Proportional mixture of two rarefaction/extrapolation curves to forecast biodiversity changes under landscape transformation

Anne Chao,[1*] Robert K. Colwell, [2,3] Nicholas J. Gotelli,[4] and Simon Thorn [5]

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

[2] Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269, USA

[3] University of Colorado Museum of Natural History, Boulder, CO 80309, USA

[4] Department of Biology, University of Vermont, Burlington, VT 05405, USA

[5] Field Station Fabrikschleichach, Biocenter, University of Würzburg, Glashüttenstr. 5, 96181 Rauhenebrach, Germany.

[*] **Correspondence**: Anne Chao, E-mail: chao@stat.nthu.edu.tw, 886-3-5711736.

Keywords: abundance data, biodiversity, extrapolation, habitat transformation, habitat modification, Hill number, incidence data, mixture, rarefaction.

Running title: Mixture of rarefaction/extrapolation curves.

Email addresses of all authors: Anne Chao (chao@stat.nthu.edu.tw), Robert K. Colwell (robertkcolwell@gmail.com), Nicholas J. Gotelli (Nicholas.Gotelli@uvm.edu), Simon Thorn (simon@thornonline.de)

Type of article: Methods

Number of words in the abstract: 149

Number of words in the main text: 4996

Number of words in Box 1: 740

Number of references: 30

Number of tables: 1

Number of figures: 3

Statement of authorship: ST motivated the project and collected the principal data. All authors conceived the statistical ideas and developed the methodology. AC and ST analyzed the data and developed the software; RKC and NJG refined the analysis and provided additional perspectives. AC and ST wrote the first draft of the paper; all authors contributed critically to manuscript writing and revisions.

Data accessibility statement: Data are available from the Dryad Digital Repository https://doi.org/10.5061/dryad.t7m9806. The R code "miNEXT" (mixed iNterpolation/EXTrapolation) is available at Anne Chao's Github (https://github.com/AnneChao).

Conflict of interest: The authors declare no conflict of interest.

Abstract

Progressive habitat transformation causes global changes in landscape biodiversity pattern but can be hard to quantify. Rarefaction/extrapolation approaches can quantify within-habitat biodiversity but may not be useful for cases in which one habitat type is progressively transformed to another habitat type. To quantify biodiversity patterns in such transformed landscapes, we use Hill numbers to analyze individual-based species abundance data or replicated, sample-based incidence data. Given biodiversity data from two distinct habitat types, when a specified proportion of original habitat is transformed, our approach utilizes a proportional mixture of two within-habitat rarefaction/extrapolation curves to predict biodiversity changes, with bootstrap confidence intervals assessing sampling uncertainty. We also derive analytic formulas for assessing species composition (i.e., the numbers of shared and unique species) for any mixture of the two habitat types. Our analytical and numerical analyses revealed that species unique to each habitat type are the most important determinants of landscape biodiversity pattern.

INTRODUCTION

Explaining landscape patterns of biodiversity has remained a central research aim of ecology for decades (Gaston 2000). In addition to abiotic factors such as precipitation and solar radiation (Hawkins *et al*. 2003), the quality and spatial organization of habitat patches in a landscape has major influences on biodiversity (Fahrig 1997). The progressive transformation of habitat patches from one habitat type into another habitat type causes significant changes in landscape patterns of biodiversity globally (Foley *et al.* 2005). Examples include not only the transformation from forests to open habitats by natural disturbances (Kulakowski *et al*. 2016) or the regeneration of mature forest from second-growth forest by natural succession (Chazdon 2014), but also anthropogenic habitat transformation (Newbold *et al.* 2015) by selective logging, clear-cutting, assisted regeneration of forest patches, or prescribed burning of forests (Betts *et al.* 2017).

Habitat transformations involve landscapes composed of habitat patches of distinct types. The biodiversity of such landscapes is hence a mixture of the assemblages of original and transformed habitat patches and changes through times with shifting proportions of original and transformed patches (Prevedello *et al*. 2016). Tracking biodiversity changes caused by the patch-wise transformation of original habitat patches to transformed habitat patches is difficult in empirical field studies, and most studies have been based on categorical comparisons of original versus transformed habitats (reviewed in McGill *et al*. 2015).

A powerful strategy for quantifying and comparing landscape patterns of biodiversity links classical rarefaction seamlessly with extrapolation curves for Hill numbers or the exponent of Rényi entropies; see Gotelli & Colwell (2001), Colwell *et al.* (2012) and Chao *et al.* (2014) for reviews. Hill numbers are parameterized by a diversity order q, which determines the measures' sensitivity to species relative abundances. Hill numbers for order $q \geq 0$ are all in units of "species" or "species equivalents" and include the three most widely

used species diversity measures as special cases: species richness (q=0), Shannon diversity (q=1, i.e., the exponential of Shannon entropy) and Simpson diversity (q=2, i.e., the inverse of Simpson concentration index); see Hill (1973) and Chao *et al.* (2014) for a review. Until now, comparisons based on Hill numbers have been performed, first, for each assemblage separately, followed by comparisons of within-assemblage curves across multiple assemblages.

Here, we develop a proportional mixture of two rarefaction/extrapolation curves derived from within-habitat biodiversity sample data from two distinct habitat types. This blended curve can then be used to analytically predict the composite biodiversity of a landscape comprising different proportions of original and transformed habitat patches. We also derive analytic formulas to forecast compositional change, i.e. to track the number of shared species and unique species in a mixed landscape. Here "shared" and "unique" species are defined based on the respective within-habitat data. Furthermore, we propose a bootstrap method for assessing the sampling uncertainty for our estimators/formulas. We use a simple example to introduce our conceptual framework (Box 1), illustrating the construction of the mixture curve and species composition analysis, with empirical biodiversity samples for ground-dwelling spiders (abundance data) and songbirds (incidence data) collected in disturbed and undisturbed mountain forests. The important role of species unique to each habitat type in landscape biodiversity change is highlighted, and an extension to a phylogenetic version of the mixture strategy is discussed.

