Theoretical and empirical results for the design-based inference on rarefaction curves

by

L. Fattorini [1] **and L. D'Alessandro** [2]

[1] Dipartimento di Metodi Quantitativi, Università di Siena, Siena (Italy)

[2] Monte dei Paschi di Siena, Siena (Italy)

Key words: plant communities; species list; probabilistic sampling; presence-absence data; accumulation curves, design-based inference; extrapolation, bootstrap confidence bands, Monte-Carlo studies.

Abstract

Statistical inference on accumulation curves is considered from a design-based perspective. Preliminaries on probabilistic sampling of plants and species are given. Rarefaction curves and its statistical properties are derived in a design-based framework, solely on the basis of the independence among the replications of the sampling scheme adopted to select plants. A design-based extrapolation of rarefaction curves is proposed. Confidence bands around the curves and the extrapolated values are constructed by means of a modified bootstrap procedure. The reliability of the bootstrap confidence bands is empirically validated by means of a simulation study. The use of some model-based procedures is also considered and compared with the design-based procedures proposed in the paper. An application to a case study is reported

1. Introduction.

Many ecological studies require the analysis of species diversity in a plant community. In this framework, Hurlbert (1971, p.584-585) points out that the use of diversity measures based on species abundance (*e.g* the Shannon index) makes sense if the plant community coincides with a taxonomic group in which species "*are likely to be about of the same size, have similar life histories and compete over both evolutionary and ecological time*". On the other hand, if the taxocene is too inclusive, the interpretation of abundance-based diversity measures becomes weak "*because individuals belonging to different species will be highly non-equivalent (e.g. in size, life history, etc.)*". When the community contains species with very different characteristics, the trivial list of species probably constitutes the simplest and most direct way to analyse diversity, thus bypassing the problem of quantifying abundance. However, since in most cases the complete list of species constitutes an unknown characteristic of the community under study, it must be compiled by means of purposive investigations of the study area, as performed traditionally by botanist, or throught sample surveys. In both cases it is essential to adopt reliable methodologies to check the effectiveness of the protocol adopted for compiling species list.

A widely-applied procedure to assess if the interest sites have been sufficiently investigated is the use of curves showing the increases of species detection as the survey effort or some related proxy increases. A curve which shows little change at its end (a horizontal or nearly horizontal line) shows that few or no species are being collected by means of a further increase of the sampling effort. On the other hand, a curve which continues to rise sharply near its end shows that many new species are still being found thanks to further efforts. In relatively new terminology these curves are referred to as

accumulation curves, even if their use in ecology has a long history and dates back to the 1920s, when they were known to taxonomists as the *collectors' curves*.

The present paper deals with statistical inference on these curves from a design-based point of view, when species are collected by means of independent replications of a probabilistic sampling scheme. As is well known, the main appeal of design-based inference is that it stems from the characteristics of the sampling scheme and, in contrast with model-based inference, it avoids assumptions about the community under study.

In section 2, preliminaries on probabilistic sampling of species are given; while in section 3 the probabilistic structure of presence-absence data is derived simply on the basis of the independence among replications, which constitutes the sole requisite adopted. In section 4, accumulation and rarefaction curves are introduced and their relations are considered from a design-based point of view. Subsequently, in section 5, the design-based characteristics of rarefaction curves, such as expectation and variance-covariance matrix, are obtained. As the number of replications increases, asymptotic results are derived in section 6. In section 7, the use of modified bootstrap confidence bands is proposed and empirically validated by means of a Monte Carlo study, Finally, section 8 is devoted to a critical comparison of the design-based approach with some widely-applied model-based procedures, while section 9 contains some concluding remarks.

Troughout the paper, the binomial coefficient $\binom{a}{b}$ will be set to 0 for any couple of integers a,b such that $a < b$. Moreover, $I(\bullet)$ will denote the indicator function equal to 1 if $\bullet$ is true and 0 otherwise, $[\bullet]$ will denote the greatest integer not exceeding $\bullet$ and

u_p will denote the *p*-quantile of the standard normal distribution function. Finally, for simplicity of notation, random vectors and random variables will not be distinguished by their realizations.

2. Prelininaires on sampling species.

Consider a plant community within a delineated study area. From a statistical point of view the community constitutes a without-frame population of *N* units spread over the area. Referring to the units by their labels, the population may be represented by the set $\mathsf{U} = \{1,2,\mathrm{K}\ ,N\}$. If *K* species are present, the community is partitioned into *K* sub-populations $\mathsf{U}_1,\mathsf{U}_2,\mathrm{K}\ ,\mathsf{U}_K$ of sizes $N_1.N_2,\mathrm{K}\ ,N_K$, where U_k denotes the set of labels identifying the N_k individuals belonging to species *k*. Usually, the frame of species partitioning the community is referred to as *species list* while *K* is referred to as *species richness*. Moreover, the N_k's and the p_k's, with $p_k = N_k / N$ $(k = 1,\mathrm{K}\ ,K)$, are referred to as *abundances* and *relative abundances*, respectively.

Owing to lack of a community frame, the most effective schemes for sampling plant populations differ from the traditional schemes, their choice being mainly determined by practical considerations on the nature of the community to be sampled. For example, when dealing with a tree population, *Bitterlich sampling*, usually referreed to as *variable circular plot* sampling is adopted, while in the case of a shrub population, *line intercept sampling* may be suitable. Alternatively, if the target population is formed by a taxocene containing very different species (*e.g.* all the vascular plants in a forest), *floating plot sampling* should be adopted, in which a point is randomly thrown onto the area and the sampled plants are those included in a circular or square plot of a pre-fixed size centered at the random point. All these sampling techniques have been developed

empirically in field investigations and have long been set apart from the core of the statistical world. More recently, some authors (*e.g.* De Vries, 1986, Thompson, 1992, Schreuder *et al.*, 1993, Overton and Stehman, 1995) have attempted to connect many of these methods with basic sampling theory.

Denote by $\mathsf{S} \subset \mathsf{U}$ the sample of units selected from the plant community by means of one of these schemes. Any scheme induces the probabilities that single plants enter the sample, say $\tau_1, \mathrm{K}, \tau_N$, which are referred to as *plant inclusion probabilities*. Usually, environmental schemes are ruled in such a way that the inclusion probabilities for (at least) the selected plants can be readily determined directly or by some field measurements. Indeed, the quantification of the inclusion probabilities, at least for the sampled plants, makes possible the Horvitz-Thompson estimation for the totals of some attributes of interest for the whole community (e.g. abundance, phytomass, basal area, etc)

However, in some situations, species rather than single plants are of interest. In this case, each sub-population of plants of the same species U_k may be viewed as a unit, in such a way that the species list may be viewed as a statistical population. Thus, referring to species by their labels, any species list may be represented by the set of species labels $\mathcal{U} = \{1,2,\mathrm{K}, K\}$. However, since species are assemblages of plants, they actually constitute unknown structures spread over the study area, which cannot be sampled directly. Thus, the most effective way for sampling species is to sample plants by means of one of the above-mentioned environmental scheme, in such a way that a species is sampled when at least one plant of that species is sampled. Practically speaking, any sample of plants $\mathsf{S} \subset \mathsf{U}$ univocally determines the corresponding sample of species, say $\mathcal{S} \subset \mathcal{U}$.

Hence, the environmental scheme adopted to sample plants, univocally determines the *species sampling design*, i.e the probability distribution over the collection $\mathbb{S}$ of all the possible samples $\mathcal{S} \subset \mathcal{U}$, which, in turn, determines the probabilities that single species or couples of species enter the sample, say π_k and π_{kl} $(l > k = 1,\mathrm{K}\,,K)$. These probabilities will be referred to as *first- and second order species inclusion probabilities*. It is worth noting that, even if these schemes are designed to quantify the inclusion probabilities of (at least) the sampled plants (as pointed out in section 2.1), the quantification of species inclusion probabilities is generally precluded. Indeed, the quantification of the π_k s should entail the knowledge of all the units belonging to a species together with their spatial distribution over the study area. As a very simple example, consider an artificial community of $N = 3$ shrubs, say $\mathsf{U} = \{1,2,3\}$, spread over a 10×10 square and suppose that the first two shrubs belong to species 1 (white species) and the remaining shrub belongs to species 2 (grey species) (see Figure1). Accordingly, $K = 2$, $\mathsf{U}_1 = \{1,2\}$, $\mathsf{U}_2 = \{3\}$ and $\mathcal{U} = \{1,2\}$. Now, suppose that the shrubs are sampled when encountered by a transect starting from a point randomly selected on the base of the square, perpendicularly crossing the whole square. All the possible samples of plants and the corresponding samples of species which can be drawn by means of that scheme are listed in Table 1, together with their corresponding probabilities. By adding the probabilities of all the samples including a unit or a species, the inclusion probabilities of that unit or species can be straightforwardly obtained. Accordingly, from Table 1 it turns out that $\tau_1 = 0.4$, $\tau_2 = 0.4$, $\tau_3 = 0.3$ while $\pi_1 = 0.6$ and $\pi_2 = 0.3$. It is worth noting that the inclusion probability of each shrub is determined by the length of its projection onto the base. Thus, the inclusion probability

of each shrub, if sampled/intercepted, can be trivially determined by measuring the maximum shrub width (perpendicular to the transect). On the other hand, the inclusion probabilities of each species are determined by the length of the union of the projections of all the shrubs of that species and as such they cannot be quantified on the basis of the sampled shrubs.

Figure 1. Graphic representation of a community of 3 shrubs of white and grey species on a square region of size 10×10.

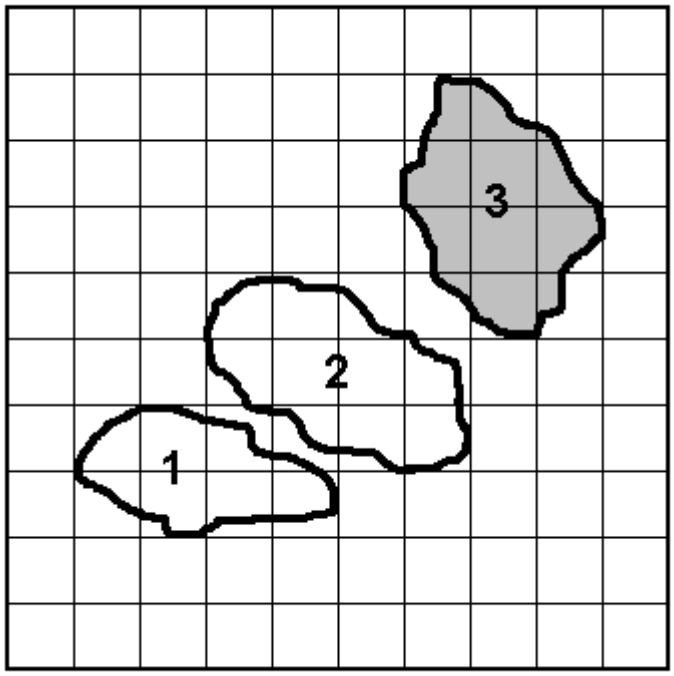

Table 1. List of all the possible samples of shrubs and species arising from line-intercept sampling performed on the shrub population in Figure 1 by means of a transect with its starting point randomly selected on the base of the square.

sample of shrubs - S	sample of species - $G(S)$	probability
$\varnothing$	$\varnothing$	0.2
$\{1\}$	$\{1\}$	0.2
$\{2\}$	$\{1\}$	0.1
$\{3\}$	$\{2\}$	0.2
$\{1,2\}$	$\{1\}$	0.2
$\{2,3\}$	$\{1,2\}$	0.1

3. Design-based properties of presence-absence data

Now consider a scheme adopted for sampling a plant community which induces a sampling design over $\mathcal{S}$. In this framework it is convenient to introduce a one-to-one mapping z from $\mathcal{S}$ to $\{0,1\}^K$, such that $\mathbf{z} = z(S)$ turns out to be a K-vector $\mathbf{z} = [z_1, z_2, \mathrm{K}, z_K]^{\mathrm{T}}$ with $z_k = I(k \in S)$ $(k = 1, \mathrm{K}, K)$. Accordingly, $\mathbf{z}$ constitutes a discrete random vector with support $\{0,1\}^K$ and probability function

$$p(\mathbf{c}_m) = \theta_m \quad , \quad \mathbf{c}_m \in \{0,1\}^K \tag{1}$$

where $\mathbf{c}_1, \mathrm{K}, \mathbf{c}_M$ represent the $M = 2^K$ points of $\{0,1\}^K$ written in a lexicographic order such as $(0, \mathrm{K}, 0), \mathrm{K}, (1, \mathrm{K}, 1)$ and $\boldsymbol{\theta} = [\theta_1, \mathrm{K}, \theta_M]^{\mathrm{T}}$ is an M-dimensional parameter varying in the parametric space $\Theta = \left\{ \boldsymbol{\theta}; 0 \le \theta_j \le 1, \sum_{j=1}^{M} \theta_j = 1 \right\}$. Practically speaking, all the uncertainty steamed from the sampling design is accomplished in (1) by means of the parameter $\boldsymbol{\theta}$. Henceforth, P_θ, E_θ, V_θ and C_θ will denote probability measure,

expectation, variance and covariance with respect to the sampling design or to (1), equivalently.

