BIOTROPICA 44(1): 44–52 2012 10.1111/j.1744-7429.2011.00777.x

Subsampling Herbarium Collections to Assess Geographic Diversity Gradients: A Case Study with Endemic Orchidaceae and Rubiaceae in Cameroon

Vincent Droissart[1,2,3,9], **Olivier J. Hardy**[4], **Bonaventure Sonké**[4,5,7], **Farid Dahdouh-Guebas**[1,3,6], and **Tariq Stévart**[3,7,8]

[1] Laboratoire de Complexité et Dynamique des Systèmes Tropicaux, Université Libre de Bruxelles – ULB, 50 Av. F. Roosevelt, CP 169, 1050 Bruxelles, Belgium

[2] Institut de Recherche pour le Développement (IRD), Unité Mixte de Recherche AMAP (Botanique et Bioinformatique de l'Architecture des Plantes), Boulevard de la Lironde, TA A-51/PS2, F-34398 Montpellier Cedex 5, France

[3] Herbarium et Bibliothèque de Botanique africaine, Université Libre de Bruxelles – ULB, 50 Av. F. Roosevelt, CP 169, 1050 Bruxelles, Belgium

[4] Service Evolution Biologique et Ecologie, Université Libre de Bruxelles – ULB, 50 Av. F. Roosevelt, CP160/12, 1050 Bruxelles, Belgium

[5] Laboratoire de Botanique systématique et d'Ecologie, Département des Sciences Biologiques, Ecole Normale Supérieure, Université de Yaoundé I, B.P. 047, Yaoundé, Cameroon

[6] Laboratory of Plant Biology and Nature Management, Vrije Universiteit Brussel – VUB, Pleinlaan 2, B-1050 Brussels, Belgium

[7] Missouri Botanical Garden, Africa and Madagascar Department, PO Box 299, St. Louis, Missouri 63166-0299, U.S.A.

[8] National Botanic Garden of Belgium, Domein van Bouchout, Nieuwelaan 38, B-1860 Meise, Belgium

ABSTRACT

We compiled herbarium specimen data to provide an improved characterization of geographic patterns of diversity using indices of species diversity and floristic similarity based on rarefaction principles. A dataset of 3650 georeferenced plant specimens belonging to Orchidaceae and Rubiaceae endemic to Atlantic Central Africa was assembled to assess species composition per half-degree or one-degree grid cells. Local diversity was measured by the expected number of species (S_k) per grid cell found in subsamples of increasing size and compared with raw species richness (S_R). A nearly unbiased estimator of the effective number of species per grid cell was also used, allowing quantification of ratios of 'true diversity' between grid cells. Species turnover was measured using a presence/absence-based similarity index (Sørensen) and an abundance-based index that corrects for sampling bias (*NNESS*). Our results confirm that the coastal region of Cameroon is more diverse in endemic species than those more inland. The southern part of this coastal forest is, however, as diverse as the more intensively inventoried northern part, and should also be recognized as an important center of endemism. A strong congruence between Sørensen and *NNESS* similarity matrices lead to similar delimitations of floristic units. Hence, heterogeneous sampling seems to confer more bias when measuring patterns of local diversity using raw species richness than species turnover using Sørensen index. Overall, we argue that subsampling methods represent a useful way to assess diversity gradients using herbarium specimens while correcting for heterogeneous sampling effort.

Abstract in French is available in the online version of this article.

Key words: *BiodivR*; central Africa; diversity patterns; effective number of species; herbarium collections; rarefaction principles; similarity index; subsampling procedure.

BIODIVERSITY STUDIES CONDUCTED USING HERBARIUM SPECIMENS AS PRIMARY DATASETS must deal with many sources of potential bias. In particular, herbarium specimens often offer an unreliable representation of the distribution of diversity because sampling effort is usually very heterogeneous in space (Prendergast *et al.* 1993). Sampling biases in botanical surveys are primarily related to the quality and the quantity of recording which mainly depend on individual abilities of botanists, on sampling methods (*i.e.*, time span of recording and scale of the survey) or on the type of plants being recorded (Rich & Woodruff 1992). Several studies have demonstrated that easily accessible or environmentally attractive areas benefit from higher sampling effort (Freitag *et al.* 1998, Reddy & Davalos 2003, Hortal *et al.* 2008). Widely recognized hotspots and areas near research facilities also receive more attention (Dennis & Thomas 2000, Reddy & Davalos 2003, Moerman & Estabrook 2006, Hortal *et al.* 2008). This happens because botanists usually aim to observe as many species as possible in a cost-effective and efficient manner. Sampling effort heterogeneity is particularly problematic in species-rich ecosystems in which local sampling at a given geographic scale records only a small fraction of the species present, a situation that can be recognized when species accumulation curves are far from saturation.

Received 16 June 2010; revision accepted 18 January 2011.

[9]Corresponding author; e-mail: vincent.droissart@ird.fr

Few attempts have been made to correct biases due to heterogeneous sampling efforts of herbarium specimens (Delisle *et al.* 2003), despite the fact that evaluating sampling bias is paramount for the design of reliable conservation strategies (Reddy & Davalos 2003, Grand *et al.* 2007). Moreover, the extent of the bias in datasets is rarely known, and few cases are described in detail (Rich & Woodruff 1992, Prendergast *et al.* 1993). Gaps and biases in biodiversity datasets are often significant enough to compromise the accurate description of diversity gradients using raw information compiled from existing data bases (Prendergast *et al.* 1993, Hortal *et al.* 2007). Ideally, sampling effort should be uniform to ensure that variations detected in distribution and abundance patterns reflect reality (Williams *et al.* 2002). An assumption of uniform sampling is also required by most methods designed to measure similarity between sampling units (Chao *et al.* 2005).

