SAMPLING DESIGNS FOR MONITORING ECOLOGICAL DIVERSITY

L. Barabesi
Department of Quantitative Methods, University of Siena, Italy

Keywords: diversity index estimation, species abundance estimation, environmental sampling, sampling design, unit sampling, area sampling.

Contents

Summary

The analysis of ecological diversity on a given study region constitutes a key topic in environmental studies. The quantitative aspects of ecological diversity are usually detected by means of the diversity indexes, which are *ad hoc* functions of the species abundance. Obviously, species abundance is unknown in practice. Hence, in order to obtain inference on diversity indexes, species abundance must be estimated by using a suitable sampling method for natural populations. The present chapter is devoted to species abundance estimation when the design-based approach is considered. Since the design-based approach avoids any assumption on the underlying population, it seems the appropriate setting when dealing with ecological populations which are usually difficult to model. More specifically, this work focuses on two large classes of sampling designs for species abundance inference, *i.e.*, unit and area sampling respectively.

1. Introduction

During recent decades a plethora of writings on the analysis of diversity has appeared in various contexts. A particular aspect of diversity regards the variety of species in an ecological population and its quantification, *i.e.* the measurement of diversity (see *Descriptive measures of ecological diversity*). Diversity indexes are closely linked with species abundance, since they are necessarily suitable transformations of the number of units belonging to each species. Hence, the knowledge of species abundance is the basis for the quantitative analysis of diversity.

The abundance vector must usually be estimated in the field. Even if it is often assumed that the estimation is based on a random sample, this assumption is striking since natural populations are sampled with designs that lend themselves to be performed in practice. In ecological settings, two classes of sampling strategies are usually suitable, *i.e.*, unit sampling or area sampling. Unit sampling is convenient when the population may be idealized as a pattern of units (points or arbitrarily shaped structures) over the study region and the first-order inclusion probabilities are assigned on the basis of an easy-to-quantify characteristic of the units. On the other hand, area sampling is advisable when the study region may be partitioned into sub-regions of known areas in which species abundance is easy to measure. In this case, the target population actually comprises the sub-regions and the first-order inclusion probabilities are assigned proportionally to sub-region sizes. As is shown in the subsequent sections, the unifying perspective for the two sampling classes is provided by the Horvitz-Thompson estimation.

The present chapter is structured as follows. In Section 2 (and relative subsections), the theoretical framework of unit sampling is described and some designs for ecological studies are presented. Moreover, in Section 3 (and relative subsections), area sampling is introduced along with some specific designs. Finally, section 4 contains the description of a two-stage sampling which adopts area sampling in the first stage and unit sampling in the second stage.

2. Unit sampling

Let us consider a well-defined planar study region R of size A and a population of N units over the region. Moreover, let us suppose that s species are present in the study region and that N_l represents the number of units belonging to the *l*-th species $(l = 1, ..., s)$, in such a way that $N = \sum_{l=1}^{s} N_l$. Thus, the abundance vector turns out to be $\mathbf{N} = (N_1, ..., N_s)^{\mathrm{T}}$. If the units are labelled by the first N integers, the population may be identified with the basic set $U = \{1, ..., N\}$. Let $\mathbf{e}_1, ..., \mathbf{e}_s$ denote the standard basis of $\mathbb{R}^S$, *i.e.*, $\mathbf{e}_l$ represents a vector of zeros except for the one in the *l*-th component $(l = 1, ..., s)$. Thus, $\mathbf{y}_1, ..., \mathbf{y}_N$ turn out to be the species marks associated with each unit, *i.e.*, $\mathbf{y}_j = \mathbf{e}_l$ if the *j*-th unit belongs to the *l*-th species $(j = 1, ..., N, \quad l = 1, ..., s)$. Accordingly, the abundance vector $\mathbf{N}$ may be represented as

$$\mathbf{N} = \sum_{j=1}^{N} \mathbf{y}_j \, , \qquad (1)$$

in such a way that the estimation of $\mathbf{N}$ reduces to the estimation of a vector of a population total, the classical problem in the setting of design-based sampling theory.

Let S denote a sample of distinct units selected from the population using a sampling design. A design-unbiased estimator for $\mathbf{N}$ is provided by the Horvitz-Thompson estimator

$$\hat{\mathbf{N}} = \sum_{j \in S} \frac{1}{\pi_j} \mathbf{y}_j \tag{2}$$

where π_j represents the first-order inclusion probability for the *j*-th unit (for an introduction to the Horvitz-Thompson estimation see *e.g*. Hedayat and Sinha, 1991, p.46 and ff.). The use of (2) is obviously advisable if the π_js are easy to quantify in practice and this often happens in many sampling designs commonly used for natural populations.