MATERIAL AND METHODS

A conceptual example

****

Box 1. A conceptual example for species richness based on abundance data.

Suppose a random sample of $n_1 = 200$ individuals of an assemblage is collected from patches of original habitat, and a random sample of $n_2 = 160$ individuals is collected from interspersed patches of transformed habitat. The two habitats are sampled in the same way and the abundance of each species is recorded for each habitat. Following the terminology of Colwell *et al.* (2012), we refer to the sample datasets taken from each of the two habitats as *reference samples* (Figure 1, black dot for the original habitat and blue dot for the transformed habitat).

Based on within-habitat reference sample data, we construct a rarefaction curve (Figure 1, black solid line) for the original habitat and an integrated rarefaction/extrapolation curve for the transformed habitat (Figure 1, blue solid line for the interpolated—rarefaction—part of the curve and blue dashed line for the extrapolated part of the curve). Here the plot is specifically for species richness (diversity of order q=0); the plots for Shannon diversity (q=1) and Simpson diversity (q=2) are presented in later sections. Changes in composite diversity—as an original habitat is progressively transformed to another habitat—will depend on the species composition and abundance in both habitat types. Hence, we ask the following question: if a specified proportion of the original habitat is transformed (equivalently, a specified proportion of individuals in the original habitat are replaced by individuals from the transformed habitat), what would be the resulting diversity after the transformation? This question can be formulated as follows: If hypothetically we choose randomly m_1 individuals from the original habitat and m_2 individuals from the transformed habitat, what would be the diversity in the combined sample of size $m_1 + m_2$? We consider a special hypothetical size $m_1 + m_2 = 200$ (matching the reference sample size from the currently original habitat), although our framework works for any pair of sample sizes (m_1, m_2). In the case $m_1 \le n_1$ and $m_2 \le n_2$, as in classic rarefaction theory, we could conduct an algorithmic approach to

assess the resulting diversity based on the two reference samples, but we provide here general analytic formulas that can be applied to all cases. The two hypothetical sample sizes (m_1, m_2) for a specified level of transformation of the currently original habitat will be clear in our simple example by considering the four cases indicated as "A, B, C, D" in Table 1 and Figure 1.

(Case A) When there is no transformation in an original habitat ($m_2 = 0, m_1 = n_1$), the expected diversity of the combined sample of size $m_1 + m_2 = n_1 = 200$ is simply estimated by the diversity of the reference sample of the original habitat (solid black dot).

(Case B) If a proportion 90% of the original habitat were retained but 10% were transformed, then we hypothetically take a random sample of size $m_1 = 180$ individuals (90% of 200) from the original habitat and simultaneously a random sample of size $m_2 = 20$ individuals (10% of 200) from the transformed habitat. This represents one of those cases that the hypothetical sample size in each habitat is less than the corresponding reference sample size (solid red line, up to 80% habitat transformation). The composite diversity for such a mixed sample of size 200 can be analytically estimated from a mixture of the two individual rarefaction curves.

(Case C) If a proportion 90% of the original habitat were transformed (i.e., only 10% of the original habitat were retained), we take a hypothetical sample of size $m_1 = 20$ individuals from the original habitat and simultaneously a hypothetical sample of size $m_2 = 180$ individuals from the transformed habitat. Because the hypothetical abundance in the transformed habitat portion exceeds the within-transformed-habitat reference sample size (i.e., 160), the resulting diversity depends on rarefaction in the original habitat and extrapolation in the transformed habitat. In this case, a mixture of rarefaction in the original

habitat and extrapolation in the transformed habitat is required to obtain the composite diversity (red dashed line in Figure 1).

(Case D) If the original habitat is completely transformed ($m_1 = 0$), we sample $m_2 = 200$ individuals from the transformed habitat only. The resulting diversity becomes the extrapolated value of a hypothetical size 200 in the transformed habitat, based on a reference sample of size 160.

In this simple illustrative example, the reference sample size for the original habitat is larger than that for the transformed habitat, i.e., $n_1 > n_2$. In the case that $n_1 \le n_2$, only the proportional mixture of two rarefaction curves is involved. See the section *Estimators of composite diversity* for the confidence bands of the mixture curve in Figure 1.

(End of Box 1)

****

(Table 1, Figure 1)

Method for abundance data

Assume that Assemblage I occupies original habitat and Assemblage II occupies transformed habitat. The two assemblages may differ not only in their species richness, but also in their species composition. Assume that there are *S* species in the *pooled* assemblage. Here *S* is an unknown parameter. The species relative abundances or frequencies in Assemblages I and II are denoted by $(p_{11}, p_{21}, ..., p_{S1})$ and $(p_{12}, p_{22}, ..., p_{S2})$, respectively, $p_{i1}, p_{i2} \ge 0$, $i = 1, 2, ..., S$. A reference sample is taken from each of the two assemblages (reference sample I with size n_1 from Assemblage I and reference sample II with size n_2 from Assemblage II). Denote the observed species abundances in the two reference samples, respectively, by $(X_{11}, X_{21}, ..., X_{S1})$ and $(X_{12}, X_{22}, ..., X_{S2})$. Assume that the observed

species sample frequencies $(X_{1k}, X_{2k}, ..., X_{Sk})$ follow a multinomial model with cell totals n_k and cell probability vector $(p_{1k}, p_{2k}, ..., p_{Sk}), k = 1, 2.$ In our framework, "shared" and "unique" species are defined by these two within-habitat reference samples.