As a consequence of this formulation, each marginal variable z_k has support $\{0,1\}$ and $\mathrm{P}_\theta(z_k = 1) = \pi_k$, in such a way that $\mathrm{E}_\theta(z_k) = \pi_k$. Accordingly, $\mathrm{E}_\theta(\mathbf{z}) = \boldsymbol{\pi}$ where $\boldsymbol{\pi} = [\pi_1, \mathrm{K}, \pi_K]^{\mathrm{T}}$ constitutes the vector of first-order species inclusion probabilities. Moreover, each marginal bivariate vector $[z_k, z_l]^{\mathrm{T}}$ has support $\{0,1\}^2$ and $\mathrm{P}_\theta(z_k = 1, z_l = 1) = \pi_{kl}$, in such a way that $\mathrm{E}_\theta(\mathbf{z}\mathbf{z}^{\mathrm{T}}) = \boldsymbol{\Pi}$, where $\boldsymbol{\Pi}$ is the *K*-simmetric matrix having π_{kl} as its *kl*-elements and π_k as its *k* diagonal element $(l > k = 1, \mathrm{K}, K)$. Obviously, $\mathrm{V}_\theta(\mathbf{z}) = \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma} = \boldsymbol{\Pi} - \boldsymbol{\pi}\boldsymbol{\pi}^{\mathrm{T}}$.

Since a study area cannot be adequately sampled by means of only one point, line or plot, usually *n* points, lines or plots are randomly and independently thrown onto the area. Obviously, the replication procedure gives rise to *n* samples of plants, say $\mathsf{S}_1, \mathrm{K}, \mathsf{S}_n$, which in turn give rise to *n* corresponding sample of species, say $\mathcal{S}_1, \mathrm{K}, \mathcal{S}_n$, which finally give rise to *n iid* realizations $\mathbf{z}_1, \mathrm{K}, \mathbf{z}_n$ from (1). The $k \times n$ matrix $\mathbf{Z}_n = [\mathbf{z}_1, \mathbf{z}_2, \mathrm{K}, \mathbf{z}_n]$, where $z_{ki} = I(k \in \mathcal{S}_i)$ $(k = 1, \mathrm{K}, K;\ i = 1, \mathrm{K}, n)$ is usually referred to as *presence-absence data*. Owing to the independence of replications, the distribution of $\mathbf{Z}_n$ is once again completely determined by the parameter $\boldsymbol{\theta}$. Accordingly $\mathrm{P}_\theta, \mathrm{E}_\theta, \mathrm{V}_\theta$ and C_θ will also be used henceforth to denote the product probability measure, expectation, variance and covariance with respect to *n* independent replications of the sampling scheme giving rise to (1).

Denote by $\mathcal{S}^n = \mathcal{S}_1 \cup \mathrm{K} \cup \mathcal{S}_n$ the pooled sample of species detected by means of *n* independent samples and by SO_n the total number of species in $\mathcal{S}^n$, usually referred to

as *species observed*. It is worth noting that only SO_n rows out of the *K* rows of $\mathbf{Z}_n$ are actually observable (all the rows having at least one 1), while the remaining $K - SO_n$ rows of zeros are lost as their corresponding species. Note also that, if at each replication the sampling scheme selects one plant only, then the *n* replications gives rise to a pooled sample $\mathcal{S}^n$ constituted by *n* plants selected with replacement from the community. For example, if like a ball in an urn, it was possible to select a plant at random from the community (i.e in such a way that all the plants have the same probability $1/N$ of being selected) , then the *n* replications of such scheme gave rise to a simple random sampling with replacement.

It is at once apparent that $\mathbf{f}_n = f(\mathbf{Z}_n)$, where $\mathbf{f}_n = [f_{1n}, \mathrm{K}\ , f_{Mn}]^{\mathrm{T}}$ is an *M*-vector in which f_{mn} denotes the absolute frequency of $\mathbf{c}_m$, constitutes the minimal sufficient statistic for $\boldsymbol{\theta}$ (see *e.g.* Lindgren, 1993, problem 7.64, p.240). Notwithstanding $\mathbf{f}_n$ provides the best synthesis of the sample data, in the subsequent developments, a more important role will be played by the vector $\mathbf{x}_n = \sum_{i=1}^{n} \mathbf{z}_i$ and the matrix $\mathbf{X}_n = \sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i^{\mathrm{T}}$. More precisely, $\mathbf{x}_n = [x_{1n}, \mathrm{K}\ , x_{Kn}]^{\mathrm{T}}$ is the *K*-vector whose *k*-component x_{kn} denotes the number of replicated samples containing species *k*. Accordingly, $x_{kn} \sim Bi(n, \pi_k)$ while the pooled sample of species can be rewritten as $\mathcal{S}^n = \{k : x_{kn} > 0\}$. Moreover, $\mathbf{X}_n$ is the *K*-symmetric matrix having x_{kln} as its *kl*-elements and x_{kn} as its *k* diagonal element, where x_{kln} denotes the number of replicated samples containing both species *k* and *l* $(l > k = 1, \mathrm{K}\ , K)$. Note that, as for the case of $\mathbf{Z}_n$, only the SO_n positive frequencies are actually observable in $\mathbf{x}_n$, while the remaining $K - SO_n$ null frequencies are lost. In the

same way, only SO_n rows/columns of $\mathbf{X}_n$ are observable, while the remaining $K - SO_n$ rows/and columns are lost.

Since $\mathbf{x}_n$ can be rewritten as $\sum_{m=1}^{M} f_{mn}\mathbf{c}_m$, it actually constitutes a function of $\mathbf{f}_n$ and as such it cannot be a sufficient statistic for $\boldsymbol{\theta}$. Rather, the following result can be proven for $\mathbf{x}_n$.

Proposition 1. $\mathbf{x}_n$ *is a complete statistic for* $\boldsymbol{\theta}$.

Proof. Consider an integrable real function *h* such that $\mathrm{E}_\theta\{h(\mathbf{x}_n)\} = 0$, for each $\boldsymbol{\theta} \in \Theta$. Accordingly, $\mathrm{E}_\theta\{h(\mathbf{x}_n)\} = 0$ for each $\boldsymbol{\theta} \in \Theta_0$, where $\Theta_0 \subset \Theta$ denotes the set of parameters $\boldsymbol{\theta}$ for which the marginal components of **z** turn out to be indipendent. In this case $\mathbf{x}_n$ is formed by *K* independent random variables $x_{kn} \sim Bi(n, \pi_k)$ $(k = 1,\mathrm{K}\,,K)$. But since $\mathbf{x}_n$ is complete for that model (see e.g. Lehmann and Casella, 1998, Example 6.26, p.44), then $h(\mathbf{x}_n) = 0$ ■

4. Design-based derivation of accumulation and rarefaction curves.

In order to avoid confusion, some unambiguous definitions regarding the curves adopted to describe the accumulation of species are necessary. From the independence of the replications, each species has probability $1-(1-\pi_k)^i$ of being sampled in at least one replication out of *i*. Accordingly, the design-based expectation of the number of species observed when *i* independent replications are performed turns out to be.

$$\gamma_i = K - \sum_{k=1}^{K}(1-\pi_k)^i \quad , \; i = 1,2,\mathrm{K} \tag{2}$$

Henceafter, the curve defined by (2) will be referred to as *design-based expected accumulation curve* while $\boldsymbol{\Gamma}_n = [\gamma_1, \mathrm{K}\ , \gamma_n]^{\mathrm{T}}$ denotes the vector containing the frst n ordinates of the curve. It is worth noting that (2) is completely determined by the first-order inclusion probabilities of species. Moreover, since the curve describes the theoretical increase in the number of species observed as the sampling effort increases, it constitutes the target parameter to be estimated from presence-absence data.

Now, quoting from Gotelli and Colwell (2001), given a collection of n samples $S_1, \mathrm{K}\ , S_n$, the *sample-based accumulation curve* is the plot of g_{in} against i $(i = 1, \mathrm{K}\ , n)$, where g_{in} represents the number of species observed in i samples, when samples are pooled in a given order. More precisely, the curve is determined by the sample statistic $\mathbf{g}_n = g_n(\mathbf{Z}_n)$, where $\mathbf{g}_n = [g_{1n}, \mathrm{K}\ , g_{nn}]^{\mathrm{T}}$ is an n-vector in which

$$g_{in} = \sum_{k \in S^n} I(x_{ki} > 0) \quad , \quad i = 1,2, \mathrm{K}\ , n \tag{3}$$

and $x_{ki} = z_{k1} + \mathrm{K} + z_{ki}$ denotes the number of replicated samples containing species k among the first i samples. If at each replication the scheme select one plant only, (3) is referred to as *individual-based accumulation curve*. In this case the accumulation units are individuals instead of samples.

For some subsequent theoretical developments, it is more convenient to rewrite (3) as a linear combination of the K random variables $I(x_{1i} > 0), \mathrm{K}\ , I(x_{Ki} > 0)$, in the sense that expression (3) is equivalent to

$$g_{in} = \sum_{k=1}^{K} I(x_{ki} > 0) \ , \quad i = 1,2, \mathrm{K}\ , n \tag{4}$$

Hence, from the independence of the replications, it is at once apparent that

$$\mathrm{E}_\theta \{ I(x_{ki} > 0) \} = \mathrm{P}_\theta (x_{ki} > 0) = 1 - \mathrm{P}_\theta (x_{ki} = 0) = 1 - (1 - \pi_k)^i \tag{5}$$

and

$$\mathrm{E}_\theta\left\{I(x_{ki}>0)I(x_{lj}>0)\right\}=\mathrm{P}_\theta(x_{ki}>0,x_{lj}>0)=1-\mathrm{P}_\theta(x_{ki}=0)-\mathrm{P}_\theta(x_{lj}=0)+$$

$$+\mathrm{P}_\theta(x_{ki}=0,x_{lj}=0)=1-(1-\pi_k)^i-(1-\pi_l)^j+(1-\pi_k-\pi_l+\pi_{kl})^i(1-\pi_l)^{j-i} \qquad (6)$$

for any $j\geq i=1,\mathrm{K}\,,n$ and $k,l=1,\mathrm{K}\,,K$. Accordingly, by using (5) and (6) to derive expectations, variances and covariances of (4), it follows that $E_\theta(\mathbf{g}_n)=\boldsymbol{\Gamma}_n$, while the design-based variance-covariance matrix $\mathrm{V}_\theta(\mathbf{g}_n)$ turns out to be a matrix in which the *ij* element is given by

$$C_\theta(g_{in},g_{jn})=\sum_{k=1}^{K}\sum_{l=1}^{K}\left\{(1-\pi_k-\pi_l+\pi_{kl})^i(1-\pi_l)^{j-i}-(1-\pi_k)^i(1-\pi_l)^j\right\}\quad,\ j\geq i=1,\mathrm{K}\,,n \qquad (7)$$

Obviously, the order in which samples are added affects the shape of the resulting curve. Practically speaking, for a given set of presence-absence data, there are $n!$ possible accumulation curves. To overcome this problem, an order-free curve should be adopted. Given a collection of *n* samples, the *rarefaction curve* is the plot of $\overline{g}_{in}$ against i $(i=1,\mathrm{K}\,,n)$, where $\overline{g}_{in}$ represents the arithmetic mean of the g_{in}s arising from all the possible $n!$ orderings. From elementary combinatorial considerations, the rarefaction curve is obtained by means of the sample statistic $\overline{\mathbf{g}}_n=\overline{g}_n(\mathbf{x}_n)$, where $\overline{\mathbf{g}}_n=\left[\overline{g}_{1n},\mathrm{K}\,,\overline{g}_{nn}\right]^{\mathrm{T}}$ is an *n*-vector in which

$$\overline{g}_{in}=SO_n-\sum_{k\in S^n}\frac{\binom{n-x_{kn}}{i}}{\binom{n}{i}}\quad,\quad i=1,\mathrm{K}\,,n \qquad (8)$$

Kobayashi (1974, p. 227) cited Shinozaki (1963) as the author deriving (8), but the same expression (or other equivalent forms) was also independently derived by Holthe

(1975), Engen (1976) and later by Smith et al (1979, p.188). Surprisingly, this result was neglected for some time in ecological works, and the computation of rarefaction curves was usually performed by means of Monte-Carlo methods in which a large set of sample orderings were randomly generated from the universe of all the $n!$ orderings. On this topic, Cotelli and Colwell (2001, p.383) state that "*Because sample-based rarefaction curve depends on the spatial distribution of individual as well as the size and placement of samples, it cannot be derived theoretically*". A full 40 years later than Shinozaki (1963), Ugland et al (2003) finally achieved expression (8)!