In order to evaluate the accuracy of (2), a further goal is the estimation of the variance-covariance matrix $\Sigma = \mathrm{Var}(\hat{\mathbf{N}})$. Since Σ is structure-dependent (*i.e.*, it depends on the locations and marks of the population units), its estimation from a unique sample S presents some undesirable features (see *e.g*. Hedayat and Sinha, 1991, p.51 and ff.). However, if n independent replications of the sampling procedure are performed (this is usually the case with unit sampling), the n samples $S_1, \ldots, S_n$ provide independent identically distributed (*iid*) unbiased estimators for $\mathbf{N}$, say $\hat{\mathbf{N}}_1, \ldots, \hat{\mathbf{N}}_n$, each with variance-covariance matrix Σ. Accordingly, a simple estimator for $\mathbf{N}$ is given by

$$\tilde{\mathbf{N}} = \frac{1}{n} \sum_{i=1}^{n} \hat{\mathbf{N}}_i \, , \tag{3}$$

which turns out to be unbiased with variance-covariance $\mathrm{Var}(\tilde{\mathbf{N}}) = n^{-1} \Sigma$. Hence, an unbiased and consistent $(n \to \infty)$ estimator for $\mathrm{Var}(\tilde{\mathbf{N}})$ is given by $\hat{\mathrm{Var}}(\tilde{\mathbf{N}}) = n^{-1} \tilde{\Sigma}$, where

$$\tilde{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (\hat{\mathbf{N}}_i - \tilde{\mathbf{N}})(\hat{\mathbf{N}}_i - \tilde{\mathbf{N}})^{\mathrm{T}} \, .$$

Finally, note that the replicated use of Horvitz-Thompson estimation even ensures the large-sample normality of (3), since as $n \to \infty$ the use of the Central Limit Theorem straightforwardly provides that

$$n^{1/2} (\tilde{\mathbf{N}} - \mathbf{N}) \xrightarrow{d} N_s(\mathbf{0}, \Sigma) \, .$$

On the basis of the findings in the present section, some common unit samplings for environmental studies will be considered in the following subsections and the structure of the relative Horvitz-Thompson estimators will be detailed.

2.1. Sampling by plots

Plot sampling designs are frequently used when the target population may be considered a point pattern over the study region, such as in the case of several plant species in a forest (see *e.g.* Schreuder *et al.* 1993). Rectangular or circular plots of suitable sizes are usually chosen in plot sampling. When rectangular plots of length L and width $2b$ are used, the sampling design may be performed in the following way: (a) select a point at random over the study region and let it be the centre of the plot; (b) choose (possibly at random) a direction θ on $(0,\pi)$, construct a rectangular plot with orientation θ and select the units included in the plot; (c) repeat the procedure n times. It is at once apparent that a unit is included in a plot if the plot centre falls within a rectangle of length L and width $2b$ with orientation θ, having the unit as its centre (see Figure 1).

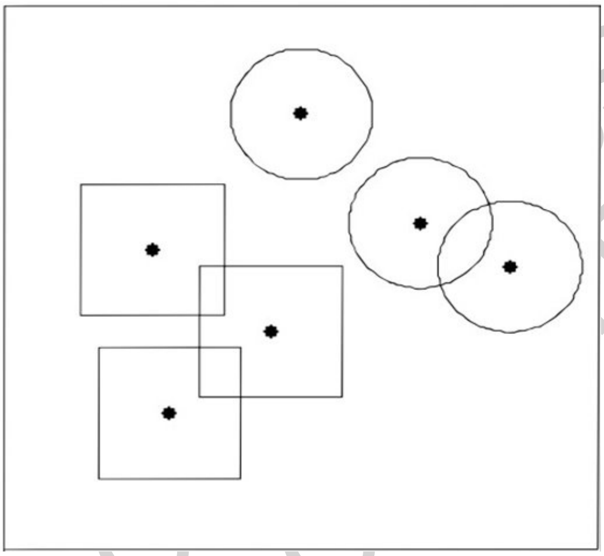

Figure 1. Population units (represented by dots) and relative inclusion regions for rectangular and circular plots respectively.

Therefore, the probability of selecting each unit turns out to be

$$\pi_j = \frac{2Lb}{A}.$$

However, it should be remarked that if a unit is near the edge of the study region, the inclusion rectangle may partially overlap the study region and sampling corrections must be made in order to remove these boundary effects (see Barabesi and Fattorini, 1998a).

