A novel statistical method for classifying habitat generalists and specialists

Robin L. Chazdon,[1,2,9] Anne Chao,[3] Robert K. Colwell,[1,2] Shang-Yi Lin,[3] Natalia Norden,[1,4] Susan G. Letcher,[5] David B. Clark,[6] Bryan Finegan,[7] and J. Pablo Arroyo[8]

[1]*Department of Ecology and Evolutionary Biology, University of Connecticut, 75 North Eagleville Road, Storrs, Connecticut 06269 USA*
[2]*Center for Macroecology, Evolution and Climate, Department of Biology, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen, Denmark*
[3]*Institute of Statistics, National Tsing Hua University, Hsin-Chu, Taiwan 30043*
[4]*Departamento de Ciencias Biológicas, Universidad de los Andes, Cra 1 #18A-10, Bogotá, Colombia*
[5]*Organization for Tropical Studies, Apartado Postal 676-2050, San Pedro de Montes de Oca, Costa Rica*
[6]*Department of Biology, University of Missouri–St. Louis, One University Boulevard, St. Louis, Missouri 63121-4400 USA*
[7]*Production and Conservation in Forests Program, Tropical Agricultural Centre for Research and Higher Education (CATIE), Apartado 93-7170, Turrialba, Costa Rica*
[8]*Department of Geography, McGill University, 805 Sherbrook Street West, Montreal, Quebec H3A 2K6 Canada*

Abstract. We develop a novel statistical approach for classifying generalists and specialists in two distinct habitats. Using a multinomial model based on estimated species relative abundance in two habitats, our method minimizes bias due to differences in sampling intensities between two habitat types as well as bias due to insufficient sampling within each habitat. The method permits a robust statistical classification of habitat specialists and generalists, without excluding rare species a priori. Based on a user-defined specialization threshold, the model classifies species into one of four groups: (1) generalist; (2) habitat A specialist; (3) habitat B specialist; and (4) too rare to classify with confidence. We illustrate our multinomial classification method using two contrasting data sets: (1) bird abundance in woodland and heath habitats in southeastern Australia and (2) tree abundance in second-growth (SG) and old-growth (OG) rain forests in the Caribbean lowlands of northeastern Costa Rica. We evaluate the multinomial model in detail for the tree data set. Our results for birds were highly concordant with a previous nonstatistical classification, but our method classified a higher fraction (57.7%) of bird species with statistical confidence. Based on a conservative specialization threshold and adjustment for multiple comparisons, 64.4% of tree species in the full sample were too rare to classify with confidence. Among the species classified, OG specialists constituted the largest class (40.6%), followed by generalist tree species (36.7%) and SG specialists (22.7%). The multinomial model was more sensitive than indicator value analysis or abundance-based phi coefficient indices in detecting habitat specialists and also detects generalists statistically. Classification of specialists and generalists based on rarefied subsamples was highly consistent with classification based on the full sample, even for sampling percentages as low as 20%. Major advantages of the new method are (1) its ability to distinguish habitat generalists (species with no significant habitat affinity) from species that are simply too rare to classify and (2) applicability to a single representative sample or a single pooled set of representative samples from each of two habitat types. The method as currently developed can be applied to no more than two habitats at a time.

Key words: diagnostic species; fidelity measures; habitat preference; habitat specificity; indicator species; indicator value; multinomial model; species classification; species distribution; succession.

Introduction

The habitat specificity of species has long been a major focus of ecological research (MacArthur and Levins 1964, Ashton 1967, Levins 1968, Rosenzweig 1981). Ecologists often classify species as generalists or specialists, based on the strength of species' affinities for particular habitats. These affinities can be determined by field observations

(Clark et al. 1999, Baker et al. 2002, Phillips et al. 2003) or by experimentation (Abramsky et al. 1990, Fine et al. 2004, Baltzer et al. 2005), but should be validated based on species abundance data derived from samples collected in different habitats. Aside from its central importance in testing and developing ecological theory, information regarding habitat specificity has many practical applications, including identifying species of concern for conservation, selecting species for restoration or reforestation projects, and identifying unique suites of traits that link species to particular habitat types (Mayfield et al. 2009).

Many statistical approaches have been used to examine species–habitat relationships, including frequency distributions of presence records, indicator value (IV) analysis, incidence- and abundance-based phi coefficient of association, ordination, Mantel correlations, multiple regression of distance matrices, and occupancy analysis (Dufrêne and Legendre 1997, Svenning 1999, Harms et al. 2001, Chytrý et al. 2002, Phillips et al. 2003, Tuomisto et al. 2003, Tichý and Chytrý 2006, De Cáceres and Legendre 2009). Although ordination techniques are commonly used to classify species–site relationships as a community-based approach, ordination does not permit a statistical classification of individual species as habitat generalists or specialists and usually demands that rare species be eliminated from data sets prior to analysis (Clarke and Warwick 2001). The SIMPER analysis, implemented in PRIMER (Clarke and Warwick 2001) quantifies the contribution of individual species to similarity between two assemblages based on the Bray-Curtis similarity between the samples, but this measure requires equal sampling fractions in both habitats and is inappropriate for many field data sets (Chao et al. 2006). Several studies have demonstrated species associations with particular habitats or topographic classes using null models that predict abundances based on random distributions across habitat types, while accounting for effects of spatial autocorrelation (Harms et al. 2001, Plotkin et al. 2002, Comita et al. 2007, Itoh et al. 2010). But these methods only allow common species to be tested for significant departure from null expectations. Moreover, existing methods that measure the strength of species associations with particular habitats or assemblages do not distinguish habitat generalists from species that are simply too rare or infrequent to assess.

Here we report on a novel statistical approach to classifying habitat generalists and specialists using a multinomial model based on the estimated relative abundance of species in two distinguishable habitats. Our model permits a robust statistical evaluation of habitat specialization for large numbers of species and does not rely on measurements of individual performance or exclude rare species a priori. As the foundation for this new approach, our objective was to develop a two-habitat species classification model that minimizes bias due to difference in sampling intensities between the two habitats as well as bias due to insufficient sampling of rare species in each habitat. These problems frequently arise in biodiversity surveys and inventories, particularly in the species-rich tropics where most species in assemblages are rare and species richness is often incompletely sampled (Colwell and Coddington 1994, Longino and Colwell 1997, Chao et al. 2005, Coddington et al. 2009). The two-habitat approach is well suited to comparative studies in which a larger number of habitat groups within particular study areas can be pooled into two major categories that are shared across all areas (e.g., high elevation/low elevation,

aquatic/terrestrial, canopy/understory, protected area/matrix, sandy soil/clay soil, forest/savanna, etc.). It is best suited for large habitat samples or for pooled replicate plots within each of two habitat types.

For species shared between two habitat types, our multinomial model distinguishes four groups: (1) habitat A specialists, (2) habitat B specialists, (3) generalists, and (4) species that are too rare to classify as either specialists or generalists. These categories would apply equally well to any comparison of the distribution of species between two ecologically meaningful categories, including not only habitat types, but also such contrasts as the prey species in the diets of two predators, the "catch" of two insect trap types, or the species that characterize two fossil-bearing strata. The method also has potential in evolutionary biology to assess contrasts between populations or species on the basis of alleles or other markers that either characterize one group or the other ("specialist" alleles) or that are confidently shared by both ("generalist" alleles).