Note that $\overline{g}_{nn} = SO_n$. Moreover, for some subsequent theoretical developments, it is worth noting that expression (8) can be rewritten as a linear combination of the K random variables $\binom{n-x_{1n}}{i}, \mathrm{K}, \binom{n-x_{Kn}}{i}$, in the sense that (8) is equivalent to

$$\overline{g}_{in} = K - \sum_{k=1}^{K} \frac{\binom{n-x_{kn}}{i}}{\binom{n}{i}} \quad , \quad i = 1, \mathrm{K}, n \tag{9}$$

Finally, as to the variances and covariances among the ordinates of the $n!$ accumulation curves arising from all the possible orderings, their analytical expressions can be straightforwardly derived. Once again, by means of elementary combinatorial considerations, the variance-covariance matrix of the $n!$ possible $\mathbf{g}_n$'s around $\overline{\mathbf{g}}_n$ turns out to be the sample statistic $\mathbf{G}_n = G_n(\mathbf{X}_n)$, in which the ij element is given by

$$g_{ijn} = (SO_n - \overline{g}_{jn}) - (SO_n - \overline{g}_{in})(SO_n - \overline{g}_{jn}) + \sum_{k \neq l \in S^n} \frac{\binom{n - x_{kn} - x_{ln} + x_{kln}}{i}}{\binom{n}{j}\binom{j}{i}} \;,\; j \geq i = 1,\mathrm{K}\,,n \qquad (10)'$$

Note that when $i = j$, expression (10) coincides with the variance expression achieved by Ugland et al (2003, Appendix 1). Moreover, when the the scheme select one plant only, in which case $x_{kln} = 0$ for any $k \neq l$, and for $i = j$, expression (10) coincides with the well-known variance expression of the individual-based accumulation curve, early achieved by Heck et al (1975).

Note that, depending on the order in which samples are added, the accumulation curve $\mathbf{g}_n$, even if is design-unbiased for $\mathbf{\Gamma}_n$, is not a function of the minimal sufficient statistic. Thus, the use of $\mathbf{g}_n$ to estimate $\mathbf{\Gamma}_n$ entails a redundant variability which should be eliminated by the Rao-Blackwell procedure, i.e by considering the expectation of $\mathbf{g}_n$ conditional to the minimal sufficient statistic $\mathbf{f}_n$**.** Accordingly, denote by $\mathrm{P}(\bullet \mid \mathbf{f}_n)$, $\mathrm{E}(\bullet \mid \mathbf{f}_n)$, $\mathrm{V}(\bullet \mid \mathbf{f}_n)$ and $\mathrm{C}(\bullet \mid \mathbf{f}_n)$ probability measure, expectation, variance and covariance conditional to $\mathbf{f}_n$. Obviously, conditional to $\mathbf{f}_n$ there are $n!/(f_{1n}!\mathrm{K}\, f_{Mn}!)$ distinct data matrices (differing for the column order) which are uniformly distributed on $\nu^{-1}(\mathbf{f}_n)$, in such a way that

$$\mathrm{P}(\mathbf{Z}_n \mid \mathbf{f}_n) = \frac{f_{1n}!\mathrm{K}\, f_{Mn}!}{n!} \;,\; \mathbf{Z}_n \in \nu^{-1}(\mathbf{f}_n) \qquad (11)$$

Thus, on the basis of (11) it can be proven that the application of Rao-Blackwell procedure to the sample accumulation curve gives rise to the rarefaction curve.

Proposition 2. $\bar{\mathbf{g}}_n = \mathrm{E}(\mathbf{g}_n \mid \mathbf{f}_n)$.

Proof. In accordance with (10), the expectation of $\mathbf{g}_n$ conditional to $\mathbf{f}_n$ turns out to be

$$\mathrm{E}(\mathbf{g}_n \mid \mathbf{f}_n) = \sum_{\mathbf{Z}_n \in \nu^{-1}(\mathbf{f}_n)} g_n(\mathbf{Z}_n) p(\mathbf{Z}_n \mid \mathbf{f}_n) = \frac{f_{1n}! \mathrm{K}\, f_{Mn}!}{n!} \sum_{\mathbf{Z}_n \in \nu^{-1}(\mathbf{f}_n)} g_n(\mathbf{Z}_n)$$

On the other hand, the rarefaction curve $\bar{\mathbf{g}}_n$ has been defined as the arithmetic mean of the $\mathbf{g}_n$ s arising from all the possible $n!$ orderings and hence it can be written as

$$\bar{\mathbf{g}}_n = \frac{1}{n!} \sum_{h_1, \mathrm{K}, h_n} g_n(\mathbf{z}_{h_1}, \mathrm{K}, \mathbf{z}_{h_n})$$

where the second summand is extended to the $n!$ permutations of the sample data. But for any distinct matrix in $\nu^{-1}(\mathbf{f}_n)$ there are $f_{1n}! \mathrm{K}\, f_{Mn}!$ permutations of its columns giving rise to the same data matrix. Accordingly,

$$\frac{1}{n!} \sum_{h_1, \mathrm{K}, h_n} g_n(\mathbf{z}_{h_1}, \mathrm{K}, \mathbf{z}_{h_n}) = \frac{f_{1n}! \mathrm{K}\, f_{Mn}!}{n!} \sum_{\mathbf{Z}_n \in \nu^{-1}(\mathbf{f}_n)} g_n(\mathbf{Z}_n) \qquad \blacksquare$$

In an very similar way, *mutatis mutandi*, it can also be proven that

$$\mathbf{G}_n = \mathrm{V}(\mathbf{g}_n \mid \mathbf{f}_n)$$

In accordance with Proposition 2, $\bar{\mathbf{g}}_n$ constitutes a design-unbiased estimator of $\mathbf{\Gamma}_n$ with design-based variance-covariance matrix $\mathrm{V}_\theta(\bar{\mathbf{g}}_n)$ such that $\mathrm{V}_\theta(\mathbf{g}_n) - \mathrm{V}_\theta(\bar{\mathbf{g}}_n) \phi\, 0$. Indeed, from the Rao-Blackwell theorem, the design-based variance-covariance matrix of $\mathbf{g}_n$ can be written as

$$\mathrm{V}_\theta(\mathbf{g}_n) = \mathrm{E}_\theta\{\mathrm{V}(\mathbf{g}_n \mid \mathbf{f}_n)\} + \mathrm{V}_\theta\{\mathrm{E}(\mathbf{g}_n \mid \mathbf{f})\} = \mathrm{E}_\theta(\mathbf{G}_n) + \mathrm{V}_\theta(\bar{\mathbf{g}}_n) \qquad (12)$$

Practically speaking, relation (12) states that the variability of $\mathbf{g}_n$ is given by the design-based variability of $\bar{\mathbf{g}}_n$ plus the design-based expectation of the variance-covariance

matrix $\mathbf{G}_n$, which simply represents the redundant variability of the *n*! order-dependent accumulation curves around the rarefaction curve $\bar{\mathbf{g}}_{in}$. Accordingly, since

$$\mathrm{V}_\theta(\bar{\mathbf{g}}_n) = \mathrm{V}_\theta(\mathbf{g}_n) - \mathrm{E}_\theta(\mathbf{G}_n) \tag{13}$$

a better estimate of $\boldsymbol{\Gamma}_n$ is always achieved by using $\bar{\mathbf{g}}_n$ instead of $\mathbf{g}_n$, while $\mathrm{E}_\theta\{\mathbf{G}_n\}$ determines the efficiency of $\bar{\mathbf{g}}_n$ with respect to $\mathbf{g}_n$. Moreover, since $\bar{\mathbf{g}}_n$ is a function of the complete statistic $\mathbf{x}_n$, it constitutes the minimum variance unbiased estimator of $\boldsymbol{\Gamma}_n$ based on $\mathbf{x}_n$ (for some $\boldsymbol{\theta} \in \Theta$ a better estimator based on $\mathbf{f}_n$ cannot be excluded).

Note also that $\mathbf{x}_n$ is the minimal sufficient statistic for the restricted model $\boldsymbol{\theta} \in \Theta_0$ in which the x_{kn}'s are independent (see e.g. Lehmann and Casella, 1998, Example 6.26, p.44); thus, in this case, $\bar{\mathbf{g}}_n$ costitute the minimum variance unbiased estimator for $\boldsymbol{\Gamma}_n$. Unfortunately, the last result is of no practical relevance since species selections are usually by far from being independent events. A similar result about the optimality of rarefaction curves is obtained in a design-based approach by Smith and Grassle (1977) at the cost of assuming that, like balls in an urn, *n* plants are selected from the plant community by means of simple random sampling with replacement (see section 3). Indeed, in this case $\mathbf{x}_n$ is multinomial with parameters *n* and $\mathbf{p} = [p_1, \mathrm{K}\ , p_K]^{\mathrm{T}}$ and $\mathbf{x}_n$ is the minimal sufficient statistic for **p**. Recently, Mao et al. (2005) achieve a similar optimality result in a model-based approach, at the cost of the unrealistic assumption that the x_{kn}'s are *iid* from a finite mixture of binomial random variables (see section 8)

5. Design-based properties of rarefaction curves.

Since $x_{kn} \sim Bi(n, \pi_k)$, it immediately follows that

$$\mathrm{E}_\theta\left\{\frac{\binom{n-x_{kn}}{i}}{\binom{n}{i}}\right\}=(1-\pi_k)^i\ ,\ k=1,\mathrm{K}\ ,K \tag{14}$$

By using (14) to determine the expectation of (9), it is trivial to show that $\overline{\mathbf{g}}_n$ is a design-unbiased estimator of $\mathbf{\Gamma}_n$, even if the results has already be proven in the previous section as a consequence of Proposition 2.

Moreover, as to the variance-covariance matrix of $\overline{\mathbf{g}}_n$, denote by $G(s;n,i,j)$ the probability generating function of an hypergeometring distribution arising by a without-replacement random selection of *j* balls from an urn of *n* balls partitioned into two groups of *i* and *n-i* balls, respectively. Then, the following result can be proved.

Proposition 3. For any $j\geq i=1,\mathrm{K}\ ,n$ and $k,l=1,\mathrm{K}\ ,K$,

$$\mathrm{E}_\theta\left\{\frac{\binom{n-x_{kn}}{i}\binom{n-x_{ln}}{j}}{\binom{n}{i}\binom{n}{j}}\right\}=(1-\pi_k)^i(1-\pi_l)^j G(r_{kl};n,i,j) \tag{15}$$

where $r_{kl}=\dfrac{1-\pi_k-\pi_l+\pi_{kl}}{1-\pi_k-\pi_l+\pi_k\pi_l}$.

Proof. From the joint distribution of (z_k,z_l), the probability generating function of $(1-z_k,1-z_l)$ turns out to be

$$\mathrm{E}_\theta(s_k^{1-z_k}s_l^{1-z_l})=\pi_{kl}+s_k(\pi_l-\pi_{kl})+s_l(\pi_l-\pi_{kl})+s_ks_l(1-\pi_k-\pi_l-\pi_{kl})$$

Thus, for $s_k=1+t_k$ and $s_l=1+t_l$, it follows that

$$\mathrm{E}_\theta\left\{(1+t_k)^{1-z_k}(1+t_l)^{1-z_l}\right\}=1+t_k(1-\pi_k)+t_l(1-\pi_l)+t_kt_l(1-\pi_k-\pi_l-\pi_{kl})$$

Accordingly, from the independence of the $\mathbf{z}_i$'s, the joint probability generating function of $n - x_{kn} = \sum_{i=1}^{n}(1 - z_{ki})$, $n - x_{ln} = \sum_{i=1}^{n}(1 - z_{li})$ at $s_k = 1 + t_k$ and $s_l = 1 + t_l$, turns out to be

$$L(t_k, t_l) = \mathrm{E}_\theta\left\{(1+t_k)^{n-x_{kn}}(1+t_l)^{n-x_{ln}}\right\} = \left\{1 + t_k(1-\pi_k) + t_l(1-\pi_l) + t_k t_l(1-\pi_k-\pi_l-\pi_{kl})\right\}^n \tag{16}$$

Moreover, it is at once apparent that

$$\frac{\partial^{i+j}}{\partial^i t_k \partial^j t_l} L(t_k, t_l) = \sum_{x_k=0}^{n-i}\sum_{x_l=0}^{n-j} \frac{(n-x_k)!}{(n-x_k-i)!}\frac{(n-x_l)!}{(n-x_l-j)!}(1+t_k)^{n-x_k-i}(1+t_l)^{n-x_l-j}\mathrm{P}_\theta(x_k, x_l)$$

from which

$$\frac{1}{i!\,j!}\left[\frac{\partial^{i+j}}{\partial^i t_k \partial^j t_l} L(t_k, t_l)\right]_{t_k=t_l=0} = \mathrm{E}_\theta\left\{\binom{n-x_{kn}}{i}\binom{n-x_{ln}}{j}\right\} \tag{17}$$

Thus, from (16) and (17), it follows that

$$\mathrm{E}_\theta\left\{\binom{n-x_{kn}}{i}\binom{n-x_{ln}}{j}\right\} = (1-\pi_k)^i(1-\pi_l)^j\binom{n}{i}\sum_{x=\max(i+j-n,0)}^{\min(i,j)}\binom{i}{x}\binom{n-i}{j-x}\left(\frac{1-\pi_k-\pi_l+\pi_{kl}}{1-\pi_k-\pi_l+\pi_k\pi_l}\right)^x$$

from which

$$\mathrm{E}_\theta\left\{\frac{\binom{n-x_{kn}}{i}\binom{n-x_{ln}}{j}}{\binom{n}{i}\binom{n}{j}}\right\} = (1-\pi_k)^i(1-\pi_l)^j\sum_{x=\max(i+j-n,0)}^{\min(i,j)}\frac{\binom{i}{x}\binom{n-i}{j-x}}{\binom{n}{j}}\left(\frac{1-\pi_k-\pi_l+\pi_{kl}}{1-\pi_k-\pi_l+\pi_k\pi_l}\right)^x$$

where the fraction of binomial coefficients in the right side constitutes the probability function at x of the hypergeometric distribution having probability generating function $G(s; n, i, j)$ ■