Proposed model and theoretical formulas for the diversity of a mixed habitat

Chao *et al.* (2014) derived analytic rarefaction/extrapolation methods for Hill numbers and developed the software iNEXT (iNterpolation/EXTrapolation) for implementation for *q*=0, 1 and 2. In formulating the species diversity (Hill numbers) of any rarefied or extrapolated sample in a single assemblage, they defined the expected diversity ${}^qD(m)$ as the Hill numbers based on the mean abundance frequency counts for a hypothetical sample of size *m*. In our extension to the two-assemblage case, as demonstrated in Table 1, our goal is to assess the expected diversity for a combined sample of size $m_1 + m_2$ when a hypothetical sample I with size m_1 and a hypothetical sample II with size m_2 are respectively taken from Assemblages I and II. Although we restrict our example to the special case $m_1 + m_2 = n_1,$ the following derivation is valid for any $m_1, m_2 \geq 0.$

For any species represented by exactly k_1 individuals in the hypothetical sample I of size m_1 and exactly k_2 individuals in the hypothetical sample II of size m_2, $k_1 = 0, 1, ..., m_1,$ $k_2 = 0, 1, ..., m_2,$ its relative frequencies in the two hypothetical samples are, respectively, $g_{k_1} \equiv g_{k_1}(m_1) = k_1 / m_1$ and $h_{k_2} \equiv h_{k_2}(m_2) = k_2 / m_2$. Letting $\alpha = m_1 /(m_1 + m_2)$ and $1 - \alpha = m_2 /(m_1 + m_2)$ represent, respectively, the proportions of the original and transformed habitats in the landscape, we can write the frequency of the species in the mixed sample as a mixture of the two individual frequencies, i.e., $\alpha g_{k_1} + (1-\alpha) h_{k_2} = (k_1 + k_2)/(m_1 + m_2).$ Assume that there are $f_{k_1,k_2}(m_1, m_2)$ such species in the mixed sample. We refer to the number $f_{k_1,k_2}(m_1, m_2)$ as the joint abundance frequency count of (k_1, k_2) and denote

${}^{q}D(m_1, m_2)$ as the Hill numbers based on the mean joint abundance frequency counts for a combined sample of size $m_1 + m_2$.

Following a similar argument in Chao *et al.* (2014), we can express the theoretical formula for the composite diversity ${}^{q}D(m_1, m_2)$ as

$$ {}^{q}D(m_1, m_2) = \left(\sum_{\substack{k_2=0 \\ k_1+k_2 \geq 1}}^{m_2} \sum_{k_1=0}^{m_1} \left(\frac{k_1 + k_2}{m_1 + m_2} \right)^{q} \times E[\, f_{k_1,k_2}(m_1, m_2)] \right)^{\frac{1}{1-q}}, \quad q \neq 1. \qquad (1) $$

For q=1, we define the composite diversity of order one as

$$ {}^{1}D(m_1, m_2) = \lim_{q \to 1} {}^{q}D(m_1, m_2) $$

$$ = \exp\left(- \sum_{\substack{k_2=0 \\ k_1+k_2 \geq 1}}^{m_2} \sum_{k_1=0}^{m_1} \left(\frac{k_1 + k_2}{m_1 + m_2} \right) \times \log\left(\frac{k_1 + k_2}{m_1 + m_2} \right) \times E[\, f_{k_1,k_2}(m_1, m_2)] \right). \qquad (2) $$

The mean joint abundance frequency count can be expressed as the following sum of the products of two binomial probabilities:

$$ E\,[\, f_{k_1,k_2}(m_1, m_2)] = \sum_{i=1}^{S} \binom{m_1}{k_1} p_{i1}^{k_1} (1 - p_{i1})^{m_1 - k_1} \times \binom{m_2}{k_2} p_{i2}^{k_2} (1 - p_{i2})^{m_2 - k_2}. \qquad (3) $$

Combining Eqs. (1), (2) and (3), we then obtain the expected or theoretical formula of our estimating target ${}^{q}D(m_1, m_2)$ in terms of the two sets of species relative abundances. The resulting formula, below, for species richness (q=0) depicts the expected two-assemblage species accumulation curve:

$$ {}^{0}D(m_1, m_2) = \sum_{i=1}^{S} [1 - (1 - p_{i1})^{m_1} (1 - p_{i2})^{m_2}]. \qquad (4a) $$

The term inside the brackets of the above formula denotes the probability that species *i* is observed in at least one of the two hypothetical samples; see Appendix S1 for derivation. The above theoretical formulas (Eqs. 1–4a) are valid for *any* pair of hypothetical sample sizes (m_1, m_2), $m_1, m_2 \geq 0$. In the special case of m_2 =0, the resulting formula ${}^{q}D(m_1, 0)$ reduces to

the one-assemblage equation (Chao *et al.* 2014), i.e., the expected number of species that would be observed if a hypothetical sample of m_1 individuals were taken from Assemblage I, disregarding the data from Assemblage II. A similar interpretation is valid for ${}^{q}D(0,m_2)$.

Estimators of composite diversity

In practice, the above formulas (Eqs. 1–4a) should be estimated from the data of the two reference samples. These estimators depend on whether m_2 is less than the reference sample size n_2. We briefly summarize the estimation separately for the two cases and derive all formulas in Appendix S1:

(i) Mixture of two rarefaction curves (Case B in Table 1, i.e., $m_1 \le n_1$ and $m_2 \le n_2$). For this mixture, an unbiased estimator of species richness in the mixed sample represents a generalization of the classic one-assemblage rarefaction to two-assemblages and can be expressed as the following sum of the estimates of the probabilities in Eq. (4a):

$${}^{0}\hat{D}(m_1,m_2) = \sum_{X_{i1}+X_{i2}\ge 1}\left(1-\frac{\binom{n_1-X_{i1}}{m_1}}{\binom{n_1}{m_1}}\times\frac{\binom{n_2-X_{i2}}{m_2}}{\binom{n_2}{m_2}}\right). \qquad (4b)$$

Also, nearly unbiased estimators exist for ${}^{q}D(m_1,m_2)$, for any q>0; see Appendix S1 for details.

(ii) Mixture of one rarefaction curve and one extrapolation curve (Case C in Table 1, i.e., $m_1 \le n_1$ but $m_2 > n_2$). When one extrapolation curve is involved, a nearly unbiased estimator for ${}^{q}D(m_1,m_2)$ exists for q=2 only, because the measure for q=2 depends on the data of the dominant species and dominant species nearly always appear in any sample. Thus, information is sufficient to estimate any measure that focuses on dominant species. For q=0 and q=1, estimation bias arises mainly due to the effect of undetected rare species in extrapolated samples. If data are sufficient (say, at least two-thirds of species are observed in

each reference sample, based on previous simulation work), then the Horvitz-Thompson (1952) adjustment for undetected species and the Chao *et al.* (2015) adjustment of the sample relative abundance can be applied to obtain approximate estimates of ${}^{q}D(m_1,m_2)$; see Appendix S1 for details.