Methods to correct sampling biases include the application of rarefaction principles, the use of species distribution modeling, and the use of richness estimators to characterize poorly sampled areas by extrapolation (Reddy & Davalos 2003). The use of a standard method, such as grid-based mapping, also minimizes potential bias due to difference in sampling effort and preference for particular sites (Petrík *et al.* 2010). Prendergast *et al.* (1993) proposed a method that uses the number of visits made in a grid square to correct species richness estimates. Data on visitor effort, however, are often not available because most data are recorded randomly or opportunistically (Freitag *et al.* 1998). Schulman *et al.* (2007) developed mechanistic corrections that explicitly include the heterogeneity of sampling effort when estimating likelihoods of species occurrences. The rarefaction approach involves subsampling a given dataset to assess diversity gradients under effectively uniform number of individuals (or herbarium specimens).

Besides the problem of sampling bias, a recent debate has occurred on how to quantify diversity when Jost (2007) proposed that 'true diversity' measures must be based on Hill numbers (Hill 1973). Hill numbers form a family of diversity measures that can be interpreted as 'effective number of species' and where a parameter, q, controls the weight given to common species vs. rare species. Interestingly, Hill numbers are transformations of classical diversity indices: the reciprocal of Simpson's concentration index for $q = 2$, the exponential of Shannon–Wiener index for $q = 1$, and it is the species richness for $q = 0$. A singular property of Hill numbers is that they conform to the replication principle, allowing a coherent partitioning of diversity into α, β, and γ components (Tuomisto 2010a). The properties of Hill numbers and their intuitive interpretation make them highly desirable to quantify and compare diversity in a most sound way. Hill numbers, however, suffer substantial bias under limited sample size, especially when $q < 2$ (Routledge 1980), limiting their application under heterogeneous sample sizes. Nevertheless, Nielsen *et al.* (2003) developed an estimator that can be applied to obtain the effective number of species, N_e, corresponding to Hill number for $q = 2$, and which is nearly unbiased when the sample size is at least equal to N_e.

Using a carefully compiled set of herbarium specimens from Cameroon of Orchidaceae and Rubiaceae endemic to the Atlantic Central African forests, we address the following question in the present paper: what are the consequences of heterogeneous sampling efforts on our perception of gradients of local diversity and on our ability to identify floristic units? Our analysis examines both local diversity and species turnover between spatially defined units. Sampling bias can result in an unreliable perception of biodiversity distribution if not corrected properly. To quantify how much effect heterogeneous collecting efforts might have in our particular case, we evaluate the bias due to variation in sampling intensity using subsampling procedures. First, we compared raw species richness with the expected number of species for standardized subsample sizes and with a nearly unbiased estimator of the effective number of species (Nielsen *et al.* 2003). Then, we compared a commonly used similarity index based on presence–absence data to a family of bias-corrected similarity indices. Our results highlight some biases in the currently accepted views of diversity gradients through Cameroonian rain forests, pointing out regions of substantial interest for their high diversity in endemic species.

METHODS

Study location and data compilation.—This study focuses on Cameroon where we examined the distribution and the diversity of Orchidaceae and Rubiaceae endemic to Atlantic Central Africa. This region covers the Lower Guinea area of endemism (White 1979) and the Gulf of Guinea islands, which exhibits the highest levels of biodiversity in tropical Africa (Myers *et al.* 2000, Kier *et al.* 2005).

We used a dataset compiled to delineate centers and areas of endemism on the basis of two large, complementary families, Orchidaceae, most of which are epiphytes and anemochore species, and Rubiaceae, most of which are shrubs and endozoochore species. Together they comprise 1159 taxa (441 Orchidaceae and 718 Rubiaceae: Govaerts *et al.* 2010a, b), made of 1070 species and 89 infra-specific taxa (subspecies or variety), that represent about 10–15 percent of the flora of Cameroon. Hereafter, for simplicity, we will used the term 'species' when referring to diversity indices computed at the level of these taxa, even if they comprise infra-specific ones. We recorded data from all herbarium specimens at BR, BRLU, K, P, SCA, WAG, YA (Holmgren & Holmgren 1998 onwards [continuously updated]), and included new collections made during recent fieldwork (mainly housed at BR, BRLU, K, and YA). Specimen identification was done in the framework of previous studies (Droissart *et al.* 2006) and supplemented as needed. Each herbarium specimen was checked for possible misidentification and all label data were recorded. We then checked the georeferencing assigning values *post facto* where required and excluding any specimen with imprecise locality information.

From the compiled data base, 3650 records with precise location (accurate to 10 km) were selected, corresponding to 751 specimens of Orchidaceae and 2899 of Rubiaceae. These data were incorporated into Arcview 3.3® (ESRI, Redland, U.S.A.),

superimposing two alternative grid sizes (0.5° × 0.5° and a 1° × 1°) on a map of Cameroon to record the presence of each species. The 1° × 1° grid resolution was used to allow maximum compatibility with previous studies focused on sub-Saharan tropical Africa (Kuper *et al.* 2004, 2006; Burgess *et al.* 2005) and the 0.5° × 0.5° resolution was used to explore the congruence between measured indices at a finer scale. For each grid cell we calculated raw species richness (S_R), total number of specimens collected, and number of different collectors who gathered material (considering only the first collector name of each specimen). Matrices of species presence–absence and abundance per grid cell were also extracted from the GIS. Hereafter we consider each grid cell as a sample, the sample size being the number of specimen recorded in each cell.

PATTERN ANALYSIS.—Local diversity within each grid cell was first estimated by the principle of rarefaction, which enables calculation of an unbiased diversity index, S_k, representing the expected number of species found in a subsample of k specimens. This index was calculated using the following analytical formula (Hurlbert 1971): for a given sample,

$$S_k = \sum_s \left(1 - \binom{N - x_s}{k} \Big/ \binom{N}{k}\right), \quad (1)$$

where N is the sample size and x_s is the abundance (number of specimens) of species s in the sample. In this formula, the subsample size k affects the importance given to rare species but it is also constrained by the sample size as S_k cannot be computed for $k > N$. Given the trade-off between the importance attributed to rare species and the number of grid cells for which S_k can be computed, we considered two subsample sizes: $S_{(k = 25)}$ and $S_{(k = 100)}$.

