ECOLOGICAL
MODELLING

Which is the optimal sampling strategy for habitat suitability modelling

Alexandre Hirzel [a,b,*], Antoine Guisan [c]

[a] *Laboratory for Conservation Biology (LBC), Institute of Ecology, University of Lausanne, Bâtiment de biologie, 1015 Lausanne, Switzerland*
[b] *Division of Conservation Biology, Institute of Zoology, University of Bern, 3012 Bern, Switzerland*
[c] *Systematic Botany and Geobotany, Institute of Ecology, University of Lausanne, Bâtiment de biologie, 1015 Lausanne, Switzerland*

Abstract

Designing an efficient sampling strategy is of crucial importance for habitat suitability modelling. This paper compares four such strategies, namely, 'random', 'regular', 'proportional-stratified' and 'equal-stratified'—to investigate (1) how they affect prediction accuracy and (2) how sensitive they are to sample size. In order to compare them, a virtual species approach (Ecol. Model. 145 (2001) 111) in a real landscape, based on reliable data, was chosen. The distribution of the virtual species was sampled 300 times using each of the four strategies in four sample sizes. The sampled data were then fed into a GLM to make two types of prediction: (1) habitat suitability and (2) presence/absence. Comparing the predictions to the known distribution of the virtual species allows model accuracy to be assessed. Habitat suitability predictions were assessed by Pearson's correlation coefficient and presence/absence predictions by Cohen's κ agreement coefficient. The results show the 'regular' and 'equal-stratified' sampling strategies to be the most accurate and most robust. We propose the following characteristics to improve sample design: (1) increase sample size, (2) prefer systematic to random sampling and (3) include environmental information in the design.
© 2002 Elsevier Science B.V. All rights reserved.

Keywords: Sampling design; Logistic model; GLM; Simulations; Virtual species; Bootstrap statistics

1. Introduction

In predictive habitat distribution models (see Guisan and Zimmermann, 2000), statistical methods are used to relate the distribution of a species to the spatial distribution of environmental pre-dictors. This is done in a 'static' or 'empirical' way by assuming that the distribution of the species is, at least within a short time frame, in equilibrium with climate and other factors and has suffered no major perturbation (Guisan and Theurillat, 2000). This contrasts with more dynamic approaches that attempt to model the fate of individuals (see e.g. Lischke et al., 1998 for comparisons of both approaches in plant ecology). Several statistical techniques can be used for this purpose. The choice depends primarily upon the type of response variable modelled. A review of some of

* Corresponding author. Tel.: +41-21-692-4176; fax: +41-21-692-4105
E-mail addresses: alexandre.hirzel@ie-zea.unil.ch (A. Hirzel), antoine.guisan@ie-bsg.unil.ch (A. Guisan).

these techniques can be found in Franklin (1995); Guisan and Zimmermann (2000).

For instance, when the response variable is binary (i.e. presence/absence), a popular approach is to use a generalised linear model (GLM; see Nicholls, 1989, Guisan et al., this issue), a particular case of multiple regression, with binomial distribution and logistic link (or other adequate links like probit). This approach, also called logistic regression, is very popular and was successfully used by e.g. Pereira and Itami (1991); Le Duc et al., (1992); Lenihan (1993); Guisan et al. (1999), as well as in this volume by e.g. Osborne and Suarez Seoane, Miller and Franklin and Moisen and Frescino, and by many authors in Scott et al. (in press) and elsewhere (see e.g. Guisan and Zimmermann, 2000).

Most statistical models, and GLM in particular, rely on strong assumptions about the data (see Guisan et al., this volume), such as their need to be independently and identically distributed (iid). Violation of these postulates may lead to biased model estimations and thus to a lessening of their predictive ability. A critical step, related to these assumptions, is hence to know how to optimally sample the data.

However, discussions on sampling design to be used for fitting predictive habitat distribution models are found in relatively few papers (e.g. Austin and Adomeit, 1991; Austin et al., 1995; Guisan and Zimmermann, 2000) or textbooks (e.g. Moore and Chapman, 1986; Jongman et al., 1995), and only few of them provide satisfactory guidelines in the specific context of single species modelling. A common statement is that, to be efficient, a sampling strategy needs to be based on those gradients that are believed to exercise major control over the distribution of a species, and these gradients should be considered primarily to stratify sampling (Mohler, 1983; Austin and Heyligers, 1989, 1991; Wessels et al., 1998). The main environmental gradients in the study area can be identified in a preliminary exploratory analysis (e.g. Dufrêne and Legendre, 1991; Aspinall and Lees, 1994) and used to define a sampling strategy that is especially designed to meet the requirements of the model objectives (Mohler, 1983).