In accordance with (14) and (15),

$$\mathrm{C}_\theta\left\{\frac{\binom{n-x_{kn}}{i}}{\binom{n}{i}},\frac{\binom{n-x_{ln}}{j}}{\binom{n}{j}}\right\}=(1-\pi_k)^i(1-\pi_l)^j\left\{G(r_{kl};n,i,j)-1\right\}$$

Then, from (9), the *ij* element of $\mathrm{V}_\theta(\overline{\mathbf{g}}_n)$ turns out to be

$$\mathrm{C}_\theta(\overline{g}_{in},\overline{g}_{jn})=\sum_{k=1}^{K}\sum_{l=1}^{K}(1-\pi_k)^i(1-\pi_l)^j\left\{G(r_{kl},n,i,j)-1\right\}\ ,\ j\geq i=1,\mathrm{K}\ ,n \qquad (18)$$

It is worth noting that expression (17) is completely determined by the number of replications and by the first- and second order inclusion probabilities of species which, in turn, depend on the sampling scheme adopted to selecte plants together with the spatial distribution of species over the study area. Indeed, species spread over the area tend to have greater inclusion probabilities then the clumped ones, while overlapping species tend to have greater second-order inclusion probabilities than species settled in different locations. As a simple example, consider an artificial community of $N=40$ shrubs spread over a 10×10 square region and apportioned into $K=2$ species in such a way that species 1 (white) has abundance $N_1=32$ and species 2 (grey) has abundance $N_2=8$ (see Figure 2). Moreover, suppose that the shrubs are sampled by means of *n* replicated transects with starting points randomly thrown onto the horizontal side of the square or, alternatively, onto the vertical side. It is at once apparent from Figure 2 that if a transect is randomly selected on the horizontal side of the quadrat, the inclusion probabilities of the two species turn out to be $\theta_1=0.1$ and $\theta_2=0.7$, with second-order inclusion probability $\theta_{12}=0.1$. However, these probabilities change completely if the transect is randomly selected on the vertical side of the quadrat, in which case, $\theta_1=\theta_2=0.4$ and $\theta_{12}=0.4$.

In most situations, plants communities in large areas are composed of a number of coexisting species, some of them overlapping and some others avoiding the same habitats. In order to have some attach to reality, the empirical investigations reported in this paper refer to an artificial community, say $\mathcal{U}(4,25)$, which is constituted by $K = 100$ species partitioned into 4 exhaustive and mutually exclusive groups of 25 nested species, say $\mathcal{U}_g(25)$ for $g = 1,2,3,4$. Thus, a sampling scheme is supposed in which second-order inclusion probabilities vanish for all the couples of species belonging to different group. On the other hand, within each group of nested species, the first-order inclusion probabilities are supposed to decrease geometrically from a maximum of 0.25 to a minimum of 0.00118 with a decreasing factor of 0.8, while the second-order inclusion probabilities are supposed to be $\pi_{kl} = \min(\pi_k, \pi_l)$ for any couple of species $l > k \in \mathcal{U}_g(25)$.

For the the species community $\mathcal{U}(4,25)$, Table 2 reports the values of γ_i as well as the coefficients of variation of g_{iin} and $\bar{g}_{iin}$, say $CV_{in} = \mathrm{V}_\theta^{1/2}(g_{in})/\gamma_i$ and $\overline{CV}_{in} = \mathrm{V}_\theta^{1/2}(\bar{g}_{in})/\gamma_i$ respectively, and the relative efficiencies of $\bar{g}_{iin}$ with respect to g_{iin}, say $EFF_{in} = \mathrm{V}_\theta(g_{in})/\mathrm{V}_\theta(\bar{g}_{in})$, for selected values of $1 \le i \le n$ and $n = 50,100$. For the same values of *i* and *n*, Table 2 also reports the limits of the inner 0.95-probability intervals of $\bar{g}_{iin}$, defined by the 0.025- and 0.975-quantiles, say $\bar{g}_{0.025in}$ and $\bar{g}_{0.975in}$ and their relative widths, say $\overline{RW}_{0.95in} = (\bar{g}_{0.975in} - \bar{g}_{0.025in})/\gamma_i$. Since the probability distribution of rarefaction curves is prohibitive to be determined analitycally, the quantiles $\bar{g}_{0.025in}$ and $\bar{g}_{0.975in}$ were determined empirically by means of 10,000 presence-absence data matrices randomly generated from the sampling scheme supposed for the

species community $\mathcal{L}(4,25)$. More precisely, each data matrix of size $n = 50,100$ was constituted by n independent vectors $\mathbf{z}_1,\mathrm{K},\mathbf{z}_n$ which, in turn, were independently generated by mean of the following algorithm: *i*) a group of species is randomly selected among the 4 groups; *ii*) a random number, say *u*, is generated from the uniform distribution on $(0,0.25)$ and all the species of the selected group having first-order inclusion probability $\pi_k > u$ are included in the sample.

Finally, Table 2 reports the minimum of the correlation coefficients between the *i*-th coordinate and the remaining $n-1$ cordinates, say $\overline{R}_{in} = \min_{j \neq i} \left\{ \mathrm{C}_\theta(\overline{g}_{in}, \overline{g}_{jn}) \mathrm{V}_\theta^{-1/2}(\overline{g}_{in}) \mathrm{V}_\theta^{-1/2}(\overline{g}_{jn}) \right\}$, together with the index, say $j_{\min}$, of the coordinate corresponding to that minimum.

Results of Table 2 show a) the increase in the expected number of species observed as the sampling effort increases (column 2): the first-order inclusion probabilities are such that about one half of the species is expected to be detected by means of 50 replications and a further replication over 50 is expected to provide an increases of about 0.3 species; on the other hand, doubling the replications, about 70% of the species is expected to be detected and a further replication over 100 is expected to provide an increases of about 0.15 species; b) the variability of accumulation and rarefaction curve ordinates (column 3 and 4): while in the accumulation curve the coefficients of variations are greater in the initial part of the curve, in the rarefaction curve the coefficients are greater at the end of the curve; in both the cases the relative variability decreases as n increases; c) the efficiency of rarefaction curve with respect to accumulation curve (column 5): while in the first part of the curve the efficiency of rarefaction curve turns out to be formidabile, little gain is achieved for i close to n,

where the variability due to ordering is negligible; obviously, no gain is achieved for $i = n$, when the variability due to ordering vanishes; d) the pattern of the 0.95-probability inner interval of rarefaction curve ordinates (column 6 and 7): the resulting interval turns out to be very wide around expectations showing a variability which is likely to deteriorate the utility of rarefaction curves in vegetation studies; after 100 replications, the 0.95 inner interval range from 50 to 87 species !; these result are quite different with those obtained by means of some design-based procedures recently appeared in literature (see Table 8); e) the strength of the linear relationship among the coordinates of the curve: the correlation coefficients between a coordinate and the neghboring ones (not reported in the tables) invariably equal one, while the smallest correlation coefficients are invariably greater than 0.5 and are obtained in correspondence with the first or the last coordinate; practically speaking a suitable and well-speced choice of few coordinates would gives rise to the same statistical information arising from the whole curve.

Table 2a. Design-based properties of the rarefaction curve arising from n=50 independent replication of the sampling scheme supposed to survey the plant community $\mathcal{U}(4,25)$

i	γ_i	CV_{in}	$\overline{CV}_{in}$	EFF_{in}	$\overline{g}_{0.025in} - \overline{g}_{0.975in}$	$\overline{RW}_{0.95in}$	$\overline{R}_{in}(j_{\min})$
1	4.98	0.88	0.12	50.00	3.86-6.28	0.49	0.70 (50)
2	9.27	0.65	0.12	26.81	7.18-11.69	0.49	0.72 (50)
3	12.99	0.54	0.13	18.66	10.02-16.41	0.49	0.74 (50)
4	16.24	0.48	0.13	14.36	12.48-20.57	0.50	0.76(50)
5	19.12	0.44	0.13	11.67	14.64-24.31	0.51	0.78 (50)
6	21.67	0.41	0.13	9.81	16.52-27.60	0.51	0.79 (50)
7	23.95	0.38	0.13	8.45	18.21-30.55	0.52	0.81 (50)
8	26.01	0.36	0.13	7.40	19.74-33.22	0.52	0.82 (50)
9	27.89	0.34	0.13	6.58	21.10-35.66	0.52	0.84 (50)
10	29.59	0.33	0.13	5.91	22.32-37.88	0.53	0.85 (50)
15	36.37	0.28	0.14	3.88	26.96-46.93	0.55	0.90 (50)
20	41.27	0.25	0.15	2.86	30.27-53.77	0.57	0.87 (1)
25	45.08	0.23	0.15	2.25	32.67-59.20	0.59	0.84 (1)
30	48.20	0.21	0.16	1.83	34.52-63.81	0.61	0.81 (1)
35	50.83	0.20	0.16	1.54	36.06-67.72	0.62	0.78 (1)
40	53.10	0.19	0.16	1.32	37.19-71.19	0.64	0.75 (1)
45	55.09	0.18	0.17	1.14	38.20-74.27	0.65	0.72 (1)
46	55.46	0.18	0.17	1.11	38.38-74.87	0.66	0.72 (1)
47	55.83	0.18	0.17	1.08	38.55-75.48	0.66	0.71 (1)
48	56.18	0.18	0.17	1.05	38.71-76.03	0.66	0.71 (1)
49	56.53	0.18	0.17	1.03	38.86-76.52	0.67	0.70 (1)
50	56.87	0.17	0.17	1.00	39.00-77.00	0.67	0.70 (1)

Table 2b. Design-based properties of the rarefaction curve arising from n=100 independent replication of the sampling scheme supposed to survey the plant community $\mathcal{U}(4,25)$

i	γ_i	CV_{in}	$\overline{CV}_{in}$	EFF_{in}	$\overline{g}_{0.025in} - \overline{g}_{0.975in}$	$\overline{RW}_{0.95in}$	$\overline{R}_{in}(j_{\min})$
1	4.98	0.88	0.09	100.00	4.16-5.86	0.34	0.59 (100)
2	9.27	0.65	0.09	53.66	7.73-10.92	0.34	0.61 (100)
3	12.99	0.54	0.09	37.41	10.82-15.32	0.35	0.63 (100)
4	16.24	0.48	0.09	28.86	13.52-19.18	0.35	0.65 (100)
5	19.12	0.44	0.09	23.51	15.88-22.62	0.35	0.67 (100)
6	21.67	0.41	0.09	19.80	17.97-25.66	0.36	0.68 (100)
7	23.95	0.38	0.09	17.09	19.83-28.42	0.36	0.70 (100)
8	26.01	0.36	0.09	15.00	21.48-30.89	0.36	0.71 (100)
9	27.89	0.34	0.09	13.36	22.99-33.13	0.36	0.72 (100)
10	29.59	0.33	0.09	12.03	24.34-35.21	0.37	0.74 (100)
15	36.37	0.28	0.10	7.98	29.67-43.53	0.38	0.79 (100)
20	41.27	0.25	0.10	5.94	33.35-49.61	0.39	0.83 (100)
25	45.08	0.23	0.10	4.71	36.06-54.45	0.41	0.86 (1)
30	48.20	0.21	0.11	3.90	38.38-58.40	0.42	0.83 (1)
35	50.83	0.20	0.11	3.32	40.20-61.84	0.43	0.81 (1)
40	53.10	0.19	0.11	2.88	41.73-64.83	0.43	0.78 (1)
45	55.09	0.18	0.11	2.54	43.03-67.50	0.44	0.76 (1)
50	56.87	0.17	0.12	2.27	44.16-69.90	0.45	0.74 (1)
60	59.92	0.16	0.12	1.85	46.07-74.18	0.47	0.71 (1)
70	62.48	0.15	0.12	1.55	47.50-77.77	0.48	0.67 (1)
80	64.68	0.15	0.13	1.33	48.66-80.94	0.50	0.65 (1)
90	66.60	0.14	0.13	1.15	49.64-83.81	0.51	0.62 (1)
95	67.47	0.14	0.13	1.07	49.88-85.28	0.52	0.61 (1)
96	67.64	0.14	0.13	1.06	49.91-85.63	0.53	0.60 (1)
97	67.81	0.14	0.13	1.04	49.94-85.97	0.53	0.60 (1)
98	67.97	0.14	0.13	1.03	49.96-86.32	0.53	0.60 (1)
99	68.13	0.14	0.13	1.01	49.98-86.66	0.54	0.60 (1)
100	68.30	0.13	0.13	1.00	50.00-87.00	0.54	0.59 (1)

6. Design-based asymptotics on rarefaction curves

The number of coordinates in the rarefaction curve obviously increases along with the number of replications in such a way that, for $n \to \infty$, the rarefaction curve becomes a sequence of random variables. It will be shown that this fact precludes the straightforward achievement of asymptotic results on the design-based distribution of the whole vector $\overline{\mathbf{g}}_n$ as the number of replications increases. Rather, asymptotic results can be readily proved only for a fixed sets of coordinates.

