Article

Betterments to biodiversity optimal sampling

Alessandro Ferrarini
Department of Evolutionary and Functional Biology, University of Parma, Via G. Saragat 4, I-43100 Parma, Italy
E-mail: sgtpm@libero.it,alessandro.ferrarini@unipr.it

Received 8 August 2012; Accepted 11 September 2012; Published online 1 December 2012
IAEES

Abstract
Biodiversity sampling is pivotal in ecology and biology. It is a complex trade-off between the need to sample ecological info, and the need to do it at low effort. In this paper, I propose an improved new solution to this challenge, which takes into account numerous aspects of biodiversity survey activities. There are two outcomes of the proposed algorithm: a) the optimal number of sampling points, and b) their coordinates in the study area. Resulting sampling points can be used for a survey exactly at that points, or as the centroids of linear and plot (areal) samplings. The proposed solution to biodiversity sampling exactly reverses the common use of ecological and biological sampling: first, one should detect the optimal strategy for his case study, then the kind (random, systematic, stratified) of sampling strategy can be *a posteriori* assessed through proper geo-statistical tests applied to the resulting optimal sampling. In other words, I suggest in this paper that the sampling strategy should be the result of an optimization procedure, not an *a priori* choice. The computational effort for the proposed sampling model is not trivial, nonetheless an optimized sampling strategy is a requirement for successive and successful steps of biodiversity sampling, analysis and preservation.

Keywords biodiversity proxies; ecological-biological sampling; genetic algorithms; optimization; Shannon's evenness index; spatial autocorrelation.

1 Introduction

In field surveys, ecologists and biologists make observations at different spatial locations. In the absence of prior knowledge or in case of outdated info about the species that they intend to sample, they usually rely on 3 kinds of sampling design (random, systematic and stratified; Cochran, 1977; Pielou 1969; Särndal et al., 1992; Scheaffer et al., 1996).

In a simple random sample, each point has the same probability of selection. However, simple random sampling is prone to sampling error because the randomness of the selection may not reflect the spatial pattern of species under study. Systematic and stratified techniques try to subdue this problem. Systematic sampling relies on selecting elements at regular intervals along the study area. However, if some kind of spatial periodicity is present with a period that is a multiple or factor of the employed interval, systematic sampling is likely to be unrepresentative of the overall spatial pattern. When species under study belong to a number of distinct categories, the survey can be organized into separate (stratified) layers. Each layer is then sampled independently, out of which samples can be randomly selected. There are, however, at least two potential drawbacks for stratified sampling. First, it requires the selection of relevant stratification variables, which can be challenging. Second, it's not useful when there are no distinct categories.

Since the results achieved by the previous approaches are often disappointing, in a previous work (Ferrarini, 2012a) I proposed a new solution to biodiversity optimal surveys by coupling information theory (Shannon and Weaver, 1962) and genetic algorithms (Holland, 1975). This algorithmic solution performed very well in two studies of biodiversity survey and conservation of plant and animal species in 101 sites of importance for conservation in the Emilia-Romagna Region (Italy; Ferrarini, 2011; Ferrarini, 2012b). However, during these two studies, I noticed that further improvements to my sampling algorithm were possible. Hence, in this paper I provide major betterments which take into account further aspects of biodiversity sampling activities.

2 The Rationale Behind the Proposed Sampling Algorithm

Conceptually, we can measure the success of biodiversity sampling (S_{bs}) in the form of a benefit-cost function:

$$S_{bs} = \frac{\text{Sampled ecological info}}{\text{Sampling effort}} \quad (1)$$

It's clear that we want to maximize this ratio, hence we are looking for:

$max(S_{bs})$ (2)

The sampling effort (denominator of S_{bs}) can be measured as a function of the number of sampling points and their average geographical distance, as follows:

$Sampling\ effort = s * D_s$ (3)

where s is the number of sampling points and D_s the average distance among sampling points.