The four strategies most frequently discussed are: (i) regular sampling, for instance along the two geographic dimensions of a grid covering the study area (i.e. grid sampling), (ii) random sampling, (iii) equal random-stratified sampling, where the study area is first subdivided into environmental strata (e.g. using a Geographic Information System) and an equal number of plots is randomly chosen in each, (iv) 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. A fifth approach, called gradsect (Austin and Heyligers, 1989; 1991), is close to a random-stratified sampling (either equal or proportional depending on its design) but is concentrated within a few geographic transects designed across the main landscape gradients, mainly to reduce study costs (money and time; cost-effective surveys). Wessels et al. (1998) have shown that the gradsect approach proves as effective as a full random-stratified design in large areas, but at reduced cost (money and time).

Le Duc et al. (1992), Guisan et al. (1998) are two examples of studies using a regular grid sampling, although, in the latter case, the choice of the sampling strategy was mainly dictated by data availability. Examples of equal random-stratified sampling are provided in Maggini et al. (in press), Wessels et al. (1998). A case of proportional random-stratified sampling is given in Goedickemeier et al. (1997), Guisan et al. (1999). The study by Goedickemeier and co-workers was again, as in Wessels et al., focussing on assessing plant diversity. We could not find any example of pure random sampling used for building habitat distribution models, but Guisan et al. (1998) used this strategy to sample the evaluation data set (i.e. for testing models). However, many examples also exist of modelling studies based on heterogeneous data (mostly stored in biological data banks; see Margules and Austin, 1994), where no design could be set up prior to the sampling (e.g. Hausser, 1995; Franklin, 1998; Jaberg and Guisan, 2001). This is usually due to the fact that, in such cases, optimal locations of survey sites cannot necessarily be suggested to the observers, especially when they

are volunteers, or when the data set is a mix of various data sources. Such data sets, especially when the number of observations is small, run the risk of being biased, particularly toward sites which are more easily accessible (e.g. avoiding cliffs), or due to their proximity to transportation pathways, or by some unknown 'preference' of the observers (e.g. for sites with rare and thus interesting species).

Finally, very few studies compare the efficiency of using different sampling approaches to reach the same goal (e.g. Austin and Adomeit, 1993). Furthermore, basing such comparisons on real data (as in Wessels et al., 1998) runs the additional risk that the truth is usually unknown, and that the comparison of results could be partly subjective, although this is certainly less of a problem when measuring species richness. Using simulated data is an interesting alternative in many regards (Hirzel et al., 2001).

Austin and Adomeit (1993) use simulations to test different sampling approaches, but in a multi-species design and with the specific aim of evaluating the respective cost of each approach for sampling an optimal number of species (i.e. diversity). Austin et al. (1995) also have recourse to a simulation approach, although they build a single simulated data set in order to compare four different modelling approaches. Hence, they do not compare different sampling designs. Finally, Goedickemeier et al. (1997), Wessels et al. (1998) also discuss various sampling designs, but their statistical analyses and discussion concentrate on the context of biodiversity modelling and assessment, and they do not have recourse to simulated data. Although each of these studies offers valuable insights, they do not answer the specific question of how to choose an optimal sampling strategy to model the distribution of a single species' suitable habitat. Yet in the early 1980s, Mohler showed already that, in practice, a random sampling could lead to truncated response curves for some species if the extremities of the main environmental gradients are under-sampled. Stratifying along these gradients and being particularly scrupulous about sampling the extremities can assure an efficient sampling of these outer limits.

Since large field surveys are rather expensive and hence not frequently conducted, related designs usually focus on sampling as many species as possible in the area (e.g. Austin and Adomeit, 1991). In the case of random-stratified sampling, this is usually carried out by choosing a proportional number of sample points in each stratum. However, there are statistical reasons to question whether or not this method constitutes an optimal strategy when the distribution of individual species is their concern. When this is the case, sampling an equal number of locations in each stratum (as done e.g. by Maggini et al., in press) might well prove more powerful. It is particularly important to answer this question since, as we have seen previously, very few papers in the literature discuss this topic, and most focus on biodiversity assessment and/or modelling.