We illustrate our method with two contrasting data sets. First, we classify 78 species of birds surveyed in adjacent woodland and heath habitats in southeastern Australia, based on data (*Ecological Archives* E083-058-A1) from a previously published regional study of the heath/woodland ecotone (Baker et al. 2002). The original study used quantitative but nonstatistical criteria to classify the habitat specialization of these birds based on species abundance, using a method that assumes equal sampling success in the two habitats.

Second, we demonstrate our approach and evaluate it in depth using a data set for tropical rain forest trees in northeastern Costa Rica. For such a highly diverse group in which rare species predominate, limited data from a single study area may be insufficient to evaluate variation in species abundance across habitat types. To test our method for this data set, we therefore used a replicated, landscape-scale approach to obtain sufficiently large samples, combining data from several studies conducted concurrently in the region. We used data on the abundance of trees ≥10 cm diameter, sampled in each of two forest types: 11.3 ha of tropical second-growth forests (SG) of various ages and land use history (7–45 yr old) and 18.3 ha of old-growth forests (OG), with no recorded history of recent major human disturbance. Second-growth forests in this region have regenerated following land clearance and burning for agriculture during the past 50 years (Letcher and Chazdon 2009). Because our statistical method compares a single, large sample (or a set of pooled samples), combining data from multiple study plots intentionally minimizes effects of within-habitat spatial heterogeneity on analysis of species habitat affinities, a strategy also applied in the Australian bird study (Baker et al. 2002).

To illustrate the advantages and limitations of the multinomial model approach, we compare our classification results for the tree data set with indicator value (IV) analysis (Dufrêne and Legendre 1997) and with

equalized and nonequalized, abundance-based phi coefficients of association (De Cáceres and Legendre 2009). A species indicator value is based on the product of mean species relative abundance within replicated samples of a habitat type and relative frequency of occurrence within those replicated samples, whereas the phi coefficient measures the degree of association between species presence/absence (or abundance counts) and a binary habitat type (De Cáceres and Legendre 2009). We also evaluate the sensitivity of the multinomial model classification to sample size, by comparing species classification results for six rarefied subsamples composing 20%, 40%, 60%, 80%, 90%, and 95% of the trees in the full sample. Across these six sampling proportions, we compare the consistency of species classifications as well as the percentages of species in the four classification groups.

METHODS

Study areas and data sets

Baker et al. (2002) surveyed bird abundance eight times at three-month intervals in Budderoo National Park, Booderee National Park, and Nadgee Nature Reserve coastal regions of southeastern Australia. Surveys were conducted in 11 sets of paired woodland and heath plots (50 × 400 m), parallel to, but 100 m distant from the sharp, linear ecotone between *Eucalyptus*-dominated woodland and treeless heath/sedgeland. These distinct heath–woodland edges divided relatively large, homogeneous areas of vegetation extending more than 200 m on each side (Baker et al. 2002). Using abundance data (total number of detections), Baker et al. (2002) classified 18 species as woodland specialists, three species as heath specialists, and 10 species as habitat generalists; an additional 55 species were sampled, but were eliminated from analysis due to low number of detections. By their classification criterion, a specialist was any species with >75% of detections in one of the two habitat types. No correction was made for unequal total detections, for all species pooled, in the two habitats. Although Baker et al. (2002) included data from ecotone (habitat transition) plots in their classification, our demonstration analysis is based only on the subset of their data from the pure woodland and pure heath plots.

The study landscape in the Sarapiquí region of northeastern Costa Rica covered an area of ~48 km², of which ~48% was forested in 2001 (Sesnie et al. 2008). The landscape consists of old-growth and second-growth fragments in a complex matrix of cattle pastures and plantations of banana, palm heart, pineapple, and tree plantations. Several reserves are located within the study area, including La Selva Biological Station and adjacent areas of Braulio Carrillo National Park, Selva Verde Lodge, Tirimbina Research Center, Finca La Martita, and Finca El Bejuco Reserve (see map in Appendix A). Reserves comprise ~28% of the area, and second-growth forests are common, mostly regenerating on abandoned pastures (Guariguata et al. 1997,

Chazdon et al. 2007, Schedlbauer and Kavanagh 2008, Letcher and Chazdon 2009).

We analyzed inventory data for trees ≥10 cm from 34 secondary forest sites (0.1–1 ha) scattered across the landscape and 11 plots in four distinct blocks of old-growth forest at elevations from 30 to 150 m above sea level. Appendix A provides detailed information on the inventory data sets. Our study areas are representative of second-growth and old-growth forests and encompass the range of edaphic and topographic variation observed within the area. For the purpose of the multinomial species classification evaluated here, all secondary forest sites were combined into a single, pooled sample and all old-growth sites were combined into a second pooled sample, together encompassing 13 689 trees (Appendix A: Table A1). These two "full samples" (as we will call them) combine the features of intensive and extensive sampling of vegetation, a robust assessment of regional assemblage composition for the analysis of tree species affinities in second- and old-growth forest. For comparison with methods that require replicated sampling (indicator value analysis and the abundance-based phi coefficients of association) we used individual data sets as replicates. Our complete data set and species list is found in Appendix B.

Multinomial model formulation

For clarity, the model will be presented in terms of the old-growth (OG) and second-growth (SG) forest habitats studied in the Costa Rican tree data set, but the model applies equally to the Australian bird data set or any other two-habitat comparison for which appropriate data are available. Suppose that there are S species in the combined area OG and SG forests and these S species are indexed by 1, 2, ..., S. Denote the true (unknown) relative abundances for the OG forest as (p_1, p_2, ..., p_S), where $\Sigma_{i=1}^{S} p_i = 1$, and the true (unknown) relative abundances for the SG forest as (π_1, π_2, ..., π_S), where $\Sigma_{i=1}^{S} \pi_i = 1$. In the sample data, let the number of stems (individuals) of the ith species in the OG forest be denoted by X_i and the corresponding number of stems of this species in the SG forest be denoted by Y_i. Thus our data include two sets of absolute abundances: (X_1, X_2, ..., X_S) for the OG forest and (Y_1, Y_2, ..., Y_S) for the SG forest. Assume the sample size in OG forest is n and the sample size in SG forest is m, i.e., $\Sigma_{i=1}^{S} X_i = n$ and $\Sigma_{i=1}^{S} Y_i = m$. Our statistical model is a multinomial model for each of the two sets of absolute abundances (X_1, X_2, ..., X_S) and (Y_1, Y_2, ..., Y_S).

Observed values of relative abundance in most studies, especially for diverse tropical biotas, are often based on incomplete samples. Our model allows that some species may be present in one or both forest habitats, even if they were not detected in one or both pooled samples. We assume that observed relative abundance in the samples reflects the true relative abundances (i.e., the data are representative of the assemblage from which they were recorded) but is

subject to random sampling error. Based on sample data, we first obtain an adequate estimator for the true relative species abundances p_i and π_i.

For species i, let $a_i = X_i/n$ be the sample relative abundance of this species in the OG forest and $b_i = Y_i/m$ be the corresponding sample relative abundance in the SG forest. A traditional approach is to use the sample relative abundance as an estimate of true relative abundance. However, the sample relative abundance works correctly for individual species only if all species present have been observed in the sample. This approach implies that any undetected species has zero relative abundance, which is not a reasonable conclusion in highly diverse assemblages with many rare species. The sample frequency underestimates the probability of observing an individual of any undetected species, while simultaneously overestimating the probabilities for the observed species, such that, overall, the expectation of the sample frequency is an unbiased estimator. Because the level of overestimation for each detected species scales inversely with abundance, however, estimates for rare, detected species are more severely biased than estimates for common, detected species (Appendix C).