Species composition information in a mixed sample

When an original habitat is progressively transformed, we can assess the species composition in a mixed landscape comprised of patches of original and transformed habitats. When "shared" and "unique" species are defined by the two within-habitat reference samples, it follows from Eq. (4b) that species richness estimator ${}^{0}\hat{D}(m_1,m_2)$ can be decomposed into the sum of three components: ${}^{0}\hat{D}(m_1,m_2) = {}^{0}\hat{D}_{shared}(m_1,m_2) + {}^{0}\hat{D}_{unique\,1}(m_1,m_2) + {}^{0}\hat{D}_{unique2}(m_1,m_2)$. Here the first term denotes an unbiased estimator of the number of shared species in a combined sample of size m_1+m_2, $m_1 \le n_1$ and $m_2 \le n_2$:

$$ {}^{0}\hat{D}_{shared}(m_1,m_2) = \sum_{X_{i1}\ge 1\ \&\ X_{i2}\ge 1} \left(1 - \frac{\binom{n_1 - X_{i1}}{m_1}}{\binom{n_1}{m_1}} \times \frac{\binom{n_2 - X_{i2}}{m_2}}{\binom{n_2}{m_2}} \right); \tag{5a} $$

the second term denotes an unbiased estimator of species unique to the original habitat:

$$ {}^{0}\hat{D}_{unique1}(m_1,m_2) = \sum_{X_{i1}\ge 1\ \&\ X_{i2}=0} \left(1 - \frac{\binom{n_1 - X_{i1}}{m_1}}{\binom{n_1}{m_1}} \right); \tag{5b} $$

and the third term denotes an unbiased estimator of the number of species unique to the transformed habitat:

$$ {}^{0}\hat{D}_{unique2}(m_1,m_2) = \sum_{X_{i1}=0\ \&\ X_{i2}\ge 1} \left(1 - \frac{\binom{n_2 - X_{i2}}{m_2}}{\binom{n_2}{m_2}} \right). \tag{5c} $$

The above decomposition can be evaluated only for the mixture of two rarefaction curves because the identities of "shared" and "unique" species are not available for any extrapolated sample.

Our analytic estimators for the composite diversity and species composition in any mixed sample are complicated functions of within-habitat species abundance frequencies. It is thus not possible to derive analytic variance estimators. Here, we generalize the one-assemblage bootstrap method (Chao *et al*. 2014) to a two-assemblage version that can be applied to obtain approximate variance for any proposed estimator in this paper and construct the associated confidence intervals to reflect sampling uncertainty (see Appendix S1 for details).

Importance of unique species for changes in species richness

Based on Eqs. (5a)–(5c), we can analytically evaluate the rate of change in species richness as transformation proceeds to increase (i.e., the hypothetical size m_1 is progressively reduced and the corresponding size m_2 is increased). Let $\hat{p}_{i1} = X_{i1}/n_1$ and $\hat{p}_{i2} = X_{i2}/n_2$ denote the sample frequencies of species *i* in reference samples I and II, respectively. For unique species from Assemblage I (i.e., species with $\hat{p}_{i1} > \hat{p}_{i2} = 0$), Eq. (5b) leads to the following negative rate of change for a unit-change in the hypothetical sample sizes:

$$ {}^0\hat{D}_{unique1}(m_1-1,m_2+1) - {}^0\hat{D}_{unique1}(m_1,m_2) \approx \sum_{X_{i1}\geq 1\ \&\ X_{i2}=0} -(1-\hat{p}_{i1})^{m_1-1}\hat{p}_{i1} < 0. \qquad (6a) $$

For unique species from Assemblage II (i.e., species with $\hat{p}_{i2} > \hat{p}_{i1} = 0$), Eq. (5c) leads to the following positive rate of change:

$$ {}^0\hat{D}_{unique2}(m_1-1,m_2+1) - {}^0\hat{D}_{unique2}(m_1,m_2) \approx \sum_{X_{i1}=0\ \&\ X_{i2}\geq 1} (1-\hat{p}_{i2})^{m_2}\hat{p}_{i2} > 0. \qquad (6b) $$

Eq. (6a) quantifies the decline rate in species richness due to the loss of unique species from the original habitat, whereas Eq. (6b) quantifies the increase rate in species richness due to the addition of unique species from the transformed habitat.

Eq. (5c) leads to the corresponding rate of change for shared species:

$$ {}^{0}\hat{D}_{shared}(m_1-1,m_2+1) - {}^{0}\hat{D}_{shared}(m_1,m_2) \approx \sum_{X_{i1}\geq 1 \,\&\, X_{i2}\geq 1} (1-\hat{p}_{i1})^{m_1-1}(1-\hat{p}_{i2})^{m_2}(\hat{p}_{i2}-\hat{p}_{i1}). \qquad (6c) $$

The above difference vanishes for shared species with $\hat{p}_{i2} = \hat{p}_{i1}$. Thus, the change for shared species richness arises mainly from the following two groups of species:

(i) For shared species with $\hat{p}_{i1} > \hat{p}_{i2} > 0,$ the difference in Eq. (6c) is negative, and thus the richness for this group of shared species declines as the proportion transformed increases. Shared richness declines because, when a shared species from the original habitat is lost in the transformation, the same species in the transformed habitat cannot "compensate" the loss due to lower abundance. Note that the absolute decline rate for a species in this group is $(1-\hat{p}_{i1})^{m_1-1}(1-\hat{p}_{i2})^{m_2}(\hat{p}_{i1}-\hat{p}_{i2}),$ which for any fixed value of $\hat{p}_{i1}$ becomes larger as $\hat{p}_{i2}$ decreases. When $\hat{p}_{i2}$ reaches its lower limit 0 so that $\hat{p}_{i1} > \hat{p}_{i2} = 0$ (i.e., a unique species), the magnitude of the rate of change approaches the maximum value $(1-\hat{p}_{i1})^{m_1-1}\hat{p}_{i1},$ i.e., the absolute decline rate derived in Eq. (6a). Consequently, unique species from the original habitat generally are the determinants of the absolute decline rate in species richness.