We also measured true diversity *sensu* Jost (2007) using an estimator of the effective number of species (Hill number of order $q = 2$) defined as following (Nielsen *et al.* 2003):

$$N_e = \frac{(N-1)^2}{3 - N + (N+1)(N-2)\sum_s (x_s/N)^2}. \quad (2)$$

Computing N_e requires a minimal sample size of three but as N_e is biased under very low sample size, it was computed only for grid cells with $N > 5$ and we noticed when $N_e < N$ because nonnegligible bias is expected. S_k and N_e were computed with the software *Biodiv*R 1.2 (Hardy 2010).

The species turnover between two grid cells, i and j, was estimated using two measures of similarity: the Sørensen index and the *NNESS* index. Sørensen similarity index C_{ij} (Sørensen 1948), regarded as one of the most effective for comparing presence/absence data between samples, was computed using *Primer6*© (PRIMER-E Ltd, Plymouth, U.K.), as follows:

$$C_{ij} = \frac{a}{[(a+b)+(a+c)]/2}, \quad (3)$$

where a is the number of species shared between both samples, b is the number of species only present in sample i, and c is the number of species only present in sample j.

The *NNESS* index, a variant of the *NESS* index (Grassle & Smith 1976) and a generalization of the Morisita–Horn similarity index, is based on species abundances and controls for sampling bias using the rarefaction principle. It is defined as

$$NNESSij/k = \frac{ESSij/k}{(ESSii/k + ESSjj/k)/2}, \quad (4)$$

where $ESS_{ij/k}$ is the expected number of species shared for random draws of k specimens from sample i and k specimens from sample j, which is estimated as

$$ESS_{ij/k} = \sum_S \left(1 - \binom{N_i - x_{is}}{k} \Big/ \binom{N_i}{k}\right)\left(1 - \binom{N_j - x_{js}}{k} \Big/ \binom{N_j}{k}\right), \quad (5)$$

where N_i is the sample size of i and x_{is} is the number of specimens of species s in sample i. Note that $NNESS_{ij/k}$ cannot be estimated if $N_i < k$ or $N_j < k$. There is an obvious analogy between the definitions of Sørensen and *NNESS* indices: their numerators represent a number of shared species between samples except that in $ESS_{ij/k}$ the size of each sample is first standardized to a common subsample size k. The denominators of Sørensen and *NNESS* indices differ somewhat as it is the mean number of species found within each sample for the former, and the mean number of shared species between two independent random draws of k specimens from the same sample for the latter. These denominators allow the indices to range between 0 for samples without shared species to 1 for strictly identical samples. Note that the Morisita–Horn index is equal to the $NNESS_{ij/k\,=\,1}$. The software *Biodiv*R 1.2 (Hardy 2010) was used to compute the *NNESS* index.

C_{ij} was computed for each pair of samples and $NNESS_{ij/k}$ was computed considering three k values ($k = 1$, $k = 25$, and $k = 100$) for each pair of samples with sample sizes at least equal to k.

Correlations between the C_{ij} matrix and the $NNESS_{ij/k}$ matrices were assessed with Mantel correlation tests using R statistical software (http://www.r-project.org/). These matrices were also correlated with a matrix of geographic distances between the centers of the grid cells. Finally, to assess how a classification of grid cells into floristic units is affected by the type of similarity measures, each similarity matrix was treated by nonmetric multidimensional scaling (NMDS) and a clustering method using *Primer6*©. Options used for NMDS were the following: Kruskal stress formula = 1; minimum stress = 0.01; number of restarts = 100. Clustering was made by the group average method (Clarke 1993).

RESULTS

SPECIES ABUNDANCE, DISTRIBUTION, AND VARIATIONS IN SAMPLING INTENSITY.—The compiled dataset contains 115 taxa of Orchidaceae and 207 of Rubiaceae, all endemic to Atlantic Central Africa. They represent, respectively, 18.8 percent and 20.2 percent of total Orchidaceae ($N = 613$) and Rubiaceae ($N = 1026$)

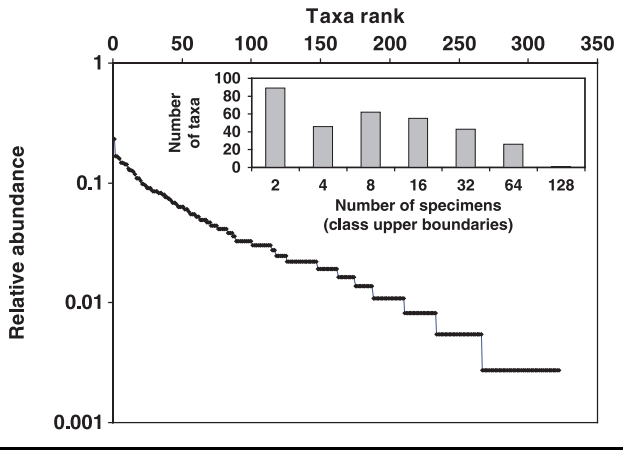

FIGURE 1. Species abundance distribution represented (i) by the relative abundance of each taxon according to its abundance rank (main graph) and (ii) by a histogram of the number of species represented by given abundance intervals following a $\log_2$ scale (inset graph). Abundance is represented here by the number of collected specimens per species.

currently recorded in the region but not necessarily endemic to Atlantic Central Africa. Many of the species were represented by very few specimens (Fig. 1), and 89 of the 322 used in our study have been collected no more than twice.

The geographic distribution of raw species richness was highly correlated with the historical sampling effort, measured as the number of specimens collected or as the number of different collectors per grid cell (Fig. 2; Tables 1 and S1). The West Province of Cameroon, and especially the Mount Cameroon area, showed the highest sampling intensity. Nearly one-third (30.7%) of all the specimens considered here were collected in the grid covering this mountain and the number of different collectors was also three times higher there than anywhere else.