A simple modification to correct for this bias for rare, detected species is based on the concept of sample coverage, attributed to Alan Turing by Good (1953, 2000). The sample coverage for a sample is defined as the proportion of the true total abundances (including undetected species) represented by the pooled abundances of the species detected by the sample. A well-known estimator (first proposed by Turing; see Good 1953) for the sample coverage is one minus the proportion of singletons in the sample (Appendix C), where a singleton is a species represented by exactly one individual in the sample. Turing's theory implies that a proper estimate for p_i is $\tilde{p}_i = C_1(X_i/n) = C_1 a_i$, where $C_1 = 1 - f_1/n$ denotes the sample coverage for OG data and f_1 denotes the number of singletons in the OG data. Similarly, a proper estimate for π_i is $\tilde{\pi}_i = C_2(Y_i/m) = C_2 b_i$, where $C_2 = 1 - g_1/m$ denotes the sample coverage for SG data and g_1 denotes the number of singletons in the SG data (Ashbridge and Goudie 2000, Chao and Shen 2003). Under this modification, the fraction of the total abundances of the undetected species in OG data and in SG data is, respectively, $1 - C_1$ and $1 - C_2$, rather than 0 as in the traditional approach.

For common species, the sample relative abundances X_i/n and Y_i/n based on observed frequencies accurately reflect the true relative abundances, and statistical adjustment is not needed (Appendix C). Thus, in practice, we adopt a mixed approach, applying the sample coverage correction to rare species (X_i or $Y_i < 10$ individuals), while using sample relative abundances for common species (X_i or $Y_i \geq 10$ individuals) (Appendix C).

Species classification algorithm

After we obtain the estimated relative abundances $\tilde{p}_i$ and $\tilde{\pi}_i$ in each type of forest (Appendix D), we compute the ratio $\tilde{p}_i/(\tilde{p}_i + \tilde{\pi}_i)$ for each species i as an estimate of the unknown parametric ratio $p_i/(p_i + \pi_i)$ and compare this estimate to a specialization threshold value. (We could just as well compute its one-complement as an estimate of $\pi_i/(p_i + \pi_i)$.) Observed species can be either shared by OG and SG samples or unshared (present in one sample but not the other). In our example, shared species can be classified into four possible groups: OG specialist, SG specialist, generalist, or too rare to classify.

Classification of shared species

If a shared species is an OG specialist, then we would expect that p_i is sufficiently higher than π_i, or equivalently, the ratio $p_i/(p_i + \pi_i)$ is sufficiently higher than 1/2. But how high is sufficiently high? We consider two different specialization thresholds to bracket our analysis. A conservative threshold, based on the concept of "supermajority" rule, uses a cut-off point $K = 2/3$, i.e., $p_i/(p_i + \pi_i) > 2/3$ (equivalently, $p_i > 2\pi_i$). Many decision-making bodies use a two-thirds majority to determine their actions, such as in the election of the Roman Catholic Pope or in parliamentary procedures of particular consequence. A liberal threshold uses a "simple majority" rule, with a cut-off point $K = 1/2$, i.e., $p_i/(p_i + \pi_i) > 1/2$ (equivalently, $p_i > \pi_i$).

Our procedure then performs one-sided statistical tests to classify species at a specified significance level, P. Both K and P are inputs to the multinomial model. Suppose that $\tilde{p}_i > 2\tilde{\pi}_i$, suggesting that species i is an OG specialist; we need to assess this classification statistically. Using the supermajority rule $K = 2/3$ to determine whether a species i belongs to the category OG specialist, we carry out a statistical test (test 1; see Appendix D) at a specified significance level P to assess:

$$H_0: p_i \leq 2\pi_i \text{ vs. } H_1: p_i > 2\pi_i. \qquad (1)$$

The choice of P depends on the user's objective. If the goal is to classify particular focal species, P could be 0.05 (or some other specified level). In contrast, if the objective is assemblage-wide classification, multiple comparisons are required and thus a much smaller P should be used, in order to yield a corresponding overall (experiment-wise) P. If test 1 is significant, then the species will be classified as an OG specialist. If the test is not significant, then the species will be classified as a generalist or as too rare to classify. Likewise, for a species for which $\tilde{\pi}_i > 2\tilde{p}_i$, to determine whether a species i belongs to the category SG specialist, we test:

$$H_0: \pi_i \leq 2p_i \text{ vs. } H_1: \pi_i > 2p_i. \qquad (2)$$

If the test is significant, then the species will be classified as an SG specialist. If test 2 is not significant, then the species will be classified as a generalist or as too rare to classify.

Based on the previous idea for classifying a species as an OG or SG specialist, for any fixed value of Y (the raw count of individuals of some particular species in SG) we
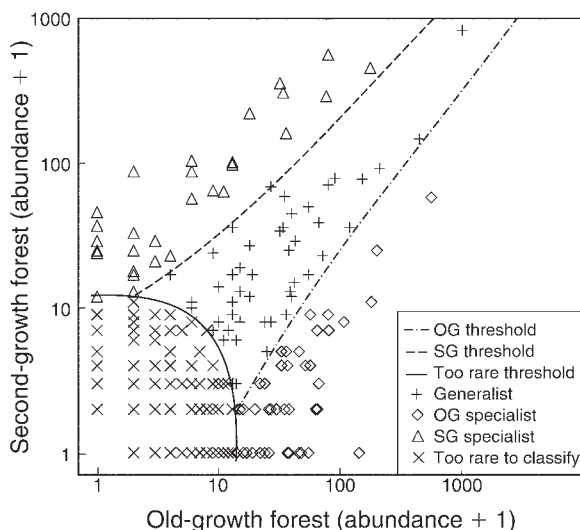
FIG. 1. Classification results for species in the full sample for the rain forest tree data set, using the super-majority specialization threshold ($K = 2/3$, $P = 0.005$), with adjustment for multiple comparisons. Key to abbreviations: SG, second growth; OG, old growth. Trees ≥ 10 cm diameter at breast height were sampled in 11.3 ha of second-growth and 18.3 ha of old-growth forests within a 48-km^2 area of northeastern Costa Rica.

can determine for given sample sizes the minimum value of X (the raw count of individuals of the same species in OG) such that the test H_0: $p_i \leq 2\pi_i$ vs. H_1: $p_i > 2\pi_i$ is significant at a particular value of K and P, assuming this species is indexed by i in the data list. For example, for the full sample in the tree data set (Fig. 1), if $Y = 1$, we require that $X \geq 13$ to assure that $p_i > 2\pi_i$ is significant so that $(X, 1)$ is declared an OG specialist (at $P = 0.005$ and $K = 2/3$). This special minimum value of X for $Y = 1$ is called X_{min}. For these data, $X_{min} = 13$. If $Y = 2$, we require that $X \geq 18$ to assure that $p_i > 2\pi_i$ is significant so that $(X, 2)$ is classified as an OG specialist. If $Y = 3$, we require that $X \geq 22$ to assure that $p_i > 2\pi_i$ is significant so that $(X, 3)$ indicates an OG specialist. Continuing this process, for any value of Y, we can determine such a minimum value of X. Thus connecting the points $(13, 1)$, $(18, 2)$, $(22, 3)$, ..., we have a boundary line between generalists and OG specialists (the dot-dashed line in Fig. 1).