(ii) For shared species with $\hat{p}_{i2} > \hat{p}_{i1} > 0$, the difference in Eq. (6c) is positive, and thus the richness for this group of shared species increases as the proportion transformed increases. Likewise, unique species from the transformed habitat generally determine the overall rate of increase in species richness.

The absolute decline/increase rates for shared species richness in each of the above two groups are relatively low, implying shared species in any mixture of two rarefactions remains

nearly unchanged, despite of the transformations. Moreover, the two-direction change rates (negative for the first group and positive for the second group) cancel out to some extent, resulting in a net change even lower than that within each group.

Simulation results

Results of simulations performed to examine the performance of our proposed estimators and species composition pattern in the mixed sample are reported under various species abundance models in Appendix S2. Simulated data were generated from several species abundance distributions for original and transformed habitats. Two structures for shared species across the two assemblages were considered: (A) the shared species include only abundant species; (B) the shared species include some abundant and some rare species.

The simulation results reveal that, for the mixture of two rarefaction curves, the proposed analytic estimators of the composite diversity ${}^{q}D(m_1,m_2)$ are nearly unbiased for all diversity orders. For the mixture of one rarefaction curve and one extrapolation curve, the proposed analytic estimators are satisfactory and are only slightly negatively biased. The analytical findings regarding the rate of change in species richness, described in the preceding section, are also supported by our numerical results in Appendix S2. As predicted, in each of the simulation scenarios, the estimated number of shared species in any mixed sample differs very little from the number of observed shared, species in the two reference samples. The R code "miNEXT" (mixed iNterpolation/EXTrapolation) for computing all estimators and plotting all curves discussed in this paper is available at Github (https://github.com/AnneChao).

An example for abundance data

The data used to illustrate the abundance-based approach were sampled in a mountain forest ecosystem in the Bavarian Forest National Park, Germany (Thorn *et al.* 2017). Here, a total of 12 experimental plots were established in *closed-forest* stands (6 plots) and *open-forest*

stands with naturally occurring gaps and edges (6 plots) to assess the effects of microclimate on communities of epigeal (ground-dwelling) spiders (details in Thorn *et al.* 2016). In this example, open forest, with naturally occurring gaps and edges, is considered an *original habitat*. Closed forest represents a *transformed habitat*—previously open forest transformed by artificially increasing canopy densities through planting of trees in gaps and patch edges. Species abundance data appear in Appendix S3.

Epigeal spiders were sampled over three years with four pitfall traps in each plot, yielding a total from the two habitats of 3171 individuals belonging to 85 species. In the open forest, there were 1760 individuals representing 74 species, whereas in the closed forest, there were 1411 individuals representing 44 species. Thirty-three species occurred in both open and closed forest plots. Thus, there were 41 unique species in open forest and 11 unique species in closed forest (33+41+11=85).

Figure 2 shows the proportional mixture of rarefaction/extrapolation curves, with corresponding 95% confidence bands, for the original (open) and the transformed (closed) forests. Note that when the hypothetical transformation of the open forest is 80%, 1760 x 80%=1408 individuals in the open forest will be replaced by individuals from closed forest. Because the reference-sample size in the closed forest is 1411, the composite diversity is thus a mixture of two rarefaction curves only up to a transformation proportion of 80%. Once the proportion of transformation in the open forest exceeds 80%, the composite diversity is a mixture of a rarefaction curve of the open forest and an extrapolation curve of the closed forest.

Figure 2 shows that, for q=0 (Panel a), species richness (red solid and dotted lines) for any mixed sample of size 1760 is less than the observed richness found in the open forest. That is, the hypothetical transformation of open forests to closed forests always resulted in a loss of species. This finding reflects the relatively large loss of unique spider species adapted

to open forest.

If we focus on common and dominant species (q=1 in Panel b, and q=2 in Panel c) diversity increases as proportions (up to 50%) of open forest are transformed to closed forest. In this scenario, common/dominant spider species in open forests are still present, even if relatively low proportions of open forest are transformed to closed forest. Also, common/dominant spider species typically found exclusively in closed forest might already be present, even with low proportions of closed forest, and thus contribute to an overall increase in diversity with small proportions of transformed habitat. The increase in diversity of order q=1 and q=2, when the transformed proportion is less than 50%, also reflects the increase of evenness of species' relative abundances due to loss of some rare species unique to open forest, together with the colonization of abundant/dominant species unique to closed forest. For these data, the confidence band-width generally decreases with the diversity order q. However, for a fixed value of q, the band-width varies modestly with the proportion of transformation, primarily because the confidence intervals refer to sampling uncertainties for mixed samples with a constant size $m_1 + m_2 = 1760$, and the expected composite diversity varies slowly with the proportion of transformation.

Figure 3 depicts the species compositional information with 95% confidence bands in any mixed sample when the proportion of transformation ranges between 0% and 80%. With no transformation, there were 33 shared species, and 41 unique species in the open forest, based on a reference sample size of 1760. As predicted by our theory, the number of shared species and the corresponding confidence bands in the mixed forest remain essentially unchanged, despite the replacement of individuals. The pattern can be intuitively understood for these data (given in Appendix S3), because most species shared between the two types of habitats represent abundant, widespread species. However, both analytical and simulation results suggest that the same pattern persists, regardless of the abundances of the shared

species. As the transformation proportion is increased, the number of species unique to open forest is reduced from 41 towards 0 with decreasing sampling uncertainty, whereas some species unique to closed forest are added from 0 towards 11 with increasing sampling uncertainty. Figure 3 further shows that the rate of decline for the species unique to open forest is greater than the rate of colonization by species unique to closed forest, leading to a net decrease in species richness.

