diversities in a unified framework. Figures 1 and 2, respectively, show the sampling curves for abundance data and incidence data. The package iNEXT provides an easy-to-use interface and efficiently uses all data to make more robust and detailed inferences about the sampled assemblages, and also to make objective comparisons of multiple assemblages. iNEXT will be soon extended to its phylogenetic generalization, `iNextPD` (https://github.com/JohnsonHsieh/iNextPD), for analysing phylogenetic data.

Acknowledgements

The authors thank the Editor (Jana Vamosi), an Associate Editor, Robert Colwell, Jonathan Lefcheck, Scott Chamberlain and an anonymous reviewer for very helpful and thoughtful comments and suggestions. This research is supported by Taiwan Ministry of Science and Technology under Contract 103-2628-M007-007. TCH was supported by a post-doctoral fellowship, Taiwan Ministry of Science and Technology, Taiwan.

Data accessibility

All data used in this paper are presented in the supporting information.

References

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

Chao, A., Gotelli, N.J., Hsieh, T.C., Sander, E.L., Ma, K.H., Colwell, R.K. & 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.

Colwell, R.K. (2013) EstimateS: Statistical estimation of species richness and shared species from samples. Version 9 and earlier. User's Guide and application. Available at: http://purl.oclc.org/estimates.

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.

Ellison, A.M. (2010) Partitioning diversity. *Ecology*, **91**, 1962–1963.

Eren, M.I., Chao, A., Chiu, C.-H., Colwell, R.K., Buchanan, B., Boulanger, M.T., Darwent, J. & O'Brien, M.J. (2016) Statistical Analysis of paradigmatic class richness supports greater paleoindian projectile-point diversity in the Southeast. *American Antiquity*, **81**, 174–192.

Hill, M.O. (1973) Diversity and evenness: a unifying notation and its consequences. *Ecology*, **54**, 427–432.

Kendrick, J.A., Ribbons, R.R., Classen, A.T. & Ellison, A.M. (2015) Changes in canopy structure and ant assemblages affect soil ecosystem variables as a foundation species declines. *Ecosphere*, **6**, art770.

Longino, J.T. & Colwell, R.K. (2011) Density compensation, species composition, and richness of ants on a neotropical elevational gradient. *Ecosphere*, **2**, art29.

Mateo-Tomás, P., Olea, P.P., Moleón, M., Vicente, J., Botella, F., Selva, N., Viñuela, J. & Sánchez-Zapata, J.A. (2015) From regional to global patterns in vertebrate scavenger communities subsidized by big game hunting. *Diversity and Distributions*, **21**, 913–924.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B. *et al.* (2015) The vegan package. Community ecology package. R package version 2.3-2. Available at: http://CRAN.R-project.org/package = vegan

Uchida, K. & Ushimaru, A. (2015) Land abandonment and intensification diminish spatial and temporal β-diversity of grassland plants and herbivorous insects within paddy terraces. *Journal of Applied Ecology*, **52**, 1033–1043.

Received 18 January 2016; accepted 16 June 2016
Handling Editor: Greg McInerny

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Appendix S1. A quick introduction to iNEXT via examples.