The aim of this paper is to test different sampling strategies for modelling individual species' distribution, by using simulations of a virtual species in a real landscape (see e.g. Hirzel et al., 2001). A simulation approach is a powerful way to provide quantitative answers to such recurrent questions, since the factual reality is evident. More specifically, this study aims to provide the answers to two main questions:

- Is there an optimal field sampling strategy for building predictive habitat distribution models?
- What is the effect of sample size on the predictive ability of a model?

2. Methods

The procedure used in this study was based on Hirzel et al. (2001). The ecological niche of a virtual species was defined as the combination of its hypothetical response curves to 11 environmental variables. This multivariate niche function was used to generate a true habitat suitability map. A true spatial distribution map of the virtual species, based on the above habitat suitability map, was then generated, by cutting suitability values at a given threshold (0.7 was used here) to obtain a presence–absence map (0/1, binary data).

No noise was added to the suitability values (as e.g. used by Mohler, 1983) in order that the comparisons of the sampling strategies should remain as precise as possible.

Both maps were then sampled according to four designs—namely random, regular, proportionally stratified and equally stratified sampling. This produced four simulated data sets, which provided the values of the environmental predictors as well as the habitat suitability and presence/absence of the virtual species for each location sampled. Four sample sizes (see below) were investigated. These data sets were then used as input into the GLM analyses to compute habitat suitability models. The quality of the models was finally assessed by comparing their predictions to those of the true habitat suitability map.

2.1. Virtual species

The virtual species was the 'equilibrium species' generated in Hirzel et al. (2001). We refer the reader to this paper for a description of the generation procedure. Only the main details will be recapitulated here.

The 11 environmental variables are taken from a real landscape, a square region of 25.6 by 25.6 km located in the Swiss Alps, digitally represented in a GIS (IDRISI32, Eastman, 1999) as raster maps of 256 by 256 cells. The variables were derived from

land-cover, topography, hydrography and road/rail network GIS databases (see Table 1).

The ecological niche of the virtual species was modelled by a weighted sum of three types of response functions: gaussian (median optimum), linear (increasing or decreasing) and truncated linear (increasing to a given threshold value, then constant). This produced a habitat suitability map with a value scaled from zero (poor habitat) to one (ideal habitat) to each cell. In contrast to Hirzel et al. (2001), no stochasticity was added to these values.

The true distribution map for the virtual species was computed by selecting from the latter map all cells with a suitability value higher than 0.7. This map indicated all cells where the virtual species was present.

2.2. Sampling

From the true distribution map, four sampling designs were simulated with four classes of sample sizes ($S \cong 100$, 200, 400 and 800 points) to generate 16 data sets. Due to design constraints, the final size of the data sets differed slightly from one sampling strategy to another. These data sets consisted in tables giving the local value of each environmental variable and the presence/absence of the virtual species for each location sampled. The four sampling strategies were applied as follows.

Table 1
Environmental variables defining the virtual species' ecological niche

Environmental predictor	Response	Weight	S-plus symbol
Forest frequency	Linear (increasing)	6	Forest fq
Elevation	Gaussian	5	Alt
Southern aspect freq.	Linear (increasing)	2	South
Distance to towns	Truncated linear	2	Towns
Distance to forests	Linear (decreasing)	1	Forest di
Slope > 30° frequency	Linear (decreasing)	1	Slope30
Distance to waters	Linear (decreasing)	1	water
Distance to villages	Gaussian	1	Towns
Distance to primary roads	Truncated linear	1	Roads1
Distance to secondary roads	Truncated linear	1	Roads2
Distance to railways	Truncated linear	1	Rails

This table indicates for each predictor (first column) its response function (second column) and its weight (third column); the fourth column gives the symbol used for GLM analyses in S-Plus. These variables were derived from the GEOSTAT and VECTOR200 Swiss databases. Reproduced from Hirzel et al. (in press).