In the same way, for any fixed value of X, we can determine for the full sample the minimum value of Y such that the test H_0: $\pi_i \leq 2p_i$ vs. H_1: $\pi_i > 2p_i$ is significant. For example, for the full sample in the tree data set, for $X = 1$, we require that $Y \geq 11$ to assure that $\pi_i > 2p_i$ is significant so that $(1, Y)$ is classified as an SG specialist. This special minimum value Y for $X = 1$ is called Y_{min}. For these data, $Y_{min} = 11$. For $X = 2$, we require that $Y \geq 14$ to assure that $\pi_i > 2p_i$ is significant (at $P = 0.005$) so that $(2, Y)$ is declared to be an SG specialist. For $X = 3$, we require that $Y \geq 17$ to assure

that $\pi_i > 2p_i$ is significant so that $(3, Y)$ belongs to an SG specialist. The line connecting points $(1, 11)$, $(2, 14)$, $(3, 17)$ ... represents the boundary between generalists and SG specialists (the dashed line in Fig. 1).

The tests used for determining the two boundary lines (Fig. 1) are the tests we need for our classification scheme. Based on our experience, 10–20 points are sufficient to plot each line. Thus a total of 20–40 tests are used in our approach (even if there are hundreds of species). Thus, for an overall (experiment-wise) $P = 0.05$, the significance level for each individual test should be controlled to be in the range of $0.05/20 = 0.0025$ and $0.05/40 = 0.00125$. For an overall $P = 0.10$, the individual level P should be controlled to be in the range of $0.10/20 = 0.005$ to $0.10/40 = 0.0025$. We recommend using $P = 0.005$ or 0.001 for experiment-wise hypotheses.

Classifying unshared species

Unshared species ($X = 0$ and $Y > 0$, or $Y = 0$ and $X > 0$) can be classified as OG specialist, SG specialist, or too rare to classify. In either case ($X = 0$ or $Y = 0$), we follow the rules for $Y = 1$ and $X = 1$ to do the classification. This approach can be statistically justified because, if the frequency $(X, 1)$ is significant for detecting $p_i > 2\pi_i$, then the frequency $(X, 0)$ must be at least as significant. That is, the threshold on the X-axis for significantly OG specialist species is X_{min}. Any unshared OG species with fewer than X_{min} individuals is too rare to classify as OG specialist (or for that matter, as a generalist), whereas an unshared OG species with abundance $X \geq X_{min}$ is declared an OG specialist, on the basis of these samples (Fig. 1). Likewise, the threshold on the Y-axis for significant SG specialist species is Y_{min}. Any unshared SG species fewer than Y_{min} individuals is too rare to classify as an SG specialist (or as a generalist), whereas an unshared SG species with abundance $Y \geq Y_{min}$ is declared an SG specialist, on the basis of these samples.

Identifying shared species that are too rare to classify

Both unshared species and shared species ($X > 0$, $Y > 0$) can be too rare to classify with confidence. Species are too rare to classify if X and Y are both relatively small or if a linear combination of X and Y is relatively small. We adopt the latter linear bound as a separation between "generalist" and "too rare." To identify these species, which lie near the origin of the (X, Y) plane, we draw a line segment connecting $(X_{min}, 0)$ to $(X_{min}, 1)$, another connecting $(0, Y_{min})$ to $(1, Y_{min})$, and a third connecting $(X_{min}, 1)$ to $(1, Y_{min})$ (the solid line in Fig. 1). This separation line can be justified as the line farthest from the origin such that all species with (X, Y) values lying between the origin and this boundary line are not significant for both test 1 and for test 2; in contrast, beyond this line, there exist (X, Y) values that are significant for test 1 or test 2. For the full samples in the tree data set (Fig. 1), we have $X_{min} = 13$ and $Y_{min} = 11$, so the four points defining the boundary are $(13, 0)$,

(13, 1), (0, 11), and (1, 11). In the (X, Y) plane, this boundary line includes three straight line segments. However, in the $(\log X, \log Y)$ plane in Fig. 1, the line connecting $(X_{min}, 1)$ and $(1, Y_{min})$ becomes a curve.

Evaluating the multinomial classification results for the tree data set

Using the procedures described above, we initially classified each of the 359 species in the full samples into one of the four classes: (1) successional generalists, (2) SG specialists, (3) OG specialists, and (4) too rare to classify (Fig. 1). We ran the multinomial model using $K = 2/3$ and $K = 1/2$ to bracket conservative and liberal classification rules, setting $P = 0.05$ and 0.005 to demonstrate both the individual species and experiment-wise approaches. The classification was based on an iterative program (CLAM) that implements all classification procedures (see Appendix G).

For the tree data set, we compared multinomial classifications to four other approaches: (1) a nonstatistical approach based on species frequency of occurrence (presence/absence) in two forest types; (2) indicator value (IV) analysis (Dufrêne and Legendre 1997) based on species frequency of occurrence in replicated samples; (3) nonequalized abundance-based phi coefficient index (i.e., the point-biserial correlation coefficient r_{pb} in De Cáceres and Legendre 2009); and (4) group equalized abundance-based phi coefficient index (i.e., the point-biserial correlation coefficient r_{pb}^{g} in De Cáceres and Legendre 2009). Significance was based on randomization tests. De Cáceres and Legendre (2009) considered two types of data (presence/absence and abundance-based) and two indices (phi coefficient index and indicator index), yielding four approaches, for each of which they also considered an "equalized" and a "nonequalized" version: a total of eight measures. We wrote R codes to perform classifications and permutation tests for indicator value analysis for these eight measures, based on a significance level of $P = 0.05$ for each species, using six sites for OG and eight sites for SG (Letcher's SG sites are pooled as a single site; Appendix A). The results do not differ much among these eight measures, so we present results for only two here: the abundance-based, nonequalized and the group equalized phi coefficient indices, which are the two measures particularly highlighted by De Cáceres and Legendre (2009). None of these other methods identifies generalists, so we could not compare this aspect of our classification method.

To assess the sensitivity of the multinomial classification model to sample size, we repeated the classification analyses on each of six rarefied subsamples for the forest data set, composed of 20%, 40%, 60%, 80%, 90%, and 95% of stems in the full samples in each forest type. Random subsamples were drawn without replacement from the full sample, performing 1000 runs for each rarefaction level, maintaining the proportion of stems in OG (0.567) and SG (0.433) found in the full sample. We examined the consistency of the multinomial classification model across different sampling fractions by quantifying the average percentage of 1000 trials in which the classification results remained perfectly consistent with classification results based on the full samples. We also compared the percentage of species in each of the four classes for each rarefaction level to the corresponding percentage within the full samples. Because even the full samples contain a large percentage of singletons (19.4% of species in OG and 30.7% of species in SG), all of which were expected to fall into the "too rare to classify" category, we expected that as the sampling fraction is reduced, the proportion of species in this category would increase and the proportions in the generalist and specialist categories would decrease.

RESULTS

Bird species classification based on the multinomial model

A total of 2482 birds (69 species) were observed in woodland plots and 1295 birds (40 species) were observed in heath plots (Baker et al. 2002); 31 of 78 (39.7%) species were shared between the two habitat types. We used a super-majority specialization threshold ($K = 2/3$) and evaluated classifications at $P = 0.05$ (appropriate for classification of individual species) and $P = 0.005$ (suitable for assessing overall pattern). Thirty-three species (42.3%) were too rare to classify using $P = 0.05$, whereas 36 species (46.2%) were too rare to classify using $P = 0.005$, a difference of only three species (Appendix E). Among the 47 bird species that Baker et al. (2002) excluded due to a low number of detections, our method classified six as woodland specialists, five as heath specialists, and three as generalists ($P = 0.05$; Appendix E), even though we used a subset of their data that excluded samples within the ecotone itself. Using $P = 0.005$, five of the excluded species were classified as woodland specialists and four as heath specialists, a difference of only one species in each category.