To this purpose, for a fixed $h \le n$, denote by $\overline{\mathbf{g}}_{hn}$ the h-vector containing the first h ordinates of the rarefaction curve, where obviously $\overline{\mathbf{g}}_{nn} = \overline{\mathbf{g}}_n$. Moreover, denote by $\overline{\mathbf{z}}_n = \mathbf{x}_n / n$ the mean vector of $\mathbf{z}_1, \mathrm{K}, \mathbf{z}_n$ and by $\mathbf{t}_n = t_n(\overline{\mathbf{z}}_n)$ the sample statistic where $\mathbf{t}_n = [t_{1n}, \mathrm{K}, t_{nn}]^{\mathrm{T}}$ is the n-vector in which

$$t_{in} = SO_n - \sum_{k \in \mathcal{G}^n} (1 - \overline{z}_{kn})^i \quad , \quad i = 1, \mathrm{K}, n \tag{19}$$

Also in this case, denote by $\mathbf{t}_{hn}$ the h-vector containing the first h ordinates of $\mathbf{t}_n$, where $\mathbf{t}_{nn} = \mathbf{t}_n$. The following convergence results hold.

Proposition 4. For any fixed integer h

$$\lim_{n \to \infty} \sqrt{n}\, \mathrm{E}_\theta \left\{ \left| \overline{\mathbf{g}}_{hn} - \mathbf{t}_{hn} \right| \right\} = \mathbf{0} \tag{20}$$

Proof. Since (19) can be rewritten as

$$t_{in} = K - \sum_{k=1}^{K} (1 - \overline{z}_{kn})^i, \quad i = 1, \mathrm{K}, h$$

then, from (5) it follows that

$$\left|\overline{g}_{in} - t_{in}\right| = \left|\sum_{k=1}^{K}\left\{(1-\overline{z}_{kn})^i - \frac{\binom{n-x_{kn}}{i}}{\binom{n}{i}}\right\}\right| \le \sum_{k=1}^{K}\left|(1-\overline{z}_{kn})^i - \frac{\binom{n-x_{kn}}{i}}{\binom{n}{i}}\right|$$

Accordingly, to prove (20) it is sufficient to prove that

$$\lim_{n\to\infty}\sqrt{n}\,\mathrm{E}_\theta\left\{\left|(1-\overline{z}_{kn})^i - \frac{\binom{n-x_{kn}}{i}}{\binom{n}{i}}\right|\right\} = 0$$

for any $i = 1,\mathrm{K}\,,h$ and any $k = 1,\mathrm{K}\,,K$. After some simple computations it follows that

$$\sqrt{n}\left|(1-\overline{z}_{kn})^i - \frac{\binom{n-x_{kn}}{i}}{\binom{n}{i}}\right| = \sqrt{n}\left|(1-\overline{z}_{kn})^i - \frac{\binom{n(1-\overline{z}_{kn})}{i}}{\binom{n}{i}}\right| =$$

$$= \sqrt{n}\left|(1-\overline{z}_{kn})^i - \frac{n(1-\overline{z}_{kn})\{n(1-\overline{z}_{kn})-1\}\mathrm{K}\ \{n(1-\overline{z}_{kn})-i+1\}}{n(n-1)\mathrm{K}\ (n-i+1)}\right| =$$

$$= \sqrt{n}\left|(1-\overline{z}_{kn})^i - \frac{(1-\overline{z}_{kn})\left\{(1-\overline{z}_{kn})-\frac{1}{n}\right\}\mathrm{K}\ \left\{(1-\overline{z}_{kn})-\frac{i-1}{n}\right\}}{\left(1-\frac{1}{n}\right)\mathrm{K}\ \left(1-\frac{i-1}{n}\right)}\right| =$$

$$= \sqrt{n}\left|(1-\overline{z}_{kn})^i - \frac{(1-\overline{z}_{kn})^i - (1-\overline{z}_{kn})^{i-1}\left\{\frac{1}{n}+\mathrm{K}\ +\frac{i-1}{n}\right\} + \frac{1}{n^2}P_{i-2}(1-\overline{z}_{kn})}{1-\left\{\frac{1}{n}+\mathrm{K}\ +\frac{i-1}{n}\right\}+o(n^{-1})}\right| =$$

$$= \sqrt{n}\left|(1-\overline{z}_{kn})^i - \frac{(1-\overline{z}_{kn})^i - (1-\overline{z}_{kn})^{i-1}\frac{i(i-1)}{2n} + \frac{1}{n^2}P_{i-2}(1-\overline{z}_{kn})}{1-\frac{i(i-1)}{2n}+o(n^{-1})}\right| \le$$

$$= \sqrt{n}\left|(1-\bar{z}_{kn})^i - \frac{(1-\bar{z}_{kn})^i - (1-\bar{z}_{kn})^{i-1}\dfrac{i(i-1)}{2n}}{1-\dfrac{i(i-1)}{2n}+o(n^{-1})}\right| + \sqrt{n}\left|\frac{\dfrac{1}{n^2}P_{i-2}(1-\bar{z}_{kn})}{1-\dfrac{i(i-1)}{2n}+o(n^{-1})}\right| =$$

$$= \sqrt{n}(1-\bar{z}_{kn})^{i-1}\left|1-\bar{z}_{kn} - \frac{1-\bar{z}_{kn}-\dfrac{i(i-1)}{2n}}{1-\dfrac{i(i-1)}{2n}+o(n^{-1})}\right| + \frac{\left|P_{i-2}(1-\bar{z}_{kn})\right|}{n^{3/2}\left\{1-\dfrac{i(i-1)}{2n}+o(n^{-1})\right\}} =$$

$$= \sqrt{n}(1-\bar{z}_{kn})^{i-1}\frac{\left|\bar{z}_{kn}\dfrac{i(i-1)}{2n}+(1-\bar{z}_{kn})o(n^{-1})\right|}{1-\dfrac{i(i-1)}{2n}+o(n^{-1})} + \frac{\left|P_{i-2}(1-\bar{z}_{kn})\right|}{n^{3/2}\left\{1-\dfrac{i(i-1)}{2n}+o(n^{-1})\right\}} \le$$

$$\le \sqrt{n}(1-\bar{z}_{kn})^{i-1}\frac{\bar{z}_{kn}\dfrac{i(i-1)}{2n}+(1-\bar{z}_{kn})o(n^{-1})}{1-\dfrac{i(i-1)}{2n}+o(n^{-1})} + \frac{\left|P_{i-2}(1-\bar{z}_{kn})\right|}{n^{3/2}\left\{1-\dfrac{i(i-1)}{2n}+o(n^{-1})\right\}} =$$

$$\le \frac{i(i-1)(1-\bar{z}_{kn})^{i-1}\bar{z}_{kn}}{2\sqrt{n}\left\{1-\dfrac{i(i-1)}{2n}+o(n^{-1})\right\}} + \frac{(1-\bar{z}_{kn})^i}{o(n^{1/2})\left\{1-\dfrac{i(i-1)}{2n}+o(n^{-1})\right\}} +$$

$$+ \frac{\left|P_{i-2}(1-\bar{z}_{kn})\right|}{n^{3/2}\left\{1-\dfrac{i(i-1)}{2n}+o(n^{-1})\right\}}$$

where $P_{i-2}(1-\bar{z}_{kn})$ is a polynomial in $(1-\bar{z}_{kn})$ of degree less than $i-1$. Accordingly

$$\sqrt{n}\mathrm{E}_\theta\left\{\left|(1-\bar{z}_{kn})^i - \frac{\binom{n-x_{kn}}{i}}{\binom{n}{i}}\right|\right\} \le \frac{i(i-1)\mathrm{E}_\theta\left\{(1-\bar{z}_{kn})^{i-1}\bar{z}_{kn}\right\}}{2\sqrt{n}\left\{1-\dfrac{i(i-1)}{2n}+o(n^{-1})\right\}} +$$

$$+ \frac{\mathrm{E}_\theta\left\{(1-\bar{z}_{kn})^i\right\}}{o(n^{1/2})\left\{1-\dfrac{i(i-1)}{2n}+o(n^{-1})\right\}} + \frac{\mathrm{E}_\theta\left\{\left|P_{i-2}(1-\bar{z}_{kn})\right|\right\}}{n^{3/2}\left\{1-\dfrac{i(i-1)}{2n}+o(n^{-1})\right\}}$$

Since $(1-\bar{z}_{kn})^{i-1}\bar{z}_{kn}<1$ and $(1-\bar{z}_{kn})^{i}<1$, their expectations are both smaller than 1. Thus, from the previous inequality it follows that

$$\sqrt{n}\mathrm{E}_\theta\left\{\left|(1-\bar{z}_{kn})^{i}-\frac{\binom{n-x_{kn}}{i}}{\binom{n}{i}}\right|\right\}\le\frac{i(i-1)}{2\sqrt{n}\left\{1-\frac{i(i-1)}{2n}+o(n^{-1})\right\}}+$$

$$+\frac{1}{o(n^{1/2})\left\{1-\frac{i(i-1)}{2n}+o(n^{-1})\right\}}+\frac{P_{\max}}{n^{3/2}\left\{1-\frac{i(i-1)}{2n}+o(n^{-1})\right\}}$$

where $P_{\max}=\max_{0\le\bar{z}_{kn}\le1}\left|P_{i-2}(1-\bar{z}_{kn})\right|$. The last inequality obviously proves (20) ■

Proposition 5. For a fixed integer h

$$\sqrt{n}(\bar{\mathbf{g}}_{hn}-\boldsymbol{\Gamma}_h)\xrightarrow{d}N(\mathbf{0},\boldsymbol{\Omega}_h) \tag{21}$$

where $\boldsymbol{\Omega}_h=\mathbf{A}_h^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{A}_h$ and $\mathbf{A}_h$ is the $h\times K$-matrix in which $a_{ik}=i(1-\pi_k)^{i-1}$ for each $i=1,\mathrm{K},h,\ k=1,\mathrm{K},K$.

Proof. It is at once apparent that

$$\sqrt{n}\,(\bar{\mathbf{g}}_{hn}-\boldsymbol{\Gamma}_h)=\sqrt{n}\,(\bar{\mathbf{g}}_{hn}-\mathbf{t}_{hn})+\sqrt{n}\,(\mathbf{t}_{hn}-\boldsymbol{\Gamma}_h)$$

where, owing to (20), $\sqrt{n}\,(\bar{\mathbf{g}}_{hn}-\mathbf{t}_{hn})\xrightarrow{p}\mathbf{0}$. Moreover, owing to Delta Method, $\sqrt{n}\,(\mathbf{t}_{hn}-\boldsymbol{\Gamma}_h)\xrightarrow{d}N(\mathbf{0},\boldsymbol{\Omega}_h)$. Thus (21) holds from a well-known convergence result ■

As already pointed out, the previous asymptotic results refer to the first h coordinates of the curve, for h fixed, and as such they cannot be extended to the whole vector $\bar{\mathbf{g}}_n$. That convergence to normality cannot be generally proven for the whole curve is at once

apparent from the very simple example of a community constituted by $K = 1$ species with inclusion probability π. Indeed, for any *n*, the last coordinate of the curve (coinciding with the number of species observed) invariably equals 0 with probability $(1-\pi)^n$ and 1 with probability $1-(1-\pi)^n$. Thus, in this case, no normality can be claimed for $\overline{g}_{nn}$ as $n \to \infty$.

7. Design-based confidence bands around rarefaction curves.

Stated the lackness of asymptotic results for the whole rarefaction curve, bootstrap seems to be the only way to achieved a confidence band around the whole curve. In this case, B bootstrap samples, say $\mathbf{Z}^*_{nb} = \left[\mathbf{z}^*_{1b}, \mathbf{z}^*_{2b}, \mathrm{K}\ , \mathbf{z}^*_{nb}\right]$ with $b = 1, \mathrm{K}\ , B$ can be constructed by randomly selecting with replacement *n* observations from $\mathbf{Z}_n$. For each bootstrap sample $\mathbf{Z}^*_{nb}$, the corresponding rarefaction curve $\overline{\mathbf{g}}^*_{nb} = \overline{g}_n(\mathbf{Z}^*_{nb})$ is computed. Then, for each ordinate, the sequence of the ordered bootstrap estimates $\overline{g}^*_{n(1)} \le \mathrm{K} \le \overline{g}^*_{n(B)}$ is considered, in such a way that $\overline{g}^*_{n(b1)}, \overline{g}^*_{n(b2)}$ with $b1 = \left[\frac{\alpha}{2} B\right]$ and $b2 = \left[\left(1 - \frac{\alpha}{2}\right) B\right]$ costitute the lower and upper limits of the bootstrap confidence interval at the nominal coverage $1-\alpha$.

From expression (3) and from the proof of Proposition 2, $\overline{g}_{in}$ can be rewritten as

$$\overline{g}_{in} = \frac{1}{n!} \sum_{h_1, \mathrm{K}, h_n} \sum_{k=1}^{K} I(z_{kh_1} + z_{kh_2} + \mathrm{K} + z_{kh_i} > 0)$$

But since there are $i!(n-i)!$ permutations of the sample data which give rise to the same value of $\sum_{k=1}^{K} I(z_{kh_1} + z_{kh_2} + \mathrm{K} + z_{kh_i} > 0)$, then $\overline{g}_{in}$ reduces to

$$\bar{g}_{in} = \binom{n}{i}^{-1} \sum_{h_1 < \mathrm{K} < h_i} \sum_{k=1}^{K} I(z_{kh_1} + z_{kh_2} + \mathrm{K} + z_{kh_i} > 0) \qquad (28)$$

where now the summand is extended to the possible choices of i observations over n. Unfortunately, even if is apparent that (28) constitutes a U-statistic, the familiar bootstrap consistency theorems for U-statistics (e.g. Bickel and Freedman, 1981) can be applied only for the first h coordinates, with h fixed. Once again, noting can be said about the actual coverage of the bootstrap confidence intervals for the last coordinates of the curve.