(Figure 2, Figure 3)

Replicated incidence data

The above framework for species abundance data can be slightly modified to handle replicated incidence data (species occurrence frequencies). Assume that T_1 sampling units (reference sample I) are randomly taken from Assemblage I, and T_2 sampling units (reference sample II) are taken from Assemblage II. In each sampling unit, only the detection or non-detection of each species is recorded. The two sets of incidence probabilities $(\pi_{11}, \pi_{21}, ..., \pi_{S1})$ and $(\pi_{12}, \pi_{22}, ..., \pi_{S2})$ for S species represent species detection probabilities in any sampling unit from Assemblages I and II, respectively, $\pi_{i1}, \pi_{i2} \geq 0,\ i = 1, 2, ..., S.$ Let Y_{i1} and Y_{i2} denote, respectively, the number of sampling units in which the ith species is detected in reference sample I and in reference-sample II. All our estimation procedures are based on the observed species incidence-based frequency sets $(Y_{11}, Y_{21}, ..., Y_{S1})$ and $(Y_{12}, Y_{22}, ..., Y_{S2})$.

We highlight the following differences from the estimation procedures for abundance data:

(a) The sample size for abundance data is replaced by the number of sampling units.

(b) The abundance vectors $(X_{11}, X_{21}, ..., X_{S1})$ and $(X_{12}, X_{22}, ..., X_{S2})$ are replaced by incidence-based frequency vectors, $(Y_{11}, Y_{21}, ..., Y_{S1})$ and $(Y_{12}, Y_{22}, ..., Y_{S2})$.

(c) The total number of individuals in abundance data is replaced by the total number of

incidences in multiple sampling units.

Under the above framework, all derivation details and estimation procedures are parallel to the abundance-based approach, to depict a mixture curve and to compute species composition in any mixed sample; see Appendix S4 for derivations and an empirical example.

CONCLUSION AND DISCUSSION

We generalized within-habitat rarefaction and extrapolation (Gotelli & Colwell 2001, Colwell *et al.* 2012, Chao *et al.* 2014) to a mixture of two rarefaction/extrapolation curves (Box 1, Figure 1). The proposed framework enables us to forecast landscape patterns of biodiversity if a known proportion of original habitat is transformed. We derived the theoretical formulas for the resulting diversity (Eqs. 1–4a) and the formulas for assessing species composition in the mixture of two habitats (Eqs. 5a–5c). We also quantified the rates of change for unique and shared species (Eqs. 6a–6c) in progressive transformations. The corresponding analytic estimators are provided in Appendix S1. The plot of the mixture curve and species composition are demonstrated in Figures 2 and 3, respectively, for ground-dwelling spider abundance data, and in Appendix S4 for song-bird incidence data. Our proposed framework is valid for the mixture of any two types of assemblages. When habitat transformation alters species abundances, and data from a partially transformed habitat are available, our model and formulas can also be applied to the mixture of the partially transformed and totally transformed assemblages. Then we can assess not only how habitat transformation alters species abundances, but also adaptively forecast landscape diversity.

Our framework and approach offer two major advances. First, previous analyses were typically based on the comparison of mean values of alpha diversity found in one habitat type compared to another habitat type (Cadotte *et al.*, 2012; Lindenmayer & Fischer 2006). Such categorical comparisons based on alpha-diversity alone do not consider beta diversity, i.e.,

they disregard the fact that assemblages found in distinct habitat patches may share a substantial fraction of species. Hence, net changes in species richness can mask changes in community composition caused by species losses and replacements (Dornelas *et al.*, 2014; Hillebrand *et al*. 2018). To address this shortcoming, many authors have quantified, separately, the responses of specialist vs. generalist species (Chazdon *et al*. 2011; Clavel *et al.*, 2011), or species belonging to different functional groups (Bihn *et al.*, 2008). Here we explicitly consider shared and unique species in each habitat type, without the need for a pre-defined classification scheme. Second, our approach is applicable to a wide range of habitat transformation scenarios and will yield results that are comparable within and among different habitat types, since it is based on a unified framework of Hill numbers.

The importance of transformed habitats for biodiversity is a controversial topic (Dent & Wright 2009; Gibson *et al.* 2011; Chazdon 2014). Categorical study-designs do not support a rigorous derivation of evidence-based thresholds needed to preserve a certain amount of biodiversity in transformed habitats. Well-designed studies that investigate changes in biodiversity over a continuous gradient of habitat transformation are scarce (but see Barlow *et al.* 2016), and thresholds for the preservation of habitats are often derived from comparisons of mean alpha diversity measures from different habitat types (e.g. Burivalova *et al.* 2014), which neglect the frequent occurrence of shared species. Here, our approach offers conservationists, land-managers and policy-makers a second advance in estimating thresholds for habitat preservation when setting up laws, guidelines or management recommendations (Kareiva *et al.* 2014). By explicitly considering shared, unique, and undetected species (in an extrapolated sample), our approach provides more accurate estimates of non-linear changes in total biodiversity that typically occur with habitat mixtures (Figure 2).

The role of unique species in progressive transformations

Based on our analytical reasoning (Eqs. 6a–6c) and numerical studies (empirical examples

and simulations in Appendix S2), a consistent pattern has emerged: the number of shared species in any mixed sample rarely changes as transformation proceeds. That is, the overall increase or decrease in species richness with increasing transformation of the original habitat depends almost exclusively on the loss of unique species from the original habitat and the addition of unique species from the transformed habitat. Our analytical and numerical findings suggest that a successful strategy for maintaining a specified level of overall species richness is equivalent to maintaining a minimum number of species unique to each habitat.

