Short Communication

Biodiversity optimal sampling: an algorithmic solution

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 13 November 2011; Accepted 20 December 2011; Published online 5 March 2012
IAEES

Abstract
Biodiversity sampling is a very serious task. When biodiversity sampling is not representative of the biodiversity spatial pattern due to few data or uncorrected sampling point locations, successive analyses, models and simulations are inevitably biased. In this work, I propose a new solution to the problem of biodiversity sampling. The proposed approach is proficient for habitats, plant and animal species, in addition it is able to answer the two pivotal questions of biodiversity sampling: 1) how many sampling points and 2) where are the sampling points.

Keywords ecological-biological sampling; optimization; cost-benefit tradeoff; Shannon's evenness index; genetic algorithms; GIS.

1 Introduction

Solutions on sampling strategy for fitting predictive biodiversity distribution models can be found in relatively few papers (e.g. Guisan and Zimmermann, 2000) or books (e.g. Jongman et al., 1995), and only few of them supply adequate guidelines. A common statement is that a sampling strategy should be based on those gradients that exercise major control over the distribution of species of interest, and these gradients should be used to stratify sampling (Austin and Heyligers, 1991). The main environmental gradients in the study area can be identified in a preliminary exploratory analysis and then used to define a sampling strategy that is especially designed to meet the requirements of the model purposes (Mohler, 1983).

The four strategies most frequently used are: 1) regular sampling, for instance along the two geographic dimensions of a grid covering the study area, 2) random sampling, 3) equal random-stratified sampling, where the study area is first split into environmental strata and an equal number of plots is randomly chosen in each, 4) proportional random-stratified sampling which is similar to the previous one, but the number of plots randomly chosen in each stratum is proportional to its coverage in the study area.

Since the results achieved by the previous approaches are usually disappointing, in this work I propose a new solution to the problem of biodiversity sampling. The proposed approach is proficient for habitats, plant and animal species, in addition it is able to answer the two pivotal questions of biodiversity sampling: 1) how many sampling points and 2) where.

2 Proposed Solution

Conceptually, we can think of biodiversity sampling optimization as the pursuit of the following benefit-cost function maximization:

$$\max \frac{\text{sampled ecological info}}{\text{sampling effort}} \quad (1)$$

where the sampling effort can be easily conceived as a function of the number of sampling points and their average geographical distance, while ecological information (habitats, plant and animal species) could be *a priori* measured using common proxies of biodiversity (topographic, land cover and land use variables). The wider the range of proxies (e.g. elevation, acclivity, slope aspects, land cover types, soil types, geomorphological types and so on) at sampling points the higher the chance to sample a wider spectrum of biodiversity (habitats, plant and animal species). Shannon's evenness index (Shannon and Weaver, 1962), calculated for each variable and summed up for *n* variables, is ideal to the aim of measuring the wideness of biodiversity proxies.

In order to put equation (1) into an algorithmic and operative form, let *n* be the number of proxies that explain the biodiversity spatial pattern (each variable should be represented by a GIS layer; e.g. a layer for elevation above sea level, a further layer for soil types and so forth); *k* the number of intervals chosen for the *n* variables (in fact Shannon's evenness index requires that variables are split into intervals); *s* the number of sampling points; D_s the average distance among sampling points. Hence, the sampling function (*SF*) to be maximized could be written in the form:

$$SF = \max \frac{\sum_{i=1}^{n} \left(-\frac{\sum_{j=1}^{k} p_{ij} * \ln p_{ij}}{\ln k} \right)}{s * D_s} \quad (2)$$

where p_{ij} is the proportion of sampling points that falls in the *j-th* interval of the *i-th* variable. 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.

Since the numerator ranges between 0 (when Shannon's evenness index is 0 for every variable) and *n* (when Shannon's evenness index is 1 for every variable) while the denominator ranges between 0 and *s*Ds*, a normalization is required to keep the denominator in the 0-*n* interval. In fact, if *SF* would be like above, the maximization algorithm would be likely induced to minimize the denominator instead of working on both the numerator and the denominator. I propose to fix this bias using the following final formula:

$$SF = \max \frac{\sum_{i=1}^{n} \left(-\frac{\sum_{j=1}^{k} p_{ij} * \ln p_{ij}}{\ln k} \right)}{n * \frac{s}{s_{\max}} * \frac{D_s}{D_{s\max}}} \quad (3)$$

where 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 (i.e., diameter of the study area). In this way, both numerator and denominator range in the 0-*n* interval.

Last, I propose to apply genetic algorithms (GAs; Holland, 1975; Goldberg, 1989) in order to solve *SF* as a function of *s* and X-Y coordinates of such *s* points. 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 *SF*. To this aim, I suggest that the study area could be partitioned into homogeneous cells (pixels) or segments in case the study area corresponds to a river or a stream, and each cell could be assigned an identification number representing a candidate solution for the optimization GAs process. Each identification number is a gene used by the GAs procedure. Hence, a chromosome would be a vector having a number of genes equal to the amount of optimized sampling points (Parolo et al., 2009). In a *s*-points scenario, each chromosome is hence composed by a string of *s* identification numbers (pixels or segments) that represent a feasible solution to the problem of biodiversity sampling optimization. In order to compute all the previous calculations, I have developed an *ad hoc* module called BOS (Biodiversity Optimal Sampler) for the free GIS GRASS (Neteler and Mitasova, 2008).

3 Conclusions

Biodiversity sampling is a very serious task. When biodiversity sampling is not representative of the biodiversity spatial pattern in the study area due to few data or uncorrected sampling point locations, successive analyses, models and simulations are inevitably biased.

In this paper, I've offered a solution to the problem of optimal biodiversity sampling. I provide consultancy to any research or working groups that decide to apply my biodiversity sampling algorithm to their study areas.

References

Austin MP, Heyligers PC. 1989. Vegetation survey design for conservation: gradsect sampling of forests in north-east New South Wales. Biological Conservation, 50: 13-32

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

Goldberg DE. 1989. Genetic Algorithms in Search Optimization and Machine Learning. Addison-Wesley, Reading, USA

Guisan A, Zimmermann NE. 2000. Predictive habitat distribution models in ecology. Ecological Modelling, 135: 147-186

Jongman RHG, ter Braak CJF, van Tongeren OFR. 1995. Data Analysis in Community and Landscape Ecology. Cambridge University Press, Cambridge, UK

Mohler CL. 1983. Effect of sampling pattern on estimation of species distributions along gradients. Vegetation, 54: 97-102

Neteler M, Mitasova H. 2008. Open Source GIS: A GRASS GIS Approach. Springer, New York, USA

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

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