The performance of the bootstrap confidence bands was empirically checked on the basis of the 10,000 samples selected from the species community $\mathcal{L}(4,25)$ by means of the procedure detailed in section 3. More precisely, for $n = 50{,}100$ and for any $i = 1,\mathrm{K}\,,n$, the design-based expectation of the lower and upper limits of the 0.95 confidence interval for γ_i, say $E_{0.025in} = \mathrm{E}_\theta(\bar{g}^*_{n(b1)})$ and $E_{0.975in} = \mathrm{E}_\theta(\bar{g}^*_{n(b2)})$, and its actual coverage, say $P_{0.95in} = \mathrm{P}_\theta(\bar{g}^*_{in(b1)} \le \gamma_i \le \bar{g}^*_{in(b2)})$ were empirically determined on the basis of the simulated data matrices. For each simulated data matrix, $B = 1{,}000$ bootstrap samples were resampled, in such a way that the limits of the 0.95 confidence intervals were given by the terms of rank $b1 = 25$ and $b2 = 975$ in the sequence of the ordered bootstrap estimates.

For selected values of $1 \le i \le n$ and $n = 50{,}100$, Table ** reports the expected limits of the confidence intervals, $E_{0.025in}$ and $E_{0.975in}$, together with their expected relative widths, say $ERW_{0.95in} = (E_{0.975in} - E_{0.025in})/\gamma_i$, and the actual probability contents of these intervals, $P_{0.95in}$.

Table 6a. Monte Carlo results on the performance of bootstrap confidence intervals based on n=50 independent replications of the sampling scheme supposed to survey the plant community $\mathcal{U}(4,25)$

i	$E_{0.025in} - E_{0.975in}$	$ERW_{0.95in}$	$P_{0.95in}$
1	3.87-6.23	0.47	0.93
2	7.12-11.47	0.47	0.92
3	9.88-15.94	0.47	0.92
4	12.24-19.79	0.46	0.91
5	14.29-23.13	0.46	0.90
6	16.08-26.07	0.46	0.89
7	17.65-28.67	0.46	0.88
8	19.05-30.89	0.46	0.87
9	20.30-33.05	0.46	0.87
10	21.43-34.91	0.46	0.86
15	25.71-41.99	0.45	0.81
20	28.60-46.69	0.44	0.77
25	30.66-50.00	0.43	0.72
30	32.21-52.41	0.42	0.67
35	33.40-54.20	0.41	0.62
40	34.32-55.52	0.40	0.58
45	35.04-56.47	0.39	0.54
46	35.16-56.62	0.39	0.53
47	35.28-56.76	0.38	0.51
48	35.40-56.89	0.38	0.51
49	35.51-57.00	0.38	0.51
50	35.61-57.11	0.38	0.51

Table 6b. Monte Carlo results on the performance of bootstrap confidence intervals based on n=100 independent replications of the sampling scheme supposed to survey the plant community $\mathcal{U}(4,25)$

i	$E_{0.025in} - E_{0.975in}$	$ERW_{0.95in}$	$P_{0.95in}$
1	4.17-5.86	0.34	0.94
2	7.72-10.86	0.34	0.93
3	10.76-15.17	0.34	0.93
4	13.39-18.91	0.34	0.93
5	15.69-22.19	0.34	0.92
6	17.72-25.10	0.34	0.92
7	19.52-27.70	0.34	0.92
8	21.13-30.02	0.34	0.91
9	22.57-32.13	0.34	0.91
10	23.88-34.04	0.34	0.90
15	28.97-41.53	0.35	0.88
20	32.50-46.80	0.35	0.86
25	35.14-50.76	0.35	0.84
30	37.21-53.87	0.35	0.82
35	38.88-56.39	0.34	0.80
40	40.28-58.46	0.34	0.78
45	41.45-60.19	0.34	0.76
50	42.45-61.65	0.34	0.73
60	44.06-63.97	0.33	0.69
70	45.30-65.67	0.33	0.64
80	46.26-66.92	0.32	0.59
90	47.01-67.82	0.31	0.54
95	47.32-68.15	0.31	0.53
96	47.38-68.21	0.31	0.53
97	47.43-68.27	0.31	0.52
98	47.49-68.32	0.31	0.49
99	47.55-68.37	0.31	0.48
100	47.60-68.42	0.30	0.48

Results from Table ** confirm the theoretical concerns about the use of bootstrap for constructing confidence bands around rarefaction curves. Indeed, bootstrap coinfidence bands provide very poor performance in the final part of the curve, with coverage decreasing from 93% to 51% when n=50 and from 94% to 48% when n=100. Practically speaking, the design-based construction of reliable confidence bands seems to be an unsolved task, since only the first part of the confidence band provides actual coverage levels near the nominal levels.

However, a very practical solution can be attempted from the analysis of the results of Table ** and **, regarding the bootstrap confidence intervals compared with the results of Table 1a and 1b. Indeed, it is at once apparent that the bootstrap intervals are very skewned around the rarefaction curve, with the lower ends of the intervals which turned out to very similar to the ends of the 0.95-probability intervals but with the upper ends which are much smaller than the theoretical ones (see column 4 of Table 3a and 3b compared with column 2 of Table 6a and 6b). Accordingly, since the 0.95 probability interval suggest symmetric confidence bands around the rarefaction curves, the bootstrap confidence intevals can be symmetrized by means of

$$\bar{g}_{in} \pm \max(\bar{g}_{in} - \bar{g}^{*}_{n(b1)}, \bar{g}^{*}_{n(b2)} - \bar{g}_{in}) \qquad (29)$$

which shall be referred to as *symmetrized bootstrap confidence intervals*.

Even if totally heuristic, the use of (29) has proven to be very effective for achievibg coverage very similar to the nominal level. Table 7 reports the expected limits of the symmetrized bootstrap intervals $E_{0.025in}$ and $E_{0.975in}$, their expected relative widths $ERW_{0.95in}$ and the actual probability contents $P_{0.95in}$. The simmetrized bootstrap intervals provide very good performance with coverage very similar to the nominal level of 0.95 for both the sample sizes..

Table 7a. Monte Carlo results on the performance of symmetrized bootstrap confidence intervals based on n=50 independent replications of the sampling scheme supposed to survey the plant community $\mathcal{U}(4,25)$

i	$E_{0.025in} - E_{0.975in}$	$ERW_{0.95in}$	$P_{0.95in}$
1	3.74-6.23	0.50	0.94
2	7.04-11.52	0.48	0.93
3	9.85-16.17	0.49	0.93
4	12.24-20.32	0.50	0.93
5	14.29-24.02	0.51	0.93
6	16.08-27.36	0.52	0.94
7	17.65-30.37	0.53	0.94
8	19.05-33.11	0.54	0.94
9	20.30-35.61	0.55	0.94
10	21.43-37.92	0.56	0.94
15	25.71-47.24	0.59	0.94
20	28.60-54.21	0.62	0.94
25	30.66-59.82	0.65	0.93
30	32.21-64.55	0.67	0.93
35	33.40-68.65	0.69	0.93
40	34.32-72.30	0.72	0.92
45	35.04-75.60	0.74	0.92
46	35.16-76.22	0.74	0.92
47	35.28-76.83	0.74	0.92
48	35.40-77.44	0.75	0.92
49	35.51-78.03	0.75	0.92
50	35.61-78.61	0.76	0.92

Table 7b. Monte Carlo results on the performance of symmetrized bootstrap confidence intervals based on n=100 independent replications of the sampling scheme supposed to survey the plant community $\mathcal{U}(4,25)$

i	$E_{0.025in} - E_{0.975in}$	$ERW_{0.95in}$	$P_{0.95in}$
1	4.11-5.87	0.35	0.94
2	7.67-10.89	0.35	0.94
3	10.74-15.28	0.35	0.94
4	13.38-19.15	0.36	0.94
5	15.69-22.60	0.36	0.94
6	17.72-25.69	0.37	0.95
7	19.52-28.47	0.37	0.95
8	21.13-31.00	0.38	0.95
9	22.57-33.30	0.38	0.95
10	23.88-35.41	0.39	0.95
15	28.97-43.92	0.41	0.95
20	32.50-50.22	0.43	0.95
25	35.14-55.23	0.45	0.95
30	37.21-59.42	0.46	0.94
35	38.88-63.01	0.47	0.94
40	40.28-66.17	0.49	0.94
45	41.45-68.99	0.50	0.94
50	42.45-71.55	0.51	0.94
60	44.06-76.05	0.53	0.94
70	45.30-79.94	0.55	0.93
80	46.26-83.36	0.57	0.93
90	47.01-86.44	0.59	0.93
95	47.32-87.87	0.60	0.93
96	47.38-88.15	0.60	0.93
97	47.43-88.43	0.60	0.93
98	47.49-88.70	0.61	0.93
99	47.55-88.97	0.61	0.93
100	47.60-89.23	0.61	0.93

8. Design-based extrapolation of rarefaction curves.

While $\bar{\mathbf{g}}_{in}$ provides a design-based unbiased estimate of the first n coordinated of the expected accumulation curve, no unbiased estimator of γ_i exists for $i > n$ in a design-

based framework. However, reliable estimates of γ_i for $i > n$ may be of great utility for predicting the gain in the expected number of species observed when the sampling effort increases over the one actually performed. Moreover, for $i \to \infty n,$ the species richness *K* may be estimasted as the curve asymptote.

The procedure is solely based on the unbiasedness of $\overline{\mathbf{g}}_n$ as an estimator of $\mathbf{\Gamma}_n$ as well as on the fact that, owing to (8), the $\overline{V}_{iin}$'s achieve their maxima when species are nested one within each other in such a way that $\pi_{kl} = \min(\pi_k, \pi_l)$. From (2) and from the unbiasedness of $\overline{\mathbf{g}}_n$ it follows that

$$\overline{g}_{in} = K - \sum_{k=1}^{K} (1 - \pi_k)^i + \varepsilon_{in} \quad , \; i = 1,\mathrm{K}\,,n \tag{12}$$

where $\boldsymbol{\varepsilon} = [\varepsilon_1,\mathrm{K}\,,\varepsilon_n]^{\mathrm{T}}$ are is the error vector with $\mathrm{E}_\theta(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathrm{V}_\theta(\boldsymbol{\varepsilon}) = \overline{\mathbf{V}}_n$. As previously pointed out, neither distributional results nor suitable estimates of $\overline{\mathbf{V}}_n$ are available in a design-based framework. Accordingly, relation (12) can be estimate only by means of ordinary least-square. However, for large K , an eccessive number of parameters is involved in (12). Then, a more parsimonious relation of type

$$\overline{g}_{in} = K - \sum_{h=1}^{H} K_h (1 - \phi_h)^i + \varepsilon_{in} \quad , \; i = 1,\mathrm{K}\,,n \tag{13}$$

is adopted, where $H \geq 1$ denotes the number of species groups containing species with the same inclusion probabilities, and $K_h > 0$ denotes the number of species with inclusion probability ϕ_h $(h = 1,\mathrm{K}\,,H)$, with $0 < \phi_1 < \mathrm{K} < \phi_H < 1$ and $K_1 + \mathrm{K} + K_H = K$. Note that (13) constitutes simply a generalization of (12) and reduces to (12) when $H = K$.

Now, for a given H, denote by $\hat{K}$, $\hat{\mathbf{K}} = \left[\hat{K}_1, \mathrm{K}, \hat{K}_H\right]^{\mathrm{T}}$ and $\hat{\mathbf{\Phi}} = \left[\hat{\phi}_1, \mathrm{K}, \hat{\phi}_H\right]$ the OLS estimates of the parameters involved in (13). Then, confidence bands for $\mathbf{\Gamma}_n$ which tend to be conservative can be achieved by generating B presence-absence data matrices, say $\mathbf{Z}^*_{n1}, \mathrm{K}, \mathbf{Z}^*_{nB}$, from a species community constituted by $\hat{K}$ nested species partitioned into H groups of sizes $\hat{K}_1, \mathrm{K}, \hat{K}_H$ with first-order inclusion probabilities $\hat{\phi}_1, \mathrm{K}, \hat{\phi}_H$. In turn, the matrices give rise to B curves, say $\overline{\mathbf{g}}^*_{n1}, \mathrm{K}, \overline{\mathbf{g}}^*_{nB}$ which can be used to determine the confidence bands. Since the nested structure generates the greatest variability around $\mathbf{\Gamma}_n$, the resulting confidence bands tend to have a coverage greater than the nominal level.