SPECIES DIVERSITY: SCALE AND GRADIENT.—S_R was relatively well correlated to N_e, $S_{(k = 25)}$ and $S_{(k = 100)}$ (Pearson's correlation ranging from 0.66 to 0.84) when considering all grid cells with at least 6, 25, or 100 specimens, respectively (Tables 1 and S1). Nearly the same correlation values were obtained for 0.5° and 1° square grid cells. However, the ranking of diversity among well-sampled grid cells (>200 specimens) differed: S_R ranking (grid cell B4 > C5 > C6 > B3 > D5; Fig. 2C) followed the ranking of sample sizes (Fig. 2A) while N_e, $S_{(k = 25)}$ and $S_{(k = 100)}$ congruently indicated a different ranking (B3 > B4 > C6 > C5 > D5; Figs. 2D, E, and F). Hence, the regular north-south gradient detected with S_R along the coast essentially disappeared when considering unbiased indices. By contrast, the west-east diversity gradient observed with S_R was conserved when considering N_e, $S_{(k = 25)}$ and $S_{(k = 100)}$ (comparisons between grid cells C5, D5 and E5 on Figs. 2C, D, E, and F). Another example of contrasted behavior between raw species richness and unbiased diversity measures is given when comparing the well-sampled grid cell B4 (1117 specimens) with the much less well-sampled grid cell D6 (72 specimens): S_R is more than three times higher in the former (160 against 47 species) while N_e and $S_{(k = 25)}$ indicate nearly identical diversity.

SPECIES SIMILARITY: *NNESS* VS. SØRENSEN.—According to the Mantel tests, similarity matrices calculated with the Sørensen and *NNESS* indices were highly positively correlated (Pearson's correlation ranging from 0.84 to 0.97; Table 2). The correlation coefficients were broadly the same for 0.5° and 1° square grid cells, except when $k = 1$, which yielded a lower coefficient with 0.5° square grid cells. Correlation coefficients increased with subsampling size. NMDS and clustering methods applied on the different similarity indices also showed congruent results (Fig. 3). For a given grid cell size, the correlation between floristic similarity and spatial distance was nearly the same using the Sørensen and *NNESS* similarity indices (Table 2). As might be expected, these correlations were less strong in absolute value when more grid cells with smaller number of specimens were considered (low k).

DISCUSSION

POTENTIAL BIASES IN DISTRIBUTION DATA AND IMPLICATIONS FOR PERCEIVING DIVERSITY.—In this study, we have assessed how heterogeneous sampling effort can affect the perception of local diversity in Cameroon. Previous authors (Reddy & Davalos 2003; Hortal *et al.* 2007, 2008) have already stressed that various assessments of patterns of species richness and endemism reflect taxonomic, temporal, or geographic biases. Despite the fact that the tropical forests of Cameroon are probably the most botanically sampled areas of Central Africa, sampling efforts are highly heterogeneous geographically and may result in a biased description of diversity gradients. Our results show that the perception of diversity patterns is partially biased when regarding endemism patterns of two large plant families (Orchidaceae and Rubiaceae) in Cameroon. While our subsampling procedure confirms that inland (eastern) forests are much less diverse in endemic species than those near the Atlantic ocean (in the west), it also demonstrates that the apparently higher species richness of the northern part of the western forests compared with their southern part is an artefact of the much higher sampling effort conducted in the Mount Cameroon region. Distortions of apparent diversity gradients due to heterogeneous sampling effort indeed occur, but the high raw species richness of the most intensively sampled area also reflects its high intrinsic diversity. The positive correlation between sampling effort and unbiased diversity indices also indicates that botanists generally focussed on areas of high diversity.

One should note that higher species richness per grid cell is probably correlated not only with the number of specimens but also with the number of collecting localities. In fact, a set of specimens can be expected to contain more species if the specimens were collected in many localities than if they were collected in few localities due to spatial autocorrelation. Another source of biases is the different sampling strategies concerning plant families (*i.e.*, most collectors collect specific families). Even a robust procedure with respect to sample size cannot compensate for these two sources of biases and what we call 'unbiased' indices