2.2.1. Regular sampling

A point was sampled at each node of a regular grid, whose cell size was adjusted so as to generate a number of points as close as possible to the desired number (Fig. 1A). This approach is often referred to as 'grid sampling'.

2.2.2. Random sampling

The sample points were drawn completely at random across the study area (Fig. 1B).

2.2.3. Equal- and proportional-stratified sampling

Four stratifying variables were selected among the most important for defining the species' ecological niche, and each was split into four classes (see Table 2). The study area was then partitioned by combining these classes to generate 256 (at most) homogenous strata. Two sampling strategies were based on these strata: (1) The

equal-stratified sampling, where an identical number of points is randomly selected in each stratum; the number chosen corresponds as closely as possible to the total number of points in the desired sample size (Fig. 1C). (2) The proportional-stratified sampling where a number of points proportional to its area A_i is drawn in each stratum (Fig. 1D). In the latter case, each stratum sample size N_i was thus computed by

$$N_i = \text{Round}\left(A_i \frac{S}{\sum A_i}\right) \tag{1}$$

Some strata were not represented or were too small. Whenever a stratum sample-size was zero (with $S = 100$), it was pooled with similar small strata to make up an adequate area. The final number of strata was 114.

2.3. Habitat modelling

Although the response is binary rather than following a true binomial distribution, we fitted GLM (see Nicholls, 1989; Guisan et al., this volume) in S-PLUS (MathSoft Inc.) by specifying a binomial distribution and a logistic link function. Guisan et al. (1999) compared this model specification to the alternative of fitting a GLM with a quasi distribution (see Guisan et al., this volume) and variance equal to $\mu(1-\mu)$, thus allowing for an automatic estimation of the scale factor rather than set it to one (the default in binomial models). This showed that both formulations actually provided very similar results. However, in many cases, defining a binomial distribution is simpler.

To compare the performance of models built from the four sampling approaches, we followed the same procedure in each case. All models were fitted on seven predictors set out in decreasing order of importance (as known from defining the virtual species' ecological preferences; Hirzel et al., 2001). Those variables having a normal or truncated–linear response curve were inserted as quadratic predictors.
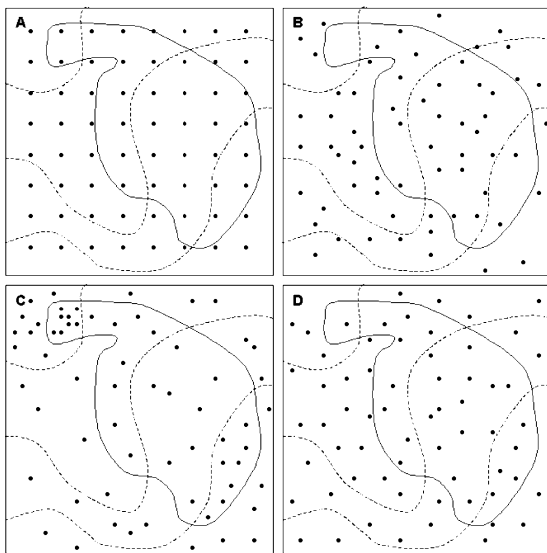

Fig. 1. Four strategies to sample 64 points throughout a landscape defined by two variables: dashed lines represent elevation classes and the plain line circumscribes a forest; these two variables therefore define eight strata. (A) Regular grid sampling (B) random sampling (C) equal-stratified sampling: each stratum contains eight sample points. (D) proportional-stratified strategy: each stratum contains a number of sample points proportional to its area.

Table 2
Each of the four stratifying variables was partitioned into four equal-range classes or strata[a]

Variable	First stratum	Second stratum	Third stratum	Fourth stratum
Forest frequency (%)	[0, 25]	[25, 50]	[50, 75]	[75, 100]
Elevation (m)	[500, 1000]	[1000, 1500]	[1500, 2000]	[2000, 2700]
Aspect (°)	[0, 90]	[90, 180]	[180, 270]	[270, 360]
Distance to towns (km)	[0, 5]	[5, 10]	[10, 15]	[15, 21]

The global stratification was obtained by crossing them to get theoretically $4^4 = 256$ strata.

[a] Actually, some strata were not represented in the study area and some other were too small to be sampled and were therefore pooled. Finally, 114 strata were used in the simulations.