Our analytical formulas show how abundant vs. rare unique species affect rates of change in species richness arising from habitat transformations. For any fixed pair of hypothetical sample sizes (m_1, m_2), Eqs. (6a) and (6b) reveal that unique species from the original habitat with abundances $\hat{p}_{i1} \approx 1/m_1$ contribute the largest absolute decline rates in species richness, whereas unique species from the transformed habitat with abundances $\hat{p}_{i2} \approx 1/(m_2+1)$ contribute the greatest rate of increase. Thus, in the initial stage of transformation (i.e., m_1 close to n_1 and m_2 close to 0), *rare* unique species from the original habitat and *abundant* unique species from the transformed habitat are the most influential. As the transformation proceeds, less-rare unique species from the original habitat and less-abundant unique species from the transformed habitat then gradually become the most influential; all these results conform to intuitive reasoning.

Generalization to a phylogenetic version

In this paper, we focus mainly on species diversity (Hill numbers) in which all species are considered equally distinct from one another. That is, only species richness and abundance are considered; species relatedness, trait differences, or contributions to ecosystem function are not incorporated in our mixture formulation. A rapidly growing literature addresses phylogenetic diversity metrics. When all species in an assemblage are connected by a rooted

phylogenetic tree, with all species as tip nodes, Chao *et al*. (2010) extended Hill numbers to incorporate phylogeny, so that evolutionary information among species can be accounted for. Hsieh & Chao (2017) further generalized the rarefaction/extrapolation of Hill numbers to phylogenetic diversity. Although it is beyond the scope of this paper, our framework can be generalized to a phylogenetic version to forecast change in phylogenetic diversity when habitats are progressively transformed.

ACKNOWLEDGEMENTS

The authors would like to thank the reviewers and in particular the Editors (John Drake and Stephan Munch) for providing very helpful and insightful comments and suggestions. This work is supported by the Taiwan Ministry of Science and Technology under Contract 106-2628-M-007-01 (for AC). ST was supported by a MOST (Ministry of Science and Technology) Taiwan Research Fellowship and received funds from the Gregor Louisoder Environmental Foundation.

REFERENCES

Barlow, J., Lennox, G.D., Ferreira, J., Berenguer, E., Lees, A.C., MacNally, R., *et al*. (2016). Anthropogenic disturbance in tropical forests can double biodiversity loss from deforestation. *Nature*, 535, 144–147. doi:10.1038/nature18326

Betts, M.G., Wolf, C., Ripple, W.J., Phalan, B., Millers, K.A., Duarte, *et al*. (2017). Global forest loss disproportionately erodes biodiversity in intact landscapes. *Nature*, 547, 441–444. doi:10.1038/nature23285

Bihn, J.H., Verhaagh, M., Brandle, M., Brandl, R., & Bra, M. (2008). Do secondary forests act as refuges for old growth forest animals? Recovery of ant diversity in the Atlantic forest of Brazil. *Biol. Conserv*., 141, 733–743. doi:10.1016/j.biocon.2007.12.028

Burivalova, Z., Şekercioğlu, Ç.H., & Koh, L.P., (2014). Thresholds of logging intensity to maintain tropical forest biodiversity. *Curr. Biol.*, 24, 1893–1898. doi:10.1016/j.cub.2014.06.065

Cadotte, M.W., Mehrkens, L.R., & Menge, D.N.L. (2012). Gauging the impact of meta-analysis on ecology. *Evol. Ecol.*, 26, 1153–1167. doi:10.1007/s10682-012-9585-z

Chao, A., Chiu C.-H. & Jost, L. (2010). Phylogenetic diversity measures based on Hill numbers. *Philos. Trans. R. Soc. Lond. B. Biol. Sci*., 365, 3599–3609.

Chao, A., Gotelli, N.J., Hsieh, T.C., Sander, E.L., Ma, K.H., Colwell, R.K. *et al*. (2014). Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecol. Monogr.*, 84, 45–67.

Chao, A., Hsieh, T.C., Chazdon, R.L., Colwell, R.K. & Gotelli, N.J. (2015). Unveiling the species-rank abundance distribution by generalizing the Good-Turing sample coverage theory. *Ecology*, 96:1189–1201.

Chazdon, R. L. (2014). Second growth: the promise of tropical forest regeneration in an age of deforestation. University of Chicago Press, Chicago, USA.

Chazdon, R.L., Chao, A., Colwell, R.K., Lin, S.-Y., Norden, N., Letcher, S.G., *et al*. (2011). A novel statistical method for classifying habitat generalists and specialists. *Ecology*, 92, 1332–1343.

Clavel, J., Julliard, R. & Devictor, V. (2011). Worldwide decline of specialist species: Toward a global functional homogenization? *Front. Ecol. Environ.*, 9, 222–228. doi:10.1890/080216

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

Dent, D.H., & Wright, J.S. (2009). The future of tropical species in secondary forests: A

quantitative review. *Biol. Conserv.*, 142, 2833–2843. doi:10.1016/j.biocon.2009.05.035

Dornelas, M., Gotelli, N.J., McGill, B., Shimadzu, H., Moyes, F., Sievers, *et al*. (2014). Assemblage time series reveal biodiversity change but not systematic loss. *Science*, 344, 296–299. doi:10.1126/science.1248484

Fahrig, L. (1997). Relative effects of habitat loss and fragmentation on population extinction. *J. Wildl. Manag.*, 61, 603–610.

Foley, J.A., Defries, R., Asner, G.P., Barford, C., Bonan, G., Carpenter, S.R. *et al.* (2005). Global consequences of land use. *Science*, 309, 570–574. doi:10.1126/science.1111772

Gaston, K.J., 2000. Global patterns in biodiversity. *Nature*, 405, 220–227. doi:10.1038/35012228

Gibson, L., Lee, T.M., Koh, L.P., Brook, B.W., Gardner, T.A, Barlow, J., *et al.* (2011). Primary forests are irreplaceable for sustaining tropical biodiversity. *Nature*, 478, 378–381. doi:10.1038/nature10425

Gotelli, N.J., & Colwell, R.K. (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.*, 4, 379–391.

Hawkins, B. A., Field, R., Cornell, H. V., Currie, D. J., Guégan, J.-F., Kaufman, D. M., *et al*. (2003). Energy, water, and broad-scale geographic patterns of species richness. *Ecology*, 84, 3105–3117.