Our results were highly concordant with the nonstatistical classification by Baker et al. (2002). The three heath specialists identified by Baker et al. (2002) were consistently classified as heath specialists using both P levels (Appendix E). Among the 18 woodland specialists identified by Baker et al. (2002), we classified 15 as woodland specialists and three as generalists ($P = 0.05$). Among the 10 habitat generalists, seven were classified as generalists and three as heath specialists ($P = 0.05$ and $P = 0.005$; Appendix E). Compared to the 21 habitat specialists identified by Baker et al. (2002), our model classified 32 and 28 as specialists for $P = 0.05$ and 0.005, respectively. Most of these additional specialists emerged from the pool of species excluded by Baker et al. (2002) as too rare to classify (Appendix E).

The percentage of species with only one or two individuals (singletons and doubletons) was 21.7 for woodland data and 25.0 for the heath data. For the
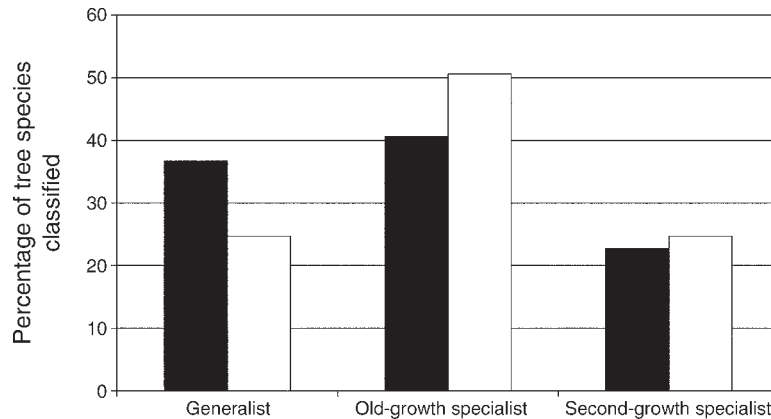
F_IG. 2. Classification results for alternative specialization thresholds for 359 tree species in the full sample, based on the percentage of species that could be significantly classified ($P = 0.005$). Results are shown as black bars using the super-majority threshold ($K = 2/3$) and as white bars using the simple-majority threshold ($K = 1/2$).

estimation of relative abundance, 44.9% of the species in woodland data and 50.0% in heath data had fewer than 10 individuals and thus required the application of the Turing-Good estimator.

Tree species classifications based on the multinomial model

The pooled data set for trees contained a total of 359 species; 299 species were observed in OG, whereas 238 species were observed in SG. Half of the species (49.6%) were shared by both forest types. Over one-third of the species were found only in OG (33.7%), whereas 16.7% of species were found only in SG. The percentage of species with only one or two individuals (singletons and doubletons) recorded in the full samples was higher for SG (47.1%) than for OG (27.1%). For the estimation of relative abundance, 71.8% of the species in SG and 61.5% in OG had fewer than 10 individuals and thus required the application of the Turing-Good estimator.

The species classification for $K = 2/3$ (super-majority threshold) and $P = 0.005$ (adjusted for experiment-wise tests) revealed that 231 species (64.4%) were too rare to classify. Among the 128 species sufficiently abundant to classify, OG specialists were the largest class (52 species, 40.6%), followed by successional generalists (47 species, 36.7%) and SG specialists (29 species, 22.7%; Fig. 2).

Setting, instead, a simple-majority threshold ($K = 1/2$) reduced the number of species too rare to classify from 231 to 205 (from 64.4% to 57.1%). Among the 154 species sufficiently abundant to classify, OG specialists were again the largest class (78 species, 50.6%), followed by generalists (38 species, 24.7%) and SG specialists (38 species, 24.7%; Fig. 2). Using the simple-majority threshold, abundant shared species tend to be classified as specialists, whereas these species are more likely to be classified as generalists using the higher threshold (Appendix F). Comparing the results for these alternative thresholds demonstrates a clear trade-off between increased sensitivity of the model to specialists and classification of a higher fraction of the total species present in samples.

Indicator species and phi coefficient of association analyses

To compare our multinomial model with the IV analysis and phi coefficient methods on the same footing, we set $P = 0.05$ for all methods, because nearly all IV and phi coefficient tests would be nonsignificant for our data for $P = 0.005$ (Appendix F). The multinomial method was more sensitive in classifying species as OG or SG than either IV or abundance-based phi coefficient analyses (Fig. 3). Of the 72 OG specialists classified by the multinomial method ($K = 2/3$), only 40 were significant by IV, 42 by nonequalized phi, and 41 by equalized phi. Likewise, of the 38 SG specialists classified by the multinomial method ($K = 2/3$), only seven were significant by IV and 10 by either of the phi-coefficient analyses.

Based on the IV analysis, 45 tree species were significant OG indicator species and seven were significant SG indicators (Appendix F). Among the 45 OG indicator species, our multinomial method classified 40 as OG specialists, two as generalists, and three as too rare to classify based on the conservative threshold ($K = 2/3$; Appendix F). All seven of the SG indicator species were classified as SG specialists by the multinomial method. The nonequalized and equalized abundance-based phi coefficient of association both pinpointed the same 10 SG indicator species. All of these species were classified as SG specialists using either specialization threshold (K) of the multinomial method (Appendix F). A total of 59 species were identified as OG indicator species based on equalized or nonequalized abundance-based phi correlations; 46 of these were OG indicator species in both phi-coefficient analyses (Appendix F). Of these 46 species, 40 were classified as OG specialists using the conservative threshold of the multinomial method ($K = 2/3$) and all 46 were classified as OG
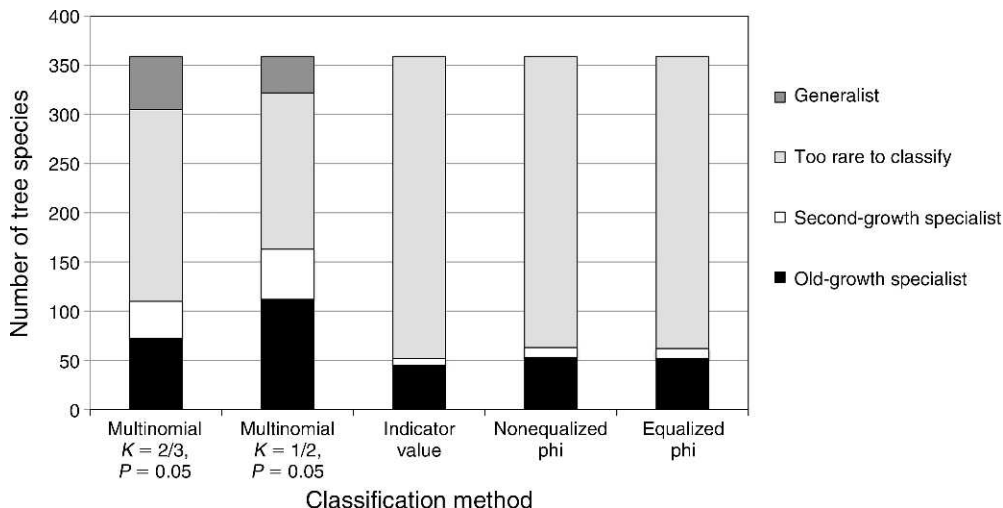
Fig. 3. Species classifications of 359 Costa Rican tree species based on the multinomial method using conservative ($K = 2/3$) and liberal ($K = 1/2$) specialization thresholds, indicator value (IV) analysis, nonequalized abundance-based phi coefficient, and equalized abundance-based phi coefficient. $P = 0.05$ for all methods.

specialists using the liberal threshold ($K = 1/2$; Appendix F).