With regard to ecological information (e.g., plant and animal species), we start from a situation where we want to sample as much species as possible, but we have just a vague or outdated idea of their spatial arrangement. To this reason, we can make use of common proxies of biodiversity (topographic, land cover and land use variables) which behave as the driving forces of species' spatial arrangement. The wider the range of proxies (e.g. elevation, acclivity, slope aspects, land cover types, soil types, geomorphological types etc.) at sampling points, the higher the chance to sample a wider spectrum of biodiversity (plant and animal species). Shannon's evenness index, calculated for each proxy and then summed up for n proxies, is proficient to the goal of measuring the wideness of sampled biodiversity proxies.

In order to put equation (1) into an operative form, let n be the number of selected proxies of biodiversity, and k the number of intervals chosen for the n variables (Shannon's evenness index requires that variables are split into intervals). Hence, the success of biodiversity sampling S_{bs} can be maximized in the form:

$$\max \frac{\sum_{i=1}^{n}\left(-\frac{\sum_{j=1}^{k} p_{ij} * \ln p_{ij}}{\ln k}\right)}{n * \frac{s}{s_{\max}} * \frac{D_s}{D_{s\max}}} \quad (4)$$

where p_{ij} is the proportion of sampling points that falls in the *j-th* interval of the *i-th* variable, S_{max} (maximum number of sampling points) is a parameter chosen at the beginning of the sampling algorithm, and D_{smax} is the maximum possible distance within the study area (diameter of the study area). There are 2 parameters (n and k that must be chosen at the beginning of the sampling algorithm) and 2 variables (s and the "hidden" variable X-Y coordinates of such s points). Instead, D_s is just a function of the s points and their correspondent coordinates. Both numerator and denominator range in the same $[0 \div n]$ interval. This is pivotal, otherwise the

optimization phase could be biased towards just numerator maximization or denominator minimization, instead of a more general maximimization of S_{bs}.

Last, genetic algorithms (GAs; Goldberg, 1989) can be used in order to solve (4) as a function of *s* and X-Y coordinates of such *s* points (Parolo et al., 2009). GAs consist of optimization procedures based on principles inspired by natural selection. GAs involve "chromosomal" representations of proposed problem solutions which undergo genetic operations such as selection, crossover and mutation. GAs can proceed by generating X-Y coordinates on the surface of the study area and, each time, by recalculating S_{bs} at the search of the highest possible value in the study area.

3 The Improved Solution

First, in equation (4) the denominator could theoretically go close to 0, for instance when D_{smax} is very high. We can fix this problem by adding 1 to both denominator and numerator. Hence, they both range in the same $[1 \div (n+1)]$ interval.

Second, it should be possible to give different weigths to numerator and denominator (α and β respectively). In case of few resources in terms of time and money for sampling activities, we could use $\beta > \alpha$ in the sampling equation (4). The opposite in case the focus is on the ecological aspect of sampling. It's clear that the case $\alpha = \beta$ resembles equation (4).

Third, the number of intervals could be different for each variable. Hence *k* for the *i-th* variable could be better expressed as k_i.

Fourth, in order to avoid spatial autocorrelation (Cliff & Ord, 1973), a minimum distance D_{min} among sampling points should be set. In spatial analyses, the simple count of sample units is not an adequate estimator of effective sample size (Parolo et al., 2008). The amount of pseudoreplication of a variable depends on the distance between sample points (i.e. a set of closely spaced observations effectively provides less information than the same number of observations more widely separated in space). Such spatial dependency is termed spatial autocorrelation and it is often overlooked (Segurado et al., 2006). The choice of D_{min} must take into account several aspects. For instance, the maximum number of sampling points (S_{max}) in the study area would become as follows:

$$S_{\max} = \frac{\text{extension of the study area}}{\pi * D_{\min}{}^2} \qquad (5)$$

because we have to consider a circle of radius equal to D_{min} (where no further sampling points are possible) around each sampling point.

In conclusion, the previous betterments modify equation (4) as follows:

$$S_{bs} = \frac{\alpha * (1 + \sum_{i=1}^{n} \left(- \frac{\sum_{j=1}^{k_i} p_{ij} * \ln p_{ij}}{\ln k_i} \right))}{\beta * (1 + (n * \frac{s}{s_{\max}} * \frac{D_s}{D_{s\max}}))} \qquad (6)$$

that must be maximized using GAs under the constraint that

$$D_{AB} \geq D_{\min} \qquad (7)$$

and

$$S \leq S_{max} \quad (8)$$

where D_{AB} is the geographical distance between two generic sampling points A and B.