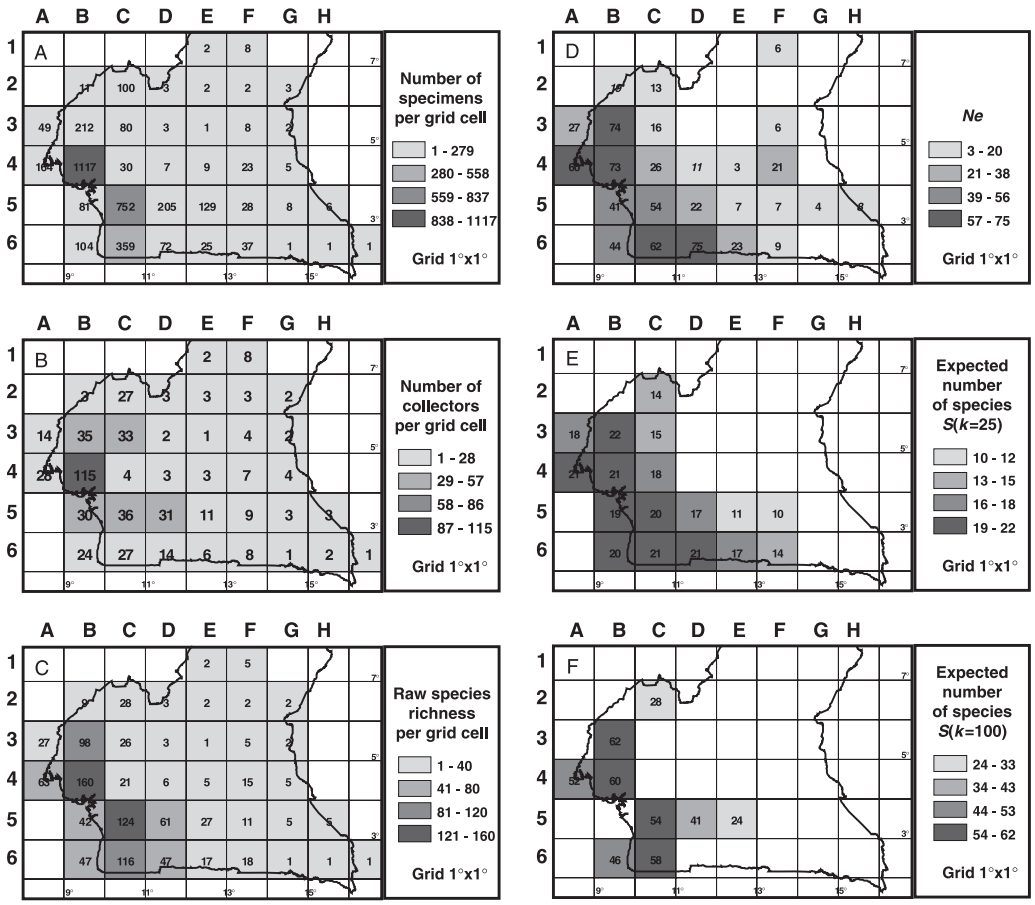

FIGURE 2. Maps with 1° grid cells showing raw data on sampling intensity and species richness (left) and indices of species diversity computed for different subsample sizes (right). (A) Number of specimens per grid cell; (B) number of collectors per grid cell; (C) raw species richness; (D) N_e calculated for cells that contain at least six specimens; (E) $S_{(k = 25)}$ calculated for cells that contain at least 25 specimens; (F) $S_{(k = 100)}$ calculated for cells that contain at least 100 specimens. N_e values higher than the number of specimens per grid cell are italicized, indicating that nonnegligible bias is expected in these particular cases.

are unbiased only with respect to sample size. In addition, one cannot correct for extremely low sampling effort. Although, in theory, S_k can be estimated even with a number of specimens as small as two for $k = 2$ (yielding a result 1 or 2), it is obvious that very small sample sizes should not be taken into consideration due to the resulting lack of accuracy of diversity estimates. When increasing k, the influence of rare species increases in S_k, which approaches the actual species richness when k becomes very large. The minimal sample sizes necessary to obtain reliable estimates of diversity depend on the current level of diversity and the relative weight attributed to rare vs. common species. S_k provides unbiased estimates and using a low k value permits to deal with sample sizes as small as k but the measure will then largely neglect the contribution of rare species to the diversity. The choice of the minimal acceptable number of specimens is a question of compromise between the spatial extent of the area that can be analyzed, the precision of the estimates and the weight given to rare species (Fig. S1). N_e attributes a relatively high weight to common species (Hill number of order $q = 2$) and can be biased under very low sample size. Simulations show that the bias in N_e becomes negligible when the sample size exceeds the true N_e (Nielsen *et al.* 2003) so that the minimal sample size required for estimating N_e should reach *ca* 100 specimens in western Cameroon, while 20 specimens might suffice in eastern Cameroon. It is also important to note that an estimate of how many specimens are needed to obtain an unbiased estimate of the number of species in a grid cell is only meaningful if it is based on a nonbiased field sampling scheme. If all specimens come from a same locality, the diversity estimates will never become representative for the entire grid cell, no matter how many specimens are produced from that locality.

Despite the risk of bias under insufficient sample size, N_e has the advantage of measuring true diversity *sensu* Jost (2007), expressing an effective number of species conforming to the replication principle (Tuomisto 2010a). The fact that N_e was often an order of magnitude lower in eastern Cameroon than in western Cameroon demonstrates the steep longitudinal gradient in endemic species richness. By comparison, the same contrasts were much less marked using S_k, especially for low k values (Fig. 2). In fact, the unbiased property of S_k makes it adequate to

TABLE 1. *Pearson's correlation between indicators of sampling intensity (number of specimens and number of collectors per grid cell) and indices of diversity (raw S_R, raw species richness; N_e, estimator of the effective number of species; $S_{(k)}$, expected number of species). The upper right half of the matrix indicates the number of grid cells considered (n) whereas the lower left half gives the Pearson's correlation coefficients. Values are given for 1° square grid cells; values obtained for 0.5° square grid cells are very similar (Table S1).*

	Number of specimens	Number of collectors	Raw S_R	N_e	$S_{(k = 25)}$	$S_{(k = 100)}$
Number of specimens	—	$n = 37$	$n = 37$	$n = 25$	$n = 17$	$n = 9$
Number of collectors	0.898*	—	$n = 37$	$n = 25$	$n = 17$	$n = 9$
Raw S_R	0.907*	0.866*	—	$n = 25$	$n = 17$	$n = 9$
N_e	0.612*	0.628*	0.832*	—	$n = 17$	$n = 9$
$S_{(k = 25)}$	0.413	0.408	0.662*	0.907*	—	$n = 9$
$S_{(k = 100)}$	0.536	0.500	0.839*	0.976*	0.964*	—

*significant tests ($P < 0.01$).

TABLE 2. *Mantel tests between Sørensen similarity, NNESS similarity and geographical distance matrices. Pearson's correlations are given for 0.5° and 1° grid cells.* n *is the number of grid cells considered.*

	Subsampling grid size = 1° × 1°			Subsampling grid size = 0.5° × 0.5°		
Correlation	$k = 1$; $n = 37$	$k = 25$; $n = 17$	$k = 100$; $n = 9$	$k = 1$; $n = 98$	$k = 25$; $n = 22$	$k = 100$; $n = 9$
(Sørensen, *NNESS*)	0.839*	0.909*	0.969*	0.911*	0.926*	0.973*
(*NNESS*, distance)	−0.271*	−0.596*	−0.779*	−0.209*	−0.623*	−0.739*
(Sørensen, distance)	−0.345*	−0.583*	−0.761*	−0.239*	−0.630*	−0.796*

*significant tests ($P < 0.01$).