Consistency of species classification in rarefied sampling fractions

Random sampling of fractional subsets from the full samples had little effect on species classifications for habitat specialists. For OG and SG specialists, the classification results for the rarefied subsamples were consistent with those based on the full sample, even for sampling fractions as small as 20% (Fig. 4). In the 20% rarefied sample (1554 stems from OG and 1184 stems from SG), on average, 98.7% of the species classified as SG specialists in the rarefied samples remained classified as SG specialists in the full samples and an average of 96.9% of the species classified as OG specialists in the rarefied samples remained classified as OG specialists in the full samples (Fig. 4). For species classified as generalists in the rarefied subsamples, the results are slightly less robust than for specialists, but species classifications remain highly consistent, especially for rarefied fractions of 40% or higher (Fig. 4). In the 20% rarefied fraction, 83.2% of the species classified as generalists in the rarefied samples maintained their classification as generalists (Fig. 4).

Distribution of classification groups in rarefied sample fractions

Not surprisingly, a higher fraction of total species can be confidently classified as generalists or specialists in larger samples. As expected, the percentage of species too rare to classify increased as the equal-sampling fraction declined from 95% to 20%, while the percentage of species classified as generalists and specialists decreased (Fig. 5). At 20% of the full sample size, an average of 82.4% of species were too rare to classify,

compared with only 64.4% in the full samples. The OG specialists formed the next largest class of species in all simulations, except in 20% and 40% rarified samples, where generalists surpassed the OG specialists. The OG specialists declined from 14.5% in full samples to only 4.6% in the 20% rarified samples (Fig. 5). The percentage of SG specialist declined from 8.1% in the full samples to only 5.4% in the 20% rarified samples. Equalizing the total abundance of stems sampled in OG and SG had no effect on the classification of specialists
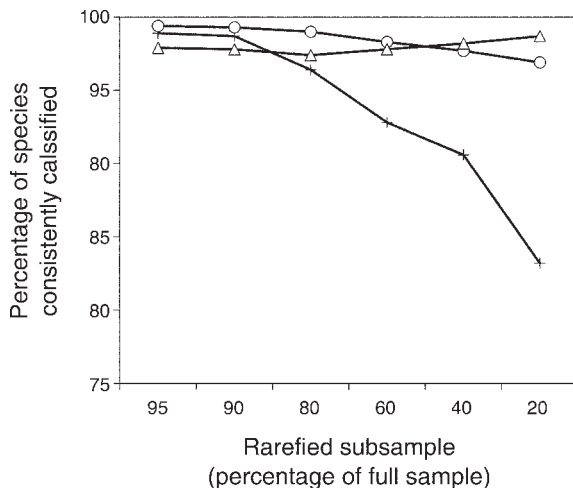

Fig. 4. The mean percentage (out of 1000 trials) of tree species that maintained the same classification of specialists and generalists as in the full sample, for six discrete levels of rarefaction. Rarefied subsamples maintained the same proportion of total tree species abundances in old growth (OG) and second growth (SG) in northeastern Costa Rica. Key to symbols: crosses, generalists; circles, OG specialists; and triangles, SG specialists.
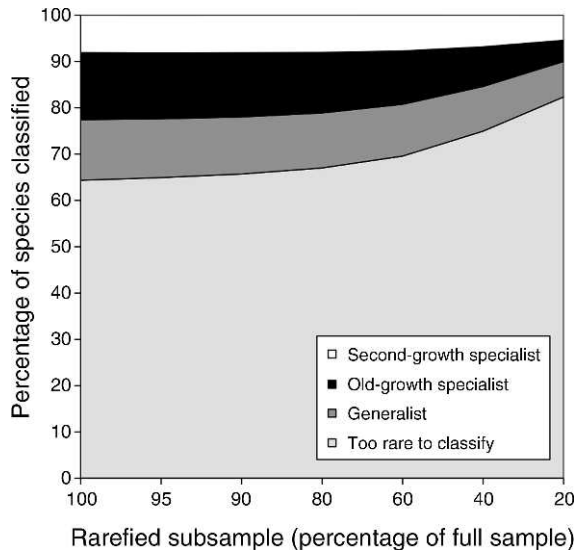
Fig. 5. The mean percentage of Costa Rican tree species (out of 1000 trials) in each of four classification groups for six discrete levels of rarefaction.

and generalists in the full sample or in any of the rarefied sample fractions (data not shown).

DISCUSSION

As illustrated by two quite different data sets, our multinomial method permits the statistically robust classification of a substantial proportion of species, based on inventory data for entire assemblages collected in representative plots in two distinct habitats. This new approach has several advantages over the application of existing methods for the binary classification of habitat specificity. Applying our method to the bird data of Baker et al. (2002) increased the number of species classified as habitat specialists compared with the quantitative but nonstatistical methods used in the published study (Appendix E). Further, the original nonstatistical classifications were strongly concordant with the multinomial model classifications for species categorized by both approaches (Baker et al. 2002). For the tree data set, the multinomial model detected more habitat specialists than either indicator value analysis or abundance-based phi correlation analyses (Fig. 3, Appendix F).

These examples demonstrate four additional advantages of the multinomial classification method using combined samples from habitat types within a region. First, species that are generalists can be classified and distinguished from species that are too rare to classify. Second, by combining a representative set of samples, the analysis avoids the complicating effects of within-habitat heterogeneity (beta diversity) on species classifications. If a species is missing or rare in one OG site but abundant in others, it can still be classified as an OG specialist. This pooling effect may well be why the method classified more habitat specialists than either

indicator value or abundance-based phi correlation methods. Thus, the method is less sensitive to the spatial distribution of species across sites, which often may reflect dispersal limitation or local historical effects, rather than true habitat specificity. Third, all species are analyzed, regardless of their abundance, and the model itself identifies those species that are too rare to classify. In our tree data set, all of the 101 tree species that were found in only one of the replicate study areas (Appendix A) were too rare to classify at $P = 0.005$ (Appendix A). Thus, to our knowledge, our method did not inappropriately classify any local site specialists as habitat specialists. Fourth, ours is the only statistically rigorous method for the binary classification of the habitat specificity that does not require replicated samples for each habitat.

The model permits flexibility in applying liberal or conservative specialization thresholds and significance levels. Our analysis illustrates the expected trade-off between "oversensitivity" of the model to specialists and classifying a higher fraction of the total species present in samples. Using the simple-majority threshold results in classification of fewer generalists and more specialists. This effect becomes more pronounced with small sample size. To compensate for this effect, we suggest using a super-majority rule (2/3 or perhaps an even stricter criterion) when using small samples. Two different classification rules can also be used to "bracket" species classification results, or for a particular species, the specialization threshold at which the species shifts from being classified as a specialist to being a generalist could be found as the threshold proportion for the test is increased.

Our simulations using rarefied fractions of the full samples confirmed that the classification of specialists and generalists based on small sampling fractions is highly consistent with classifications based on the full samples for our data (Fig. 4). As expected, when smaller and smaller random samples are drawn from the full samples for OG and SG, an increasing number of species become too rare to classify, and (for the tree data) this effect is most pronounced for OG specialists and generalists (Fig. 5). This effect of "undersampling" will always be stronger when the full sample has few dominant species and many relatively rare species, as is clearly the case with tropical forest tree communities (Ashton et al. 1990, Pitman et al. 1999).