As to the numerical procedure adopted for obtaining the OLS estimates, it is worth noting that, for a given $\mathbf{\Phi}$, relation (13) can be rewritten as

$$\overline{\mathbf{g}}_{in} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \text{subject to } \mathbf{1}^{\mathrm{T}}\boldsymbol{\beta} = 0$$

where $\mathbf{X}$ is an $(n+1) \times K$ matrix depending on $\mathbf{\Phi}$, in which $x_{ik} = (1-\phi_k)^i$ $(i = 0,1,\mathrm{K}, n, k = 1,\mathrm{K}, K)$ and $\boldsymbol{\beta} = \left[K, -K_1, \mathrm{K}, -K_H\right]^{\mathrm{T}}$. Thus, if $H < n-1$, the restricted OLS estimate of $\boldsymbol{\beta}$ can be straightforwardly computed together with the corresponding sum of squares, say $SS_H(\mathbf{\Phi})$. Accordingly, starting with an initial vector $\mathbf{\Phi}_0$, a Monte Carlo minimization procedure is performed in which, at any step $t = 1,2,\mathrm{K}$ and for each $h = 1,\mathrm{K}, H$, a pre-fixed number, say r, of random points are independently thrown onto the r intervals of equal size partitioning the range of ϕ_{th}, i.e. the interval $(\phi_{th-1}, \phi_{th+1})$, where $\phi_{t0} = 0$ and $\phi_{tH+1} = 1$. At any improvement of the sum of squares, the OLS estimates are updated until no improvement takes place for a prefixed

number, say *v*, of consecutive steps. Starting with $H = 1$, the *H* value is finally selected as the integer maximizing the adjusted R^2

Confidence bands arising from the proposed procedure were empirically compared with bootstrap confidence bands and with confidence bands arising from the two model-based techniques by Mao et al (2005). To this purpose, 10,000 presence-absence data matrices constituted by $n = 100$ independent replications were randomly selected from the species community $\mathcal{U}(4,25)$ and the four bands were computed. The OLS procedure for estimating the parameters of (13) was performed adopting $r = 100$ intervals and $v = 20$ failures. All the resampling procedures were performed by meas of $B = 1{,}000$ replications. Table 2 reports the empirical 0.95-band around $\mathbf{\Gamma}_n$ together with the expected limits of the four bands and their coverage for selected values of $1 \le i \le n$. Results of the table confirm the inadequacy of the bootstrap technique as i approaches *n* and emphasize the inadequacy of the model-based procedures under the failure of the independence among species selections. On the other hand, the coverage of the design-based procedure turns out to be invariably greater than the nominal level, even if at the cost of moderate enlargements of the band widths. All the FORTRAN routines adopted in the simulation are available by request from the authors.

To this purposes, the design-based procedure described in the previous section can be straightforwardly adapted to extrapolate the rarefaction curve and estimate γ_i for any $i > n$. Indeed, once the OLS estimates $\hat{K}$, $\hat{\mathbf{K}} = \left[\hat{K}_1, \mathrm{K}, \hat{K}_H\right]^{\mathrm{T}}$ and $\hat{\mathbf{\Phi}} = \left\lfloor \hat{\phi}_1, \mathrm{K}, \hat{\phi}_H \right\rfloor$ have been determined, then, on the basis of (13)

$$\hat{\gamma}_i = \hat{K} - \sum_{h=1}^{H} \hat{K}_h (1 - \hat{\phi}_h)^i \quad i = n+1, n+2\mathrm{K} \tag{14}$$

constitutes a very natural estimator of γ_i for $i > n$.

Then, confidence bands around the extrapolated values can be constructed, once again, by means of the B curves $\overline{\mathbf{g}}^*_{n1}, \mathrm{K}, \overline{\mathbf{g}}^*_{nB}$. In this case, for each curve $\overline{\mathbf{g}}^*_{nb}$ ($b = 1, \mathrm{K}, B$), the corresponding OLS estimates $\hat{K}^*_b$, $\hat{\mathbf{K}}^*_b$ and $\hat{\mathbf{\Phi}}^*_b$ are obtained (adopting the same procedure performed to obtain $\hat{K}$, $\hat{\mathbf{K}}$ and $\hat{\mathbf{\Phi}}$). In turn, for any $i > n$, the B parameter estimates give rise to B estimates of γ_i, say $\hat{\gamma}^*_{i1}, \mathrm{K}, \gamma^*_{iB}$, which can be used to determine the confidence band around the extrapolated values. It should be noticed that procedure involves a Monte Carlo minimization for each one of the B samples, and as such it may turn out to be highly computer-intensive,

The performance of (14) as estimator of γ_i for $i > n$ and the properties of the confidence bands constructed around the extrapolated values were empirically checked by means of a simulation study. In order to achieve a sustainable computational time, the simulation was reduced to 1,000 presence-absence data matrices constituted by $n = 100$ independent replications which were randomly selected from the species community $\mathcal{U}(4,25)$. The OLS procedure was performed by means $r = 100$ intervals and $v = 20$ failures once again, while bands were constructed on the basis of $B = 1{,}000$ replications. Table 3 reports the relative bias and the relative mean error of the extrapolated values together with the expected limits of the corresponding band for selected values of $i > n$. The table also reports the performance of the asymptote estimator $\hat{K}$ as an estimator of species richness. All the FORTRAN routines adopted in the simulation are available by request from the authors.

Results of the table show that the design-based extrapolation of the rarefaction curve with the objective of predicting the number of additional species for i not too greater

than n, give rise to quite accurate predictions and reliable confidence bands. On the other hand the estimation of species richness as the asymptote of the curve give rise to a remarkable underestimation owing to the presence in the community $\mathcal{U}(4,25)$ of species with very small inclusion probabilities. Indeed, in these situations, curves approach K very slowly and they seem to level out at a number of species which may be much smaller than the actual asymptote K. As a trivial example, consider an artificial population of plants containing $K = 250$ species and suppose the presence of 100 species with inclusion probabilities equal to 0.4, 100 species with inclusion probabilities equal to 0.1 and 50 rare species with inclusion probability equal to 5×10^{-7}. Then, for $i = 100$ the sampling is able to detect all the species with inclusion probabilities 0.4 and 0.1, while all the rare species are missed. The situation remains practically the same even when the sampling effort is increased to the prohibitive level of 100,000 replicated plots. More than half of the rare species are missed with 1,000,000 plots while an approximately complete species list requires more than 10,000,000 plots!. On this topic, D'Alessandro and Fattorini (2002) have theoretically proven that the presence small inclusion probabilities also deteriorates the performance of some common procedures adopted to estimate K.

8. Model-based vs design-based approaches.

As already pointed out, the inference adopted in this paper is solely based on the independence of replications, a feature which may be easily ensured by the random selection of points, lines or plots, and avoids any assumption on the species community as well as on species detection. Apart from the work by Grassle and Smith (1977), which is of little practical relevance as it is based on simple random sampling (with

replacement) of plants, no further design-based proposals can be found in literature. On the other hand, several cases of model-based inference have appeared in recent years.

Soberon and Llorente (1993) and, subsequently, Diaz-Francés and Gorostiza (2002) consider surveys performed by *n* plots visited sequentially over time, so that resulting accumulation curves are viewed as time series arising from pure birth processes. On the basis of biological arguments, the authors propose several formulations for the birth rate of the process, leading to several equations for the model-based expectation of the accumulation curve. Nakamura and Peraza (1998) and Christen and Nakamura (2000) also consider *n* plots visited sequentially over time. Moreover, the authors suppose that the first-order species inclusion probabilities π_1, K, π_K constitute *K* iid realizations of a random variable on $(0,1)$, and that, conditional on the resulting probabilities the species detections are independent events among replications as well as among the species. In this way, as pointed out by Nakamura and Peraza (1998, p.20) the detection of each species over the *n* occasions is viewed as an independent replication of the same experiment, thus leading from a very simple likelihood of the presence-absence data. Beyond the presence of these unreliable assumptions, which will be discussed later, the major shortcoming of the above mentioned works lies in the fact that inference, a relevant shortcoming of the above mentioned works lies in the fact that inference is based on accumulation rather than rarefaction curves. As a consequence, inference results strictly depend on the order in which the sample effort is accumulated.

More recently, an order-free procedure based on rarefaction curves has been proposed by Mao et al. (2005) (see also Colwell, et al, 2004). The authors adopt a model very similar to the model by Nakamura and Peraza (1998), with the further assumptions that the π_k's are iid realizations of a discrete random variable with finite support

$0 < \phi_1 < \mathrm{K} < \phi_H \leq 1$ and probability function $\rho_1, \mathrm{K} < \rho_H$. In analogy with previous notations, P_{MCC}, E_{MCC}, V_{MCC} and C_{MCC} now denote probability, expectation, variance and covariance induced by the assumptions of the model by Mao,Colwell and Chang. From the independence of the replications, $x_{1n}, \mathrm{K}, x_{Kn}$ are identically distributed as a mixture of binomial random variables $Bi(n, \phi_h)$. Accordingly, the probability function of each x_{kn} is given by

$$P_{xn} = \mathrm{P}_{MCC}(x_{kn} = x) = \binom{n}{x} \sum_{h=1}^{H} \phi_h^x (1-\phi_h)^{n-x} \rho_h \quad , \; x = 0, \mathrm{K}, n \tag{30}$$

while the probability that a species is detected in *i* replications reduces to

$$1 - P_{0i} = 1 - \sum_{h=1}^{H} (1-\phi_h)^i \rho_h$$

in such a way that the model-based expectation of the number of species observed in *i* replications turns out to be

$$\mu_i = K(1 - P_{0i}) = K - K \sum_{h=1}^{H} (1-\phi_h)^i \rho_h \quad , \; i = 1, 2, \mathrm{K} \tag{31}$$

Moreover, from (30) and from the assumption that species are detected independently each others, any species is once again viewed as an independent replication of the same experiment in which the possible outcomes are the number of detections $0, 1, \mathrm{K}, n$. Hence, if k_{xn} denotes the number of species detected *x* times out of *n*, the random vector $[k_{0n}, k_{1n}, \mathrm{K}, k_{nn}]^{\mathrm{T}}$ is multinomial with probability function

$$p(k_{0n}, k_{1n}, \mathrm{K}, k_{nn}) = \frac{K!}{\prod_{x=0}^{n} k_{xn}!} \prod_{x=0}^{n} P_{xn}^{k_{xn}}$$

From these model-based results, the authors prove that the rarefaction curve turns out to be the UMVUE of the first *n* ordinates of the model-expected accumulation curve of

type (31), with model-based variance covariance matrix $\mathrm{V}_m(\overline{\mathbf{g}}_{in})$ whose *ij* element is given by

$$\mathrm{C}_m(\overline{g}_{in},\overline{g}_{jn}) = K\sum_{x=1}^{n}\left\{1-\frac{\binom{n-i}{x}}{\binom{n}{x}}\right\}\left\{1-\frac{\binom{n-j}{x}}{\binom{n}{x}}\right\}P_{xn}-\frac{\mu_i\mu_h}{K}\ ;\quad i,j=1,\mathrm{K}\ ,n$$

Mao et al (2005) propose the matrix $\mathbf{V}_{n(MCC)}$ with *ij* element

$$v_{ijn(MCC)} = \sum_{x=1}^{n}\left\{1-\frac{\binom{n-i}{x}}{\binom{n}{x}}\right\}\left\{1-\frac{\binom{n-j}{x}}{\binom{n}{x}}\right\}k_x-\frac{\overline{g}_{in}\overline{g}_{jn}}{\hat{K}}\ ;\quad i,j=1,\mathrm{K}\ ,n \qquad (32)$$

as a model-based estimator for $\mathrm{V}_m(\overline{\mathbf{g}}_{in})$, where $\hat{K}$ is a sample estimator of the number of species in the community. Estimator (32) is referred by the authors to as the *moment estimator*. Moreover, as the number of species increases, the authors also prove the model-based convergence of $\overline{\mathbf{g}}_{in}$ to normality, in such a way that for a sufficiently large *K*,

$$\overline{g}_{in}\pm z_{1-\alpha/2}v_{iin(MCC)}^{1/2}\ ;\quad i=1,\mathrm{K}\ ,n \qquad (33)$$

is proposed as a confidence band of nominal coverage $1-\alpha$. The procedure is also implemented in the widely-applied EstimateS 8.0 software (Colwell, 2006) by using the following estimator of species richness

$$\hat{K}=\begin{cases} SO+\dfrac{(n-1)k_{1n}^2}{2nk_{2n}} & if\ k_{2n}>0 \\ S0+\dfrac{(n-1)k_{1n}(k_{1n}-1)}{2n(k_{2n}+1)} & if\ k_{2n}=0 \end{cases} \qquad (34)$$

Unfortunately, the model by Mao, Colwell and Chang is still based on the assumption that species detections are independent events, and as such it turns out to be highly

unrealistic. While the independence among the replications of the sampling scheme can be trivially ensured by the random placement of plots, lines or points, the independence among species detections never holds in practice. Indeed, plot-, line- and point-sampling constitute without-replacement schemes which exclude independence among the sampled plants owing to the spatial association of species, which invariably occurs in any community. If well delineated in its practical sense, such an assumption should alarm any ecologist and should sound like an oxymoron for any botanist who is familiar with the concept of *inter-specific association.* Instead, the assumption is introduced in Mao et al (2005, p.434) by means of a probabilistic consideration, simply stating that the x_{kn}s "*arise as a random sample from the binomial mixture*". Obviously, such a sentence is likely to sound obscure to any ecologists not well-trained in statistical modelling.