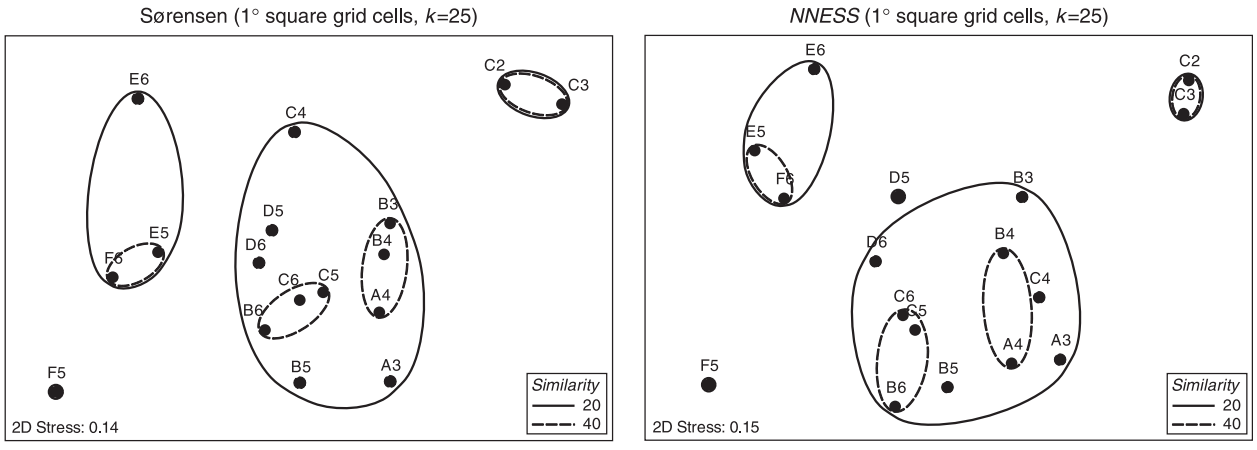

FIGURE 3. Ordination of 1° grid cells combining results from Nonmetric multidimensional scaling (points) and clustering (line-delineated groups) obtained from Sørensen similarity (left panel) and $NNESS_{(k = 25)}$ similarity (right panel) matrices. Continuous and dotted lines represent, respectively, 20 and 30 percent similarity groups from clustering. The same congruence between Sørensen and *NNESS* similarity is observed using 0.5° grid cells (results not shown).

rank diversity but because S_k does not conform to the replication principle, ratios of S_k among samples cannot be interpreted as ratios of true diversity *sensu* Jost (2007). Ongoing theoretical work for transforming S_k into measures of effective number of species with good statistical properties and conforming to the replication principle should lead to improved estimators of true diversity.

SPECIES SIMILARITY: PRESENCE–ABSENCE VS. ABUNDANCE.—There are numerous concepts and methods for measuring β-diversity or species turnover (Tuomisto 2010a), among which (dis)similarity measures are the simplest and most commonly used to describe species turnover (Jurasinski *et al.* 2009). One reason for their popularity is that they are easy to calculate and their results are easy to interpret. Many of these indices use presence/absence data and do not take into account the relative abundances of species. The fact that all species have the same weight in presence/absence indices is a fundamental, and in many cases a desirable, property of these indices. The Sørensen index is one of the oldest

and most widely used similarity indices for assessing compositional similarity of assemblages (Chao *et al.* 2005). Despite its broad application in ecological studies, the Sørensen index has been shown to perform poorly as a measure of similarity between assemblages that include a substantial fraction of rare species (Plotkin & Muller-Landau 2002). Its underestimation of similarity occurs because of the failure to account for unseen shared species, *i.e.*, species that are likely to be present in a larger homogeneous sample of the assemblage, but that are missing from actual sample data (Chao *et al.* 2005). Moreover, the main drawback of presence/absence similarity indices remains the fact that a species dominating an assemblage carries no more weight in this species turnover measure than a species represented by a singleton. This has led to the development of a large number of similarity measures based on abundance data, among which the Morisita–Horn measure is widely used (Magurran 2004). The *NESS* and *NNESS* indices, which are generalizations of the Morisita index and the Morisita–Horn index, respectively, have received relatively little attention despite their advantage for correcting unequal sample sizes. As for S_k, rare species are better accounted by $NNESS_{ij/k}$ when the subsampling size k increases (Fig. S1).

Researchers are usually not interested in the smallest sampling unit they work on (such as a set of available herbarium specimens), but they want to extrapolate results to larger units (such as a 1° square grid cell). The fact that the Sørensen index is sensitive to rare species means that its value is difficult to extrapolate, but when interpreted in terms of the available data, it is just as accurate as any other measure (see Tuomisto 2010b for a discussion on extrapolation problems related to species turnover). Despite its inherent sampling bias, our comparison of the Sørensen similarity index with $NNESS_{ij/k}$ shows that they are highly correlated (Table 2) and that the identification of floristic units is highly congruent in both cases. This suggests that the use of the Sørensen index with heterogeneous number of specimens does not convey an overly biased perception of species turnover, in contrast to the use of raw species richness for describing local diversity patterns. Comparing Sørensen and *NESS* indices for tropical moth ensembles, Brehm and Fiedler (2004) show that 'at least under certain conditions' simple presence–absence measures such as Sorensen's index can be useful. Nevertheless, the classical Jaccard or Sørensen indices are known to depend, often strongly, on sampling intensities and diversity (Wolda 1981, Lande 1996, Chao *et al.* 2006), and we recommend the use of unbiased similarity measures whenever possible. We expect that the partially de-biased estimators of Sørensen or related coefficients (*e.g.*, Chao *et al.* 2005) should also perform satisfactorily.