Classification methods based on species abundance data assume that abundance is a proxy measure for species adaptedness or fitness within the particular habitats examined. But species abundance can be limited locally in a suitable habitat due to dispersal limitation and therefore does not always reflect the suitability of a particular habitat for a species. Although relatively less important for the bird example, dispersal limitation is particularly common for large-seeded, animal-dispersed tree species in early successional habitats. Seedlings of many of the trees species classified here as OG specialists

are present in the understory of SG forests (Norden et al. 2009, Chazdon et al. 2010). Thus, tree species classification as OG specialists is attributed to the later colonization in SG rather than to the inability of these tree species to grow and survive in SG.

Our model does not explicitly account for spatial autocorrelation, but rather relies on study design to avoid the pitfalls of confounding habitat differences with unmeasured spatial variables. We emphasize the importance of sampling species abundance in well-dispersed and representative study sites. Baker et al. (2002) strategically placed replicate plots over a wide area of southeast Australia. Tree data from SG and OG sites were interspersed across a broad landscape including a wide range of edaphic variation, thereby reducing the potential effect of unmeasured environmental variables, but the abundance of many tropical tree species is known to be patchy, even at the landscape scale (He et al. 1997, Clark et al. 1998, Condit et al. 2000, Plotkin et al. 2000). Spatial aggregation can lead to the appearance of specialization to a particular forest type or condition, but may arise from some other underlying effect, such as dispersal limitation, recruitment subsidies (Franklin and Rey 2007, Zuidema et al. 2010), or unique microhabitats. We would expect spatial autocorrelation to be a more serious problem classifying species within a restricted area (a 50-ha plot, for example; Condit et al. 2000) than from widely scattered sites across a landscape or region. We recommend testing for spatial autocorrelation of species abundance patterns across sampling plots prior to the application of the multinomial classification method. If significant spatial autocorrelation is found, species abundance data should be replaced with frequency data (presence or absence of species in spatially gridded subsamples), thereby removing strong effects of spatial clumping on the classification results.

Using our model as currently developed, specialization can be assessed for no more than two habitats. Characterizing the scope of a continuum by means of a dichotomy is a longstanding practice in ecology (consider, for example, r- vs. K-selected species, "small" vs. "large" and "near" vs. "far" islands in the theory of island biogeography, or temperate vs. tropical climates). Indeed, binary classifications in science in general are ubiquitous and often more useful than more complex conceptualizations. In statistics, two-sample models and tests are widely applied to contrasts of all kinds. Nevertheless, unlike IV analysis, the current model cannot be used to compare species' relative abundance patterns across a larger number of habitats. The model could be used to compare species distributions across more than two habitats by classifying species in different combinations of habitat pairs, but this approach still considers only two habitats within each classification and interpretation could quickly become unmanageable. We aim to develop an expanded model for more than two habitats in the future. Finally, in our model, as with any discrete classification model, biological variables are usually continuous and gradients of species characteristics and responses are not taken explicitly into account, although they could be part of the design.

Ultimately, classification methods are limited by the nature and extent of the underlying data. If sampling is sparse and incomplete, all but a few highly abundant species will be too rare to classify. As applied to successional forests, classification results will also be highly sensitive to the successional stage of forests sampled, as the tree species in old-growth forests are slow to colonize and recruit to 10-cm size classes in old fields and young second-growth forests (Letcher and Chazdon 2009, Chazdon et al. 2010).

Species rarity presents challenges for any model of classification. About 40% of the bird species and more than half of the tree species in the data sets we analyzed here proved too rare to classify (Figs. 1 and 5; Appendices E and F for details). For these species, it is not possible to determine habitat affinities with statistical confidence based on relative abundance data. But rare species are often those of greatest conservation concern (for example, the Eastern Bristlebird, *Dasyornis brachypterus*, in the Baker et al. [2002] study), and most tree species are rare in tropical forests. In upper Amazonian forests, for example, 88% of tree species occur at densities of less than 1 tree/ha (Pitman et al. 1999). Of the 231 species too rare to classify in the full samples, 54 occurred only in SG forest, 101 occurred only in OG forest, and 76 were shared. In these cases, incidence data (as opposed to relative abundance) may be the only criterion of habitat affinity, and determination of habitat specialization will require more detailed, species-level information.

Conclusions

Classifying species into generalist and specialist groups is the first step toward examining the underlying biological and ecological factors leading to the differential distribution of species between habitats. Results of classification models have broad significance for testing ecological theory, for planning and communicating conservation or reintroduction programs, and for assessing effects of climate change and succession on species distributions. We have demonstrated that our multinomial model holds much promise for classifying species habitat affinities at a landscape scale. The utility and broad applicability of our new abundance-based model should be tested with a wider variety of data and taxa.

Literature Cited

Abramsky, Z., M. Rosenzweig, B. Pinshow, J. Brown, B. Kotler, and W. Mitchell. 1990. Habitat selection: an experimental field test with two gerbil species. Ecology 71:2358–2369.

Ashbridge, J., and I. B. J. Goudie. 2000. Coverage-adjusted estimators for mark–recapture in heterogeneous populations. Communications in Statistics: Simulation 29:1215–1237.

Ashton, P. S. 1967. Climate versus soil in the classification of south-east Asian tropical lowland vegetation. Journal of Ecology 55:67–68.

Ashton, P., L. Holm-Nielsen, I. Nielsen, and H. Balslev. 1990. Tropical forests: botanical dynamics, speciation and diversity. Academic Press, London, UK.

Baker, J., K. French, and R. J. Whelan. 2002. The edge effect and ecotonal species: bird communities across a natural edge in southeastern Australia. Ecology 83:3048–3059.

Baltzer, J. L., S. C. Thomas, R. Nilus, and D. F. R. P. Burslem. 2005. Edaphic specialization in tropical trees: physiological correlates and responses to reciprocal transplantation. Ecology 86:3063–3077.

Chao, A., R. L. Chazdon, R. K. Colwell, and T.-J. Shen. 2005. A new statistical approach for assessing compositional similarity based on incidence and abundance data. Ecology Letters 8:148–159.

Chao, A., R. L. Chazdon, R. K. Colwell, and T.-J. Shen. 2006. Abundance-based similarity indices and their estimation when there are unseen species in samples. Biometrics 62:361–371.

Chao, A., and T.-J. Shen. 2003. Nonparametric estimation of Shannon's index of diversity when there are unseen species in samples. Environmental and Ecological Statistics 10:429–443.

Chazdon, R. L., B. Finegan, R. S. Capers, B. Salgado-Negret, and F. Casanoves. 2010. Composition and dynamics of functional groups of trees during tropical forest succession. Biotropica 42:31–40.

Chazdon, R. L., S. G. Letcher, M. van Breugel, M. Martinez-Ramos, F. Bongers, and B. Finegan. 2007. Rates of change in tree communities of secondary Neotropical forests following major disturbances. Philosophical Transactions of the Royal Society B 362:273–289.

Chytrý, M., L. Tichý, J. Holt, and Z. Botta-Dukát. 2002. Determination of diagnostic species with statistical fidelity measures. Journal of Vegetation Science 13:79–90.