Hillebrand, H., Blasius, B., Borer, E. T., Chase, J. M., Downing, J. A., Eriksson, B. K., *et al.* (2018). Biodiversity change is uncoupled from species richness trends: Consequences for conservation and monitoring. *J. Appl. Ecol.*, 55, 169–184.

Hsieh, T.C. & Chao, A. (2017). Rarefaction and extrapolation: making fair comparison of abundance-sensitive phylogenetic diversity among multiple assemblages. *Syst. Biol.*, 66, 100–111.

Kareiva, P., Groves, C., & Marvier, M. (2014). The evolving linkage between conservation

science and practice at the nature conservancy. *J. Appl. Ecol.*, 51, 1137–1147. doi:10.1111/1365-2664.12259

Kulakowski, D., Seidl, R., Holeksa, J., Kuuluvainen, T., Nagel, T.A., Panayotov, M., *et al*. (2016). A walk on the wild side: Disturbance dynamics and the conservation and management of European mountain forest ecosystems. *For. Ecol. Manage.*, 388, 120–131. doi:10.1016/j.foreco.2016.07.037

Lindenmayer, D. B. & Fischer, J. (2006). Habitat fragmentation and landscape change: an ecological and conservation synthesis. Island Press, Washington, DC.

McGill, B.J., Dornelas, M., Gotelli, N.J., Magurran, A.E., (2015). Fifteen forms of biodiversity trend in the Anthropocene. *Trends Ecol. Evol*., 30, 104–113. doi:10.1016/j.tree.2014.11.006

Newbold, T., Hudson, L.N., Hill, S.L.L., Contu, S., Lysenko, I., Senior, R.A., *et al.* (2015). Global effects of land use on local terrestrial biodiversity. *Nature*, 520, 45–50. doi:10.1038/nature14324

Prevedello, J.A., Gotelli, N.J., & Metzger, J.P., (2016). A stochastic model for landscape patterns of biodiversity. *Ecol. Monogr.*, 86, 462–479. doi:10.1002/ecm.1223

Thorn, S., Bässler, C., Svoboda, M., & Müller, J. (2017). Effects of natural disturbances and salvage logging on biodiversity - Lessons from the Bohemian Forest. *For. Ecol. Manege.*, 388, 113–119. doi:10.1016/j.comnet.2006.11.031

Thorn, S., Bußler, H., Fritze, M.-A., Goeder, P., Müller, J., Weiß, I., *et al*. (2016). Canopy closure determines arthropod assemblages in microhabitats created by windstorms and salvage logging. *For. Ecol. Manage*., 381, 188–195. doi:10.1016/j.foreco.2016.09.029

Table 1. Four typical cases when a specified hypothetical proportion of an original habitat is transformed; the reference sample sizes for the original and transformed habitats are respectively 200 and 160. The sample sizes for the two hypothetical samples are m_1 and m_2, $m_1+m_2=200$.

Case	Proportion of original habitat in mixture	Proportion of transformed habitat in mixture	Hypothetical sample size from the original habitat (m_1)	Hypothetical sample size from the transformed habitat (m_2)	Composite diversity in the mixture
A	100%	0%	200	0	Observed diversity in the reference sample of the original habitat
B	90%	10%	180	20	Mixture of two rarefaction curves
C	10%	90%	20	180	Mixture of one rarefaction curve and one extrapolation curve
D	0%	100%	0	200	Extrapolated value for a size of 200 for the transformed-habitat reference sample

Figure Legends

Figure 1. Proportional mixture of two rarefaction/extrapolation curves for an original habitat (black) and a transformed habitat (blue). The rarefaction curve for the original habitat (black solid line) and the integrated rarefaction/extrapolation curve for the transformed habitat (blue solid line for the rarefaction part and blue dashed line for the extrapolated part) are plotted for a given sample size in the X-axis (i.e., number of individuals, with proportion in parenthesis). The closed dots denote the two reference samples and the open dot denotes the extrapolated value for a sample size of 200 individuals for the transformed habitat, matching the sample size from the original habitat. Any point in the red solid line (a mixture of two rarefaction curves) or the red dashed line (a mixture of a rarefaction curve and an extrapolation curve) represents the composite diversity when a hypothetical sample with size m_1 individuals (using the same X-axis as the two individual rarefaction/extrapolation curves) and a hypothetical sample with size m_2 (m_1+m_2=200) are respectively taken from the original and transformed habitats. The 95% confidence intervals (shaded bands) are obtained by a bootstrap method based on 100 replications. The lower horizontal grey dotted line represents the level of the extrapolated diversity of sample size 200 in the transformed habitat, and the upper horizontal grey line represents the level of the diversity of the reference sample of size 200 in the original habitat. The four letters, A, B, C and D correspond to the four cases in Box 1 and Table 1. Case A corresponds to the diversity of the reference sample of size 200 in the original habitat, Case B corresponds to the diversity from a mixture of two rarefaction curves with a combined size of 200, Case C corresponds to the diversity from a mixture of a rarefaction curve and an extrapolation curve with a combined size of 200, and Case D corresponds to the extrapolated diversity value of a hypothetical size of 200 for the transformed habitat (i.e., same level as the blue open circle).

Figure 2. Mixed rarefaction/extrapolation curves of spiders sampled in original open forest and transformed closed forest stands (Thorn *et al*. 2016). Mixed rarefaction/extrapolation is calculated for (a) q=0 (species richness), (b) q=1 (the effective number of common species), and (c) q=2 (the effective number of dominant species). The 95% confidence intervals (shaded bands) are obtained by a bootstrap method based on 100 replications. See Figure 1 legend for more details.

Figure 3. Number of species found in both habitat types (i.e. shared, purple dotted curve), unique-to-original species (black dash-dotted curve), and unique-to-transformed species (blue dash-double-dotted curve) in a mixed sample when the open forest (original habitat) is progressively transformed to a closed forest. The 95% confidence intervals (shaded bands) are obtained by a bootstrap method based on 100 bootstrap replications. The black solid curve denotes the rarefaction curve for the original habitat, and the red curve denotes the mixture curve; see Figure 2 (a).