SOUTHERN CAMEROON, A BIODIVERSITY HOTSPOT NEGLECTED BY SCIENTISTS.—The spatial resolution of prioritization schemes quantifying the global distribution of biodiversity (Myers *et al.* 2000) is very low (Soria-Auza & Kessler 2008), a fact that necessitates a finer scale approach such as the one we developed here. The designation and management of protected areas are usually of national concern (Ceballos & Brown 1995), and the 192 countries that have signed the convention on biological diversity have national focal points charged with following up national strategies and plans for conserving biodiversity. In addition, universities and research institutions located in these countries are also subsidized at the national level. To manage and preserve biodiversity at this scale, more precise analyses are needed to identify areas that encompass remarkable and/or complementary biodiversity to those that have been previously highlighted. In most tropical countries, however, identification of priority areas for conservation is almost always based on spatially heterogeneous levels and intensities of biodiversity inventory work.

In Cameroon, politicians and scientists have focused particular attention on several mountain massifs including Mt. Cameroon, Mt. Oku, and Kupe/Bakossi. In contrast to these well-inventoried areas, the Bipindi/Akom II massif in South Province (cells C5 and C6 in Fig. 2) as well as the Ntem basin (cells C6 and D6 in Fig. 2) remain largely neglected, even though our results suggest that they are probably as diverse. The same situation occurs in Central Province where the remaining small patches of primary vegetation are now under considerable threat (results not shown, observed in our 0.5° square grid cells analyses). During the last decade, field inventories focusing on Orchidaceae and Rubiaceae conducted in the Central and South Provinces led to the discovery of many new and rare species (*e.g.*, Droissart *et al.* 2006, Simo *et al.* 2009, Sonké *et al.* 2009), which strongly supports the results of the analyses presented here.

Our study confirms that detecting and analyzing sampling bias can highlight geographical areas where further research is likely to be particularly productive (Reddy & Davalos 2003). Our comparison shows that botanical exploration has largely been concentrated in western Cameroon while the southeastern part of the country has been so neglected that it is not even possible to generate reliable estimates of species diversity (Figs. 2E and F). The tendency for botanists to collect where they expect to find the highest level of diversity could have negative consequences by reinforcing existing biases rather that overcoming them. The timely use of the kind of approach presented here can play a key role in avoiding such situations, thereby ensuring that future inventory work is focused in a way that contributes maximally to understanding geographic patterns of biodiversity and informing conservation planning and priority setting.

ACKNOWLEDGMENTS

We express our gratitude to Professor Jean Lejoly from the Université Libre de Bruxelles for his support in his laboratory and to Dr Porter P. Lowry II for helpful comments. We are also grateful to three anonymous reviewers for constructive remarks on earlier versions of this manuscript. The surveys carried out in Central Africa were supported by the ECOFAC Program (EC-DG8), DIVEAC (CUD-ULB), the Fonds Leopold III, the Wildlife Conservation Society (WCS), the Fonds de la Recherche Scientifique (F.R.S.-FNRS), the American Orchid Society (AOS), and the 'Sud Expert Plantes' project under the French Ministry of Foreign Affairs. Visits by Droissart to the herbaria of Wageningen, Paris,

and Kew were funded by the European Commission's Research Infrastructure action via the SYNTHESYS Project (applications FR-TAF-2418 and NL-TAF-1611) and by the F.R.S.-FNRS where Hardy is a research associate. The curators of these herbaria are gratefully acknowledged. Stévart's participation was supported by the U.S. National Science Foundation (1051547, TS as PI).

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

TABLE S1. *Pearson's correlation between indicators of sampling intensity and indices of diversity. Values are given for 0.5° square grid cells.*

FIGURE S1. Schematic behavior of species diversity(S_k) and floristic similarity($NNESS_{ij/k}$) indices in connection with subsampling size variation.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

LITERATURE CITED

Brehm, G., and K. Fiedler. 2004. Ordinating tropical moth ensembles from an elevational gradient: A comparison of common methods. J. Trop. Ecol. 20: 165–172.

Burgess, N., W. Kuper, J. Mutke, J. Brown, S. Westaway, S. Turpie, C. Meshack, J. Taplin, C. McClean, and J. C. Lovett. 2005. Major gaps in the distribution of protected areas for threatened and narrow range afrotropical plants. Biodiversity Conserv. 14: 1877–1894.

Ceballos, G., and J. H. Brown. 1995. Global patterns of mammalian diversity, endemism, and endangerment. Conserv. Biol. 9: 559–568.

Chao, A., R. L. Chazdon, R. K. Colwell, and T. J. Shen. 2005. A new statistical approach for assessing similarity of species composition with incidence and abundance data. Ecol. Lett. 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.

Clarke, K. R. 1993. Non-parametric multivariate analyses of changes in community structure. Aust. J. Ecol. 18: 117–143.

Delisle, F., C. Lavoie, M. Jean, and D. Lachance. 2003. Reconstructing the spread of invasive plants: Taking into account biases associated with herbarium specimens. J. Biogeogr. 30: 1033–1042.

Dennis, R. L. H., and C. D. Thomas. 2000. Bias in butterfly distribution maps: The influence of hot spots and recorder's home range. J. Insect Conserv. 4: 73–77.

Droissart, V., B. Sonké, and T. Stévart. 2006. Les Orchidaceae endémiques d'Afrique centrale atlantique présentes au Cameroun. Syst. Geogr. Plants 76: 3–84.

Freitag, S., C. Hobson, H. C. Biggs, and A. S. Van Jaarsveld. 1998. Testing for potential survey bias: The effect of roads, urban areas and nature reserves on a southern African mammal data set. Anim. Conserv. 1: 119–127.

Govaerts, R., M. A. Campaccil, D. H. Baptista, P. J. Cribb, A. George, K. Kreuz, and J. Wood. 2010a. World checklist of Orchidaceae. Available at http://www.kew.org/wcsp/ (accessed 5 October 2010).

Govaerts, R., M. Ruhsam, L. Andersson, E. Robbrecht, D. Bridson, A. Davis, I. Schanzer, and B. Sonké. 2010b. World checklist of Rubiaceae. Available at http://www.kew.org/wcsp/ (accessed 5 October 2010).

Grand, J., M. P. Cummings, T. G. Rebelo, T. H. Ricketts, and M. C. Neel. 2007. Biased data reduce efficiency and effectiveness of conservation reserve networks. Ecol. Lett. 10: 364–374.