Clark, D. B., D. A. Clark, and J. M. Read. 1998. Edaphic variation and the mesoscale distribution of tree species in a neotropical rain forest. Journal of Ecology 86:101–112.

Clark, D. B., M. W. Palmer, and D. A. Clark. 1999. Edaphic factors and the landscape-scale distributions of tropical rain forest trees. Ecology 80:2662–2675.

Clarke, K. R., and R. M. Warwick. 2001. Change in marine communities: an approach to statistical analysis and inter-pretation. Second edition. Primer-E, Plymouth Marine Laboratory, Plymouth, UK.

Coddington, J., I. Agnarsson, J. Miller, M. Kuntner, and G. Hormiga. 2009. Undersampling bias: the null hypothesis for singleton species in tropical arthropod surveys. Journal of Animal Ecology 78:573–584.

Colwell, R. K., and J. A. Coddington. 1994. Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society B 345:101–118.

Comita, L. S., R. Condit, and S. P. Hubbell. 2007. Developmental changes in habitat associations of tropical trees. Journal of Ecology 95:482–492.

Condit, R., P. Ashton, P. Baker, S. Bunyavejchewin, S. Gunatilleke, N. Gunatilleke, S. Hubbell, R. Foster, A. Itoh, and J. LaFrankie. 2000. Spatial patterns in the distribution of tropical tree species. Science 288:1414.

De Cáceres, M. D., and P. Legendre. 2009. Associations between species and groups of sites: indices and statistical inference. Ecology 90:3566–3574.

Dufrêne, M., and P. Legendre. 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecological Monographs 67:345–366.

Fine, P. V. A., I. Mesones, and P. D. Coley. 2004. Herbivores promote habitat specialization by trees in Amazonian forests. Science 305:663–665.

Franklin, J., and S. J. Rey. 2007. Spatial patterns of tropical forest trees in Western Polynesia suggest recruitment limitations during secondary succession. Journal of Tropical Ecology 23:1–12.

Good, I. J. 1953. The population frequencies of species and the estimation of population parameters. Biometrika 40:237–264.

Good, I. J. 2000. Turing's anticipation of empirical Bayes in connection with the cryptanalysis of the naval Enigma. Journal of Statistical Computation and Simulation 66:101–111.

Guariguata, M., R. Chazdon, J. Denslow, J. Dupuy, and L. Anderson. 1997. Structure and floristics of secondary and old-growth forest stands in lowland Costa Rica. Plant Ecology 132:107–120.

Harms, K. E., R. Condit, S. P. Hubbell, and R. B. Foster. 2001. Habitat associations of trees and shrubs in a 50-ha neotropical forest plot. Journal of Ecology 89:947–959.

He, F., P. Legendre, and J. V. LaFrankie. 1997. Distribution patterns of tree species in a Malaysian tropical rain forest. Journal of Vegetation Science 8:105–114.

Itoh, A., T. Ohkubo, S. Nanami, S. Tan, and T. Yamakura. 2010. Comparison of statistical tests for habitat associations in tropical forests: a case study of sympatric dipterocarp trees in a Bornean forest. Forest Ecology and Management 259:323–332.

Letcher, S. G., and R. L. Chazdon. 2009. Rapid recovery of biomass, species richness, and species composition in a forest chronosequence in northeastern Costa Rica. Biotropica 41:608–617.

Levins, R. 1968. Evolution in changing environments. Princeton University Press, Princeton, New Jersey, USA.

Longino, J. T., and R. K. Colwell. 1997. Biodiversity assessment using structured inventory: capturing the ant fauna of a lowland tropical rainforest. Ecological Applications 7:1263–1277.

MacArthur, R., and R. Levins. 1964. Competition, habitat selection, and character displacement in a patchy environment. Proceedings of the National Academy of Sciences 51:1207.

Mayfield, M. M., M. F. Boni, and D. D. Ackerly. 2009. Traits, habitats, and clades: identifying traits of potential importance to environmental filtering. American Naturalist 174:E1–E22.

Norden, N., R. L. Chazdon, A. Chao, Y.-H. Jiang, and B. Vilchez Alvarado. 2009. Resilience of tropical rain forests:

tree community reassembly in secondary forests. Ecology Letters 12:384–394.

Phillips, O. L., P. N. Vargas, A. L. Monteagudo, A. P. Cruz, M. E. C. Zans, W. G. Sanchez, M. Yli-Halla, and S. Rose. 2003. Habitat association among Amazonian tree species: a landscape-scale approach. Journal of Ecology 91:757–775.

Pitman, N. C. A., J. Terborgh, M. R. Silman, and P. Nuñez V. 1999. Tree species distributions in an Upper Amazonian Forest. Ecology 80:2651–2661.

Plotkin, J., J. Chave, and P. Ashton. 2002. Cluster analysis of spatial patterns in Malaysian tree species. American Naturalist 160:629–644.

Plotkin, J. B., M. D. Potts, N. Leslie, N. Manokaran, J. LaFrankie, and P. S. Ashton. 2000. Species–area curves, spatial aggregation, and habitat specialization in tropical forests. Journal of Theoretical Biology 207:81–99.

Rosenzweig, M. 1981. A theory of habitat selection. Ecology 62:327–335.

Schedlbauer, J. L., and K. L. Kavanagh. 2008. Soil carbon dynamics in a chronosequence of secondary forests in northeastern Costa Rica. Forest Ecology and Management 255:1326–1335.

Sesnie, S. E., P. E. Gessler, B. Finegan, and S. Thessler. 2008. Integrating Landsat TM and SRTM-DEM derived variables with decision trees for habitat classification and change detection in complex neotropical environments. Remote Sensing of the Environment 112:2145–2159.

Svenning, J.-C. 1999. Microhabitat specialization in a species-rich palm community in Amazonian Ecuador. Journal of Ecology 87:55–65.

Tichý, L., and M. Chytrý. 2006. Statistical determination of diagnostic species for site groups of unequal size. Journal of Vegetation Science 17:809–818.

Tuomisto, H., K. Ruokolainen, M. Aguilar, and A. Sarmiento. 2003. Floristic patterns along a 43-km long transect in an Amazonian rain forest. Journal of Ecology 91:743–756.

Zuidema, P., T. Yamada, H. During, A. Itoh, T. Yamakura, T. Ohkubo, M. Kanzaki, S. Tan, and P. Ashton. 2010. Recruitment subsidies support tree subpopulations in non-preferred tropical forest habitats. Journal of Ecology 98:636–644.

APPENDIX A

Description of vegetation inventory data sets used in this study (*Ecological Archives* E092-112-A1).

APPENDIX B

Tree species ≥10 cm dbh and their abundance in old-growth and second-growth study areas described in Appendix A (*Ecological Archives* E092-112-A2).

APPENDIX C

The effect of undetected species on estimated species frequencies (*Ecological Archives* E092-112-A3).

APPENDIX D

Testing for specialization under the multinomial model, using a super-majority threshold (*Ecological Archives* E092-112-A4).

APPENDIX E

Classifications of 78 bird species surveyed by Baker et al. (2002) in woodland and heath plots in southeastern Australia (*Ecological Archives* E092-112-A5).

APPENDIX F

Comparison of results of Indicator Species Analysis, equalized and non-equalized abundance-based phi correlation analysis and multinomial classification of 359 tree species using super-majority threshold and simple-majority threshold using $P = 0.05$ or $P = 0.005$ (*Ecological Archives* E092-112-A6).

APPENDIX G

User's Guide for Program CLAM, with instructions and a link for downloading the program (*Ecological Archives* E092-112-A7).