Stratified random sampling for spatial sampling

Sampling is fundamental to most ecological studies and a representative sampling design is of high importance for biodiversity monitoring. It was previously recommended that the ecological sampling design should be stratified to improve precision, accuracy, and to ensure proper spatial coverage (Gregory et al. 2004). Hence, stratified random sampling has been one of the designs frequently applied in ecological studies. Among the various options for stratified random sampling, Latin Hypercube Sampling (LHS) is promising. It efficiently samples variables from their multivariate distributions and can be conditioned in the multidimensional space defined by environmental covariates, then called conditioned Latin Hypercube Sampling (cLHS) (Minasny and McBratney 2006). The cLHS approach may be used to optimize the sampling design and improve predictions of species distributions by introducing spatial structures of explanatory variables and their cross-spatial structures into the cLHS optimization procedure. This is of special importance as overestimates in species distribution models often result from a lack of relevant explanatory variables or spatial autocorrelation (Lobo and Tognelli 2011). Environmental variables and species distribution data are frequently recorded in different cell (grain) sizes (Lauzeral et al. 2013). Therefore, spatial resolution is critical in any examination of distributions of species (Lauzeral et al. 2013). Reliable methods to downscale environmental variables or species distributions from coarse to fine grain resolutions have potential benefits for ecology and conservation studies (Keil et al. 2013). In regard to spatial resolution, species distribution models (SDMs) are impacted by the fact that environmental descriptors of samples are frequently recorded at different resolutions (Lauzeral et al. 2013) and may thus require scaling to the same resolution. A method called Area-to-Point (ATP) kriging uses spatial structures of predictors for downscaling to predict species distributions (Keil et al. 2013) by taking spatial dependence of predictors into account. We illustrate the approach using Swinhoe’s blue pheasants (Lophura swinhoii) in Taiwan as an example.

Combining spatial downscaling with conditioned Latin hypercube sampling (sdcLHS)

In Latin Hypercube Sampling, one must first decide how many sample points to use, and to remember for each sample point from which row and column the sample point was

Figure 1. (a) Flowchart of the procedure of spatial downscaling conditioned Latin Hypercube Sampling (sdcLHS); (b) Interface of the Windows-based tool of spatial conditioned Latin Hypercube Sampling (scLHS).
taken. Statistically expressed, the sd-cLHS approach will address the following optimization problem: Given \( N \) sample sites with environmental variables \((Z)\), select \( n \) sample sites \((n<<N)\) such that the sampled sites form a Latin hypercube. For \( k \) continuous variables, each component of \( Z \) is divided into \( n \) equally probable strata based on their distributions and \( \chi \) denotes a sub-sample of \( Z \). The steps of the sd-cLHS algorithm, which are based on those in cLHS (Minasny and McBratney 2006), are as follows (Figure 1a):

Step 1. ATP kriging to downscale environmental variables \( Z \) from coarse scale to the fine scale.
Step 2. Division of the quantile distribution of \( Z \) into \( n \) strata; calculation of the quantile distribution for each variable.
Step 3. Selection of \( n \) random samples from \( N \); calculation of the correlation matrix of \( \chi \).
Step 4. Calculation of the objective functions. The overall objective function integrates four different components (objective functions) (for details see Lin et al. 2014). For general applications, the weight assigned to each component in the overall objective function is equal.

Steps 5 to 7. Steps 5-7 are optimization procedures (for details see Minasny and McBratney 2006).
Step 8. Repetition of steps 4 to 7 until either the objective function value falls beyond a given stop criterion or 10,000 iterations are completed.

The spatial conditioned Latin Hypercube Sampling (scLHS) (Lin et al. 2014) is the sampling part (steps 2 to 8) of sd-cLHS (Figure 1a), developed as a Windows-based tool to select optimal sampling sites (Figure 1b).

Illustrative example

We applied the optimal sampling method with a downscaling approach to locate optimal sampling sites at the \( 2 \times 2 \) km scale and to improve the identification of the spatial structure of the distribution of Swinhoe’s blue pheasants (Figure 2) in Taiwan. The distribution of the focal species was estimated by Maximum Entropy (see Maxent; Phillips et al. 2009) based on the existing 803 \( 2 \times 2 \) km samples and separately based on 725 \( 1 \times 1 \) km samples (Figure 3). The estimated distributions were assumed to be the real distribution of the focal species.
species for evaluating our proposed approach. The presence data of the species at certain locations, determined from a set of samples based on presence-absence data, combined with the values of a selected set of environmental variables were used as input for the calculations. The resulting output represented the distribution of maximum entropy among all distributions satisfying the set of constraints (Phillips et al. 2009). The performances of Maximum Entropy were validated by the Kappa and AUC values.

The sample locations at 2 × 2 km and 1 × 1 km resolution were partially clustered due to similar spatial patterns and structures (variograms) of the variation of several environmental parameters (Figure 4). The Kappa value of the Maximum Entropy model was 0.38 and the AUC value was 0.86 in model validations using 401 samples at the 2 × 2 km resolution. The Kappa and AUC values in the Maximum Entropy method were slightly higher when using 362 samples (Kappa= 0.58; AUC= 0.92) at the 1 × 1 km resolution. The predictions with 200, 400 and 600 optimal samples taken from the assumed real distributions showed a consistently high performance, with AUC values of 0.99 and Kappa values of 0.97-1.00 for 1 × 1 km cells and AUC values of 0.98 and Kappa values of 0.96-0.98 for 2 × 2 km cells.

**Concluding remarks**

Incorporating spatial dependency of variables with different resolution into sampling approaches is critical to achieve efficient, unbiased spatial sampling. In the frame of the EU project SCALES, we have tested here an optimal sampling approach using the spatial downsampling sdcLHS based on selected environmental variables without pre-sampled species data, and used a Maximum Entropy approach to show the efficiency of the proposed approach in capturing the distribution of the endemic Swinhoe’s blue pheasant in Taiwan. Our analysis showed that fine scale data yielded accurate presence/absence maps using a subset of presence/absence data that were optimally located. Locations of samples tended to be non-randomly spatially distributed when sample size increased at a coarser cell size. In regards to cost and resource efficiency without the loss of spatial structures (variograms) of focal species, our method with a sufficiently large sample size, 200 optimal samples in this case, performed well in capturing the spatial structure and predicting the spatial distribution of the focal species.

We conclude that the proposed sdcLHS approach considers the statistical distributions and effectively exploits the spatial structures of the selected environmental variables to capture spatial correlations in the original data recorded at various cell sizes. In addition, our approach does not require pre-sampled species data to select spatially unbiased sample locations based on information of parameters collected at various scales.
References