Grassle, J. F., and W. Smith. 1976. A similarity measure sensitive to the contribution of rare species and its use in investigation of variation in marine benthic communities. Oecologia 25: 13–22.

Hardy, O. J. 2010. BiodivR 1.2. A program to compute statistically unbiased indices of species diversity within sample and species similarity between samples using rarefaction principles. Available at http://ebe.ulb.ac.be/ebe/Software.html (accessed 21 March 2011).

Hill, M. O. 1973. Diversity and evenness: A unifying notation and its consequences. Ecology 54: 427–432.

Holmgren, P. K., and N. H. Holmgren. 1998 onwards [continuously updated]. Index Herbariorum. Available at http://sciweb.nybg.org/science2/IndexHerbariorum.asp (accessed 1 November 2007).

Hortal, J., A. Jimenez-Valverde, J. F. Gomez, J. M. Lobo, and A. Baselga. 2008. Historical bias in biodiversity inventories affects the observed environmental niche of the species. Oikos 117: 847–858.

Hortal, J., J. M. Lobo, and A. Jimenez-Valverde. 2007. Limitations of biodiversity databases: Case study on seed-plant diversity in Tenerife, Canary Islands. Conserv. Biol. 21: 853–863.

Hurlbert, S. H. 1971. The nonconcept of species diversity: A critique and alternative parameters. Ecology 52: 577–586.

Jost, L. 2007. Partitioning diversity into independent alpha and beta components. Ecology 88: 2427–2439.

Jurasinski, G., V. Retzer, and C. Beierkuhnlein. 2009. Inventory, differentiation, and proportional diversity: A consistent terminology for quantifying species diversity. Oecologia 159: 15–26.

Kier, G., J. Mutke, E. Dinerstein, T. H. Ricketts, W. Kuper, H. Kreft, and W. Barthlott. 2005. Global patterns of plant diversity and floristic knowledge. J. Biogeogr. 32: 1107–1116.

Kuper, W., J. H. Sommer, J. C. Lovett, and W. Barthlott. 2006. Deficiency in African plant distribution data—Missing pieces of the puzzle. Bot. J. Linn. Soc. 150: 355–368.

Kuper, W., J. H. Sommer, J. C. Lovett, J. Mutke, H. P. Linder, H. J. Beentje, R. Van Rompaey, C. Chatelain, M. Sosef, and W. Barthlott. 2004. Africa's hotspots of biodiversity redefined. Ann. Mo. Bot. Gard. 91: 525–535.

Lande, R. 1996. Statistics and partitioning of species diversity, and similarity among multiple communities. Oikos 76: 5–13.

Magurran, A. E. 2004. Measuring biological diversity. Blackwell Publishing, Oxford, U.K.

Moerman, D. E., and G. F. Estabrook. 2006. The botanist effect: Counties with maximal species richness tend to be home to universities and botanists. J. Biogeogr. 33: 1969–1974.

Myers, N., R. A. Mittermeier, C. G. Mittermeier, G. A. B. da Fonseca, and J. Kent. 2000. Biodiversity hotspots for conservation priorities. Nature 403: 853–858.

Nielsen, R., D. R. Tarpy, and H. K. Reeve. 2003. Estimating effective paternity number in social insects and the effective number of alleles in a population. Mol. Ecol. 12: 3157–3164.

Petrík, P., J. Pergl, and J. Wild. 2010. Recording effort biases the species richness cited in plant distribution atlases. Perspect. Plant Ecol. Evol. Syst. 12: 57–65.

Plotkin, J. B., and H. C. Muller-Landau. 2002. Sampling the species composition of a landscape. Ecology 83: 3344–3356.

Prendergast, J. R., S. N. Wood, J. H. Lawton, and B. C. Eversham. 1993. Correcting for variation in recording effort in analyses of diversity hotspots. Biodiversity Lett. 1: 39–53.

Reddy, S., and L. M. Davalos. 2003. Geographical sampling bias and its implications for conservation priorities in Africa. J. Biogeogr. 30: 1719–1727.

Rich, T. C. G., and E. R. Woodruff. 1992. Recording bias in botanical surveys. Watsonia 19: 73–95.

Routledge, R. D. 1980. Bias in estimating the diversity of large, uncensused communities. Ecology 61: 276–281.

Schulman, L., T. Toivonen, and K. Ruokolainen. 2007. Analysing botanical collecting effort in Amazonia and correcting for it in species range estimation. J. Biogeogr. 34: 1388–1399.

Simo, M., V. Droissart, B. Sonké, and T. Stévart. 2009. The orchid flora of the Mbam Minkom Hills (Yaoundé, Cameroon). Belg. J. Bot. 142: 111–123.

Sonké, B., M. Simo, and S. Dessein. 2009. Synopsis of the genus *Mitriostigma* (Rubiaceae) with a new monocaulous species from Southern Cameroon. Nord. J. Bot. 27: 305–312.

Sørensen, T. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. Biologiske Skrifter 5: 1–34.

Soria-Auza, R. W., and M. Kessler. 2008. The influence of sampling intensity on the perception of the spatial distribution of tropical diversity and endemism: A case study of ferns from Bolivia. Divers. Distrib. 14: 123–130.

Tuomisto, H. 2010a. A diversity of beta diversities: Straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. Ecography 33: 2–22.

Tuomisto, H. 2010b. A diversity of beta diversities: Straightening up a concept gone awry. Part 2. Quantifying beta diversity and related phenomena. Ecography 33: 23–45.

White, F. 1979. The Guineo-Congolian Region and its relationships to other phytochoria. Bull. Jard. Bot. Nat. Belg. 49: 11–55.

Williams, P. H., C. R. Margules, and D. W. Hilbert. 2002. Data requirements and data sources for biodiversity priority area selection. J. Biosci. 27: 327–338.

Wolda, H. 1981. Similarity indices, sample size and diversity. Oecologia 50: 296–302.