Resulting sampling points can be used for a sampling exactly at that points, or as the centroids of linear or plot (areal) samplings.

4 Further Possible Improvements

I conceive that further, minor improvements are possible to the algorithmic solution of equations (6), (7) and (8). With regard to the measurement of sampling effort, a more complex definition could also take into account the distance of sampling points from roads (the more the distance, the more the sampling effort), or the steepness of mountain areas that results very difficult to sample.

With regard to weigths, a deterministic choice at the beginning of the sampling algortihm could decide whether α must be greater β or vice versa. Successively, a stochastic simulation could generate several values for both α and β, and correspondent optimizations of equation (6) via GAs could be verified in order to decide the most appropriate values for α and β.

With regard to the constraints for equation (6), several restraints can be added. For instance, if we are just looking for hygrophytic plant species, a maximum distance for sampling points from rivers and streams could be set. In addition, areas with excessive steepness or altitude a.s.l. could be *a priori* excluded for logistic reasons.

5 Conclusions

Biodiversity sampling is pivotal in ecology and biology. It is a complex trade-off between the need to sample ecological info, and the need to do it at low effort.

In this paper, I proposed an improved solution to this challenge, which take into account numerous aspects of biodiversity sampling activities. The proposed solution to biodiversity sampling exactly reverses the common use of ecological and biological sampling: first, one should detect the optimal strategy for his case study, then the kind (random, systematic, stratified) of sampling strategy can be *a posteriori* assessed through proper geo-statistical tests applied to the resulting optimal sampling. In other words, I have suggested in this paper that the sampling strategy should be the result of an optimization procedure, not an *a priori* choice.

There are two outcomes of my algorithm: a) the number of sampling points, and b) their coordinates in the study area. Resulting sampling points can be used for a survey exactly at that points, or as the centroids of linear and plot (areal) samplings.

The computational effort for this sampling algorithm is not trivial, by the way the sampling strategy is a requirement for successive and successful steps of biodiversity sampling, analysis and preservation.

References

Cliff AD, Ord JK. 1973. Spatial Autocorrelation. Pion Limited, London, UK

Cochran WG. 1977. Sampling Techniques (3rd edition). Wiley, USA

Ferrarini A. 2011. Campionamento della biodiversità animale in 101 aree protette della Regione Emilia-Romagna. Technical Report, Italy

Ferrarini A. 2012a. Biodiversity optimal sampling: an algorithmic solution. Proceedings of the International Academy of Ecology and Environmental Sciences, 2: 50-52

Ferrarini A. 2012b. Campionamento della biodiversità vegetale in 101 aree protette della Regione Emilia-Romagna. Technical Report, Italy

Holland JH. 1975. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence. University of Michigan Press, Ann Arbor, USA

Parolo G, Rossi G, Ferrarini A. 2008. Toward improved species niche modelling. Arnica montana in the Alps as a case study. Journal of Applied Ecology, 45: 1410-1418

Parolo G, Ferrarini A, Rossi G. 2009. Optimization of tourism impacts within protected areas by means of genetic algorithms. Ecological Modelling, 220: 1138-1147

Pielou EC. 1969. An Introduction to Mathematical Ecology. Wiley, USA

Särndal CE, Swensson B, Wretman J. 1992. Model Assisted Survey Sampling. Springer-Verlag, New York, USA

Scheaffer RL, Mendenhal W, Lyman Ott R. 1996. Elementary Survey Sampling (fifth edition). Duxbury Press, CA, USA

Segurado P, Araujo MB, Kunin W.E. 2006. Consequences of spatial autocorrelation for niche-based models. Journal of Applied Ecology, 43: 433-444

Shannon C, Weaver W. 1962. The Mathematical Theory of Communication. University of Illinois Press, Urbana, Illinois, USA