The design-based performance of confidence intervals of type (33), was empirically checked on the basis of the 10,000 samples selected from the species community $\mathcal{L}(4,25)$ by means of the procedure detailed in section 3 and using expression (34) for estimating *K*.. For $n = 50,100$ and for any $i = 1,\mathrm{K}\ ,n$, the design-based expectation for the lower and upper limits of the 0.95 confidence interval for γ_i, say $E_{0.025in} = \mathrm{E}_\theta(\overline{g}_{in} - 1.96v^{1/2}_{iin(MCC)})$ and $E_{0.975in} = \mathrm{E}_\theta(\overline{g}_{in} + 1.96v^{1/2}_{iin(MCC)})$, as well as its actual coverage, say $P_{0.95in} = \mathrm{P}_\theta(\overline{g}_{in} - 1.96v^{1/2}_{iin(MCC)} \le \gamma_i \le \overline{g}_{in} + 1.96v^{1/2}_{iin(MCC)})$, were empirically derived on the basis of the 10,000 simulated data matrices. The resulting values of these indexes are reported in Table 8 for $n = 50,100$, and selected values of $1 \le i \le n$.

Table 8a. Monte Carlo results on the performance of Mao-Colwell-Chang confidence intervals based on n=50 independent replications of the sampling scheme supposed to survey the plant community $\mathcal{U}(4,25)$

i	$E_{0.025in} - E_{0.975in}$	$ERW_{0.95in}$	$P_{0.95in}$
1	3.69-6.27	0.52	0.96
2	6.98-11.55	0.49	0.95
3	9.93-16.04	0.47	0.93
4	12.58-19.82	0.45	0.91
5	14.97-23.24	0.43	0.89
6	17.15-26.17	0.42	0.88
7	19.13-28.75	0.40	0.86
8	20.94-31.05	0.39	0.85
9	22.62-33.12	0.38	0.83
10	24.16-34.99	0.37	0.81
15	30.44-42.23	0.32	0.73
20	35.12-47.34	0.30	0.67
25	38.80-51.27	0.28	0.62
30	41.81-54.48	0.26	0.58
35	44.31-57.20	0.25	0.55
40	46.44-59.59	0.25	0.52
45	48.26-61.73	0.24	0.50
46	48.59-62.13	0.24	0.50
47	48.92-62.53	0.24	0.50
48	49.23-62.92	0.24	0.50
49	49.53-63.31	0.24	0.49
50	49.82-63.69	0.24	0.49

Table 8b. Monte Carlo results on the performance of Mao-Colwell-Chang confidence intervals based on n=100 independent replications of the sampling scheme supposed to survey the plant community $\mathcal{U}(4,25)$

i	$E_{0.025in} - E_{0.975in}$	$ERW_{0.95in}$	$P_{0.95in}$
1	3.68-6.28	0.52	0.99
2	6.95-11.58	0.50	0.99
3	9.87-16.10	0.48	0.99
4	12.50-19.98	0.46	0.99
5	14.86-23.36	0.44	0.98
6	17.01-26.32	0.43	0.98
7	18.97-28.93	0.42	0.97
8	20.76-31.26	0.41	0.96
9	22.41-33.35	0.39	0.96
10	23.93-35.25	0.38	0.95
15	30.15-42.57	0.34	0.90
20	34.82-47.72	0.31	0.86
25	38.53-51.64	0.29	0.81
30	41.61-54.80	0.27	0.78
35	44.23-57.44	0.26	0.74
40	46.50-59.71	0.25	0.71
45	48.51-61.69	0.24	0.68
50	50.29-63.46	0.23	0.66
60	53.36-66.51	0.22	0.61
70	55.89-69.09	0.21	0.58
80	58.03-71.36	0.21	0.56
90	59.84-73.38	0.20	0.54
95	60.64-74.33	0.20	0.53
96	60.79-74.51	0.20	0.53
97	60.95-74.69	0.20	0.53
98	61.09-74.88	0.20	0.53
99	61.24-75.06	0.20	0.53
100	61.38-75.23	0.20	0.52

As empirically showed in Table 8, the confidence bands of type (33) turn out to be much narrow than the intervals of Table 3 (column 4) which ensure an actual coverage of 95%. Obviously, the spatial aggregation of species is likely to produce an actual

variability in detections and subsequently in the rarefaction curve ordinates which is much higher than that achieved by the authors on the basis of independence among species detections. As a consequence, the Mao-Colwell-Chang confidence intervals provide unsatisfactory performance, mostly at the end of the curve, with coverage decreasing from 96% to 41% when n=50 and from 99% to 52% when n=100.

These consideration are in contrast with Colwell et al. (2004, p.2721), which attempt to show "*by means of simple but definitive examples*" that rarefaction curves are robust with respect to the failure of the assumption that species detections are independent events. It is however apparent hat rarefaction curves constitute descriptive devices which are completely determined by the data, and as such do not require any assumption. Rather, when rarefaction curves are adopted to estimate expected accumulation curves, their actual efiiciency is much worse than that presumed under independence among species detections

9 An application to a case study.

The design based procedure proposed in the previous sections was checked at first on a site of size $30 \times 30\ m^2$ in the South Eneabba Flora Reserve (Western Australia). All the $N = 12{,}877$ plants within the area were censused and mapped and a complete list of $K = 104$ species was compiled. Subsequently, $n = 50$ circular plots of radius $0.5\,m$ were independently thrown onto the area, giving rise to a presence-absence data matrix constituted of 77 rows (the number of species observed) and 50 columns. The OLS interpolation of the resulting rarefaction curve was performed by using $r = 200$ intervals and $v = 50$. Confidence bands around rarefaction curves and extrapolated values were constructed on the basis of $B = 5{,}000$ replications. The fitted model

involved *** groups of $\hat{K} = **$ species with inclusion probabilities $\hat{\phi}_1 = **$, $\hat{\phi}_2 = **$, $\hat{\phi}_3 = **$, $\hat{\phi}_4 = **$ and sizes $\hat{K}_1 = **$, $\hat{K}_2 = **$, $\hat{K}_3 = **$, $\hat{K}_4 = **$. The prediction of the expected number of species observed by means of $n = 100$ plots turned out to be $\hat{\gamma}_{100} = **$ with 0.95 interval: ****, while ***** was the 0.95-interval around the asymptote estimate.

Subsequently, the design-based procedure was adpoted in a sample survey performed in the Poggio all'Olmo Reserve, an area of 423 ha. in the administrative province of Grosseto (Central Italy). The reserve was surveyed by means of $n = 50$ circular plots of size 50 m^2, randomly thrown onto the study area. A total number of 342 species was observed. The OLS interpolation of the rarefaction curve was performed by means of $r = 200$ intervals and $v = 50$ failures. Confidence bands around were constructed on the basis of $B = 5{,}000$ replications. The fitted model involved involved *** groups of $\hat{K} = **$ species with inclusion probabilities $\hat{\phi}_1 = **$, $\hat{\phi}_2 = **$, $\hat{\phi}_3 = **$, $\hat{\phi}_4 = **$ and sizes $\hat{K}_1 = **$, $\hat{K}_2 = **$, $\hat{K}_3 = **$, $\hat{K}_4 = **$. The prediction of the expected number of species observed by means of $n = 100$ plots turned out to be $\hat{\gamma}_{100} = **$ with 0.95 interval: ****, while ***** was the 0.95-interval around the asymptote estimate. The substantial gain of *** species predicted for a sampling effort increasing from 50 to 100 plots stimulated a further sampling investigation of 50 additional random plots. The subsequent survey gave rise to an overall number of 409 species observed, a value which was very near to the value extrapolated on the basis of 50 plots. Further increases in the number of plots were judged unnecessary in that no substantial gains in the number of species observed were predicted by increasing the sampling effort to over 100 plots.

10. Conclusions

When studies involve plant communities many different species, the species list represents the most convenient way to analyse diversity. Since species cannot be sampled directly but rather by sampling plants, the replications of plots or transects may also be adopted for compiling species list. Unfortunately, the resulting lists are proved to be highly incomplete in the presence of species with very low inclusion probabilities.

In this framework, accumulation curves can be adopted as effective tools for assessing if a forest area has been sufficiently sampled , in which case the curves flatten out at their ends, suggesting that there are only a few species to be found.

If the survey has been performed by making n independent replications of a sampling scheme, the rarefaction curve ordinates are proved to be design-unbiased estimator of the first n ordinates of the design-expected accumulation curve with a design-based variance-covariance matrix determined by the first- and second order species inclusion probabilities. Theoretical and empirical results on finite-sample and asymptotic distributions of rarefaction curves are achieved, and the use of symmetrized bootstrap is proposed in order to empirically achieve confidence bands with actual coverage very similar to the nominal level. It is worth noting that the results achieved in the paper are solely based on the independence of the replications, but they do not require assumptions about species detections. Apart from the work by Grassle and Smith (1977), which is of little practical relevance as it is based on unrealistic simple random sampling of plants, the paper is at the moment the sole proposal for performing design-based inference on rarefaction curves. In most related works, inference on rarefaction curve is performed in a model-based framework, at the cost of introducing unreliable assumptions about species detection.

In accordance with these considerations, the design-based approach proposed in the paper seems to be a very reliable way to perform statistical inference on accumulation curves. Indeed, as pointed out by Sarndal et al (1992, p.21) "*Design-based inference is objective, nobody can challenge that the sample was really selected according to the given dampling design. The probability distribution associated with the design is real, not modelled or assumed*".

Aknowledgements

The authors wishes to thank Prof. Luca Pratelli for its helpful suggestions in deriving the theoretical results of the paper.

References

Bickel, P.J. and Freedman, D.A. (1981) Some Asymptotic Theory for the Bootstrap. *Annals of Statistics*, **9**, 1196-1217.

Christen, J.A. and Nakamura, M. (2000) On the analysis of accumulation curves. *Biometrics*, **56**, 748-754.

Colwell, R.K. (2006) *EstimateS 8.0 User's Guide*. *Statistical Estimation of Species Richness and Shared Species from Samples*. http://viceroy.eeb.uconn.edu/Colwell

Colwell, R.K., Mao, C.X. and Chang, J. (2004) Interpolating, extrapolating, and comparing incidence-based species accumulation curves. *Ecology*, **85**, 2717-2727.

De Vries, P.G. (1986) *Sampling Theory for Forest Inventories*. Berlin, Springer-Verlag.

Diaz-Francés, E. and Gorostiza, L.G. (2002) Inference and model comparison for species accumulation functions using approximating pure birth processes. *Journal of Agricoltural, Biological and Environmental Statistics*, **7**, 335-349.

Engen, S. (1976). A note on the estimation of the species-area curve. *Journal du Conseil International pour l'Exploration de la Mer*, **36**, 286-288.

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

Heck, K.L., van Belle, G. and Simberloff, D. (1975) Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. *Ecology*, **56**, 1459-1461.

Holthe, T. (1975) A method for the calculation of ordinate values of the cumulative species-area curve. *Journal du Conseil International pour l'Exploration de la Mer*, **36**, 183-184.

Hurlbert, S.H. (1971) The nonconcept of species diversity: a critique and alternative parameters. *Ecology* , **52,** 577-586.

Kobayashi, S. (1974) The Species-Area Relation I. A Model for Discrete Sampling. *Researches on Population Ecology*, **15**, 223-237.

Lehmann, E.L. and Casella, G. (1998) *Theory of Point Estimation* (2nd ed.). New York, Springer-Verlag.

Lindgren, B.W. (1993) *Statistical Theory* (4th ed.). London, Chapman & Hall.

Mao, C.X., Colwell, R.K. and Chang, J. (2005) Estimating the species accumulation curve using mixtures. *Biometrics*, **61**, 433-441

Nakamura, M. and Peraza, F. (1998) Species accumulation for beta distributed recording probabilities. *Journal of Agricoltural, Biological and Environmental Statistics*, **3**, 17-36.

Overton, W.S. and Stehman, S.V. (1995). The Horvitz-Thompson theorem as a unifying perspective for probability sampling with examples from natural resource sampling. *American Statistician*, **49**, 261-268.

Schreuder, H.T., Gregoire, T.G. and Wood, G. (1993) *Sampling Methods for Multiresources Forest Inventories*. New York, Wiley.

Shao,J. and Tu, D. (1995) *The Jackknife and the Bootstrap*. New York, Springer-Verlag.

Shinozaki, K. (1963) Note on the Species Area Curve. In *Proceedings of the Annual Meeting of the Ecological Society of Japan*, p. 5 (in Japanese).

Smith, W. and Grassle, J.F. (1977) Sampling properties of a family of diversity measures. *Biometrics*, **33**, 283-292.

Smith, W., Grassle, J.F. and Kravitz, D. (1979) Measures of diversity with unbiased estimates. In Grassle, J.F., Patil, G.P., Smith, W. and Taille,C., editors. *Ecological Diversity in Theory and Practice*. Fairland (MD), International Co-operative Publishing House. p 177-191.

Soberon, J. and Llorente, J. (1993) The use of species accumulation functions for the prediction of species richness. *Conservation Biology*, **7**, 480-488.

Thompson, S.K. (1992) *Sampling*. New York: Wiley.

Ugland, K.I., Gray, J.S. and Ellingsen, K.E. (2003) The species-accumulation curve and estimation of species richness. *Journal of Animal Ecology*, **72**, 888-897.