Similar results are obtained when circular plots of radius b are used. In this case, the circular plot sampling design reduces to: (a) select a point at random over the study region, let it be the centre of a circle of radius b and select the units included in the plot; (b) repeat the procedure n times. It should be remarked that a unit is included in a plot if the plot centre falls within a circle of radius b, having the unit as its centre (see again Figure 1). Hence, a unit is included in a plot with probability

$$\pi_j = \frac{\pi b^2}{A} .$$

Again, sampling corrections must be made in the case of boundary effects (see Barabesi and Fattorini, 1998a).

Finally, note that both the procedures give rise to n iid unbiased estimators of type (2) which can be combined in the following overall estimators of type (3), in such a way to obtain

$$\tilde{\mathbf{N}} = \frac{A}{2nLb} \sum_{i=1}^{n} \sum_{j \in S_i} \mathbf{y}_j \qquad (4)$$

in the case of rectangular plots, and

$$\tilde{\mathbf{N}} = \frac{A}{n\pi b^2} \sum_{i=1}^{n} \sum_{j \in S_i} \mathbf{y}_j \qquad (5)$$

in the case of circular plots. Obviously, estimators (4) and (5) are simply the sampling vectors of the total species counts in the plots, weighted by factors which depend on the plot size.

-
-
-

Bibliography

Barabesi, L. (2000a). Plant intensity estimation by using point-to-plant sampling: model-based and design-based inference. Proceedings of the XL Meeting of the Italian Statistical Society, Florence (Italy), April 26-28, 2000. [This contains the theoretical basis for design-based inference of diversity indexes when point-to-plant sampling is adopted].

Barabesi, L. (2000b). Local likelihood density estimation in line transect sampling. *Environmetrics* **11**, 413-422. [This provides a semiparametric method for abundance estimation in line transect sampling].

Barabesi, L. and Fattorini, L. (1998a). The use of replicated plot, line and point sampling for estimating species abundance and ecological diversity. *Environmental and Ecological Statistics* **5**, 353-370. [This introduces design-based inference of diversity indexes].

Barabesi L., Fattorini L. and Ridolfi, G. (1998b) Two phase surveys of elusive populations, in *New directions in surveys and censuses*, Symposium 97, Statistics Canada, 285-287. [This introduces the use of two-stage sampling].

Berger, Y. (1998). Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *Journal of Statistical Planning ad Inference* **67**, 209-226. [This provides a theoretical framework for large-sample distribution of Horvitz-Thompson estimators].

Buckland, S.T., Anderson, D.R., Burnham, K.P. and Laake, J.L. (1993). *Distance sampling: estimating abundance of biological populations*. London: Chapman and Hall. [This book comprehensively treats line transect sampling].

Burnham, K.P. and Anderson, D.R. (1976). Mathematical models for nonparametric inferences from line transect data. *Biometrics* **32**, 325-336. [This introduces the key detection model for line transect sampling].

Deville, J.C. and Tillé, Y. (1998) Unequal probability sampling without replacement through a splitting method. *Biometrika* **85**, 89-101. [This paper provides a constructive method for generating $\pi\mathrm{ps}$ sampling].

Hájek, J. (1981). *Sampling from a finite population*. New York: Marcel Dekker. [This is an advanced text on design-based inference].

Hayne, D.W. (1949). An examination of the strip census method for estimating animal populations. *Journal of Wildlife Management* **13**, 145-157. [This introduces a famous detection model for line transect].

Hedayat, A.S. and Sinha, B.K. (1991). *Design and inference in finite population sampling*. New York: Wiley. [This is an introductory text on design-based inference].

Overton, W.S. (1969). Estimating the number of animals in wildlife populations. *Wildlife management techniques* (ed. R.H. Giles), 405-455. Washington: The Wildlife Society. [This is a seminal paper on design-based inference for ecological populations].

Sampford, M.R. (1967). On sampling without replacement with unequal probability of selection. *Biometrika* **54**, 499-513. [This proposes an effective and easy-to-implement $\pi\mathrm{ps}$ design].

Schreuder, H.T., Gregoire T.G. and Wood, G. (1993). *Sampling methods for multiresource forest inventories*. New York: Wiley. [This book presents all the essential aspects of design-based inference in forestry and focuses on plot and Bitterlich sampling].

Scott, A. and Wu, C.F. (1981). On the asymptotic distribution of ratio and regression estimators. *Journal of the American Statistical Association* **76**, 98-102. [This provides some technical results on the large-sample distribution of ratio estimators].

Thompson, S.K. (1992). *Sampling*. New York: Wiley. [This book deals with design-based inference for ecological populations].

Biographical Sketch

Lucio Barabesi is Professor of Statistics at the University of Siena, Italy. His research interests encompass three main areas, environmental sampling, nonparametric density estimation *i.e.*and computational methods in Statistics. He has published numerous articles dealing with these topics on many statistical journals.