The GLM formula was therefore:

$$\frac{p}{a} = \alpha_1 \, \text{alt} + \alpha_2 \, \text{alt}^2 + \alpha_3 \, \text{south} + \alpha_4 \, \text{forest.fq}$$

$$+ \alpha_5 \, \text{forest.di} + \alpha_6 \, \text{water} + \alpha_7 \, \text{roads2}$$

$$+ \alpha_8 \, \text{roads2}^2 + \alpha_9 \, \text{roads1} + \alpha_{10} \, \text{roads1}^2 \qquad (2)$$

where p/a is the sampled presence/absence of the virtual species and α_i are numerical regression coefficients; the predictors are described in Table 1.

Only the random sampling strategy allowed the sample size to be fixed. The target sample sizes were therefore chosen so as to be as close as possible (cf. Table 3).

2.4. Model evaluation

When working with a virtual species, the fit of the model can obviously be accurately estimated. The true value of habitat suitability is actually known at each location of the study area and its correlation with the modelled value can be as-

Table 3
Comparison of the prediction accuracy resulting from the four sampling strategies and four classes of sample sizes

Sampling	Sample size	Pearson		κ		Errors[b]
		Mean[a]	S.D.[a]	Mean[a]	S.D.[a]	
Regular	110	0.575	0.044	0.686	0.068	13
Equal	114	0.583	0.054	0.69	0.067	6
Proportional	125	0.575	0.051	0.688	0.066	8
Random	116	0.571	0.049	0.67	0.073	10
Regular	247	0.636	0.028	0.780	0.024	4
Equal	228	0.634	0.033	0.777	0.028	4
Proportional	240	0.624	0.032	0.771	0.032	4
Random	240	0.621	0.035	0.765	0.034	7
Regular	440	0.656	0.017	0.805	0.016	0
Equal	456	0.660	0.018	0.806	0.014	0
Proportional	453	0.650	0.021	0.801	0.017	0
Random	450	0.653	0.023	0.801	0.017	0
Regular	782	0.664	0.010	0.819	0.007	0
Equal	798	0.668	0.012	0.817	0.008	0
Proportional	812	0.662	0.014	0.815	0.009	0
Random	810	0.661	0.014	0.814	0.011	0

Habitat suitability prediction accuracy is assessed by Pearson's correlation coefficient; presence/absence prediction accuracy is assessed by κ coefficient of agreement.

[a] Means and S.D. are computed on 300 replicates.

[b] A few replicates caused the GLM algorithm to crash; their number is given in this column.

sessed. Model accuracy was therefore assessed by two measures.

1) True and predicted habitat suitability indices were compared by means of a Pearson's correlation coefficient computed on all map cells.
2) The predicted habitat suitability maps—with values ranging from 0 to 1—were rounded into Boolean maps, thus predicting presence/absence, using the same 0.7 threshold used for defining the true presence–absence map of the virtual species. The accuracy of this classification was then compared with the true distribution map by mean of the κ coefficient of agreement (Cohen, 1960; see Fielding and Bell, 1997).

2.5. Simulations

This sampling-modelling-evaluation process was applied to all 16 sampling situations (4 strategies × 4 sample sizes) and in each case it was replicated 300 times. Each replicate was obtained by drawing a new sample from the whole area, according to the procedures described above; in the case of the regular grid sampling strategy, each replicate was obtained by translating the grid regularly; in this case the replicate number was restricted, for each sample size, respectively, to 289 ($=17^2$), 169 ($=13^2$) and 100 ($=10^2$).

These replicates were used to compare modelling efficiency reached by the four sampling strategies. Average Pearson's and κ coefficients were computed for the 4 × 4 cases. Finally, for each sample size, pairs of sampling strategies were tested by mean of bootstrap statistics (2000 resamplings), corrected by the sequential Bonferroni method (Rice, 1989; base significance level = 0.05).

3. Results

The resulting distributions are summarised graphically in Fig. 2 (habitat suitability prediction accuracy, assessed by Pearson's correlation coefficient) and in Fig. 3 (presence/absence prediction accuracy, assessed by κ agreement coefficient).

Pearson's coefficients range from 0.30 to 0.71. κ Ranges from 0.23 to 0.83. The mean and standard deviations of prediction accuracy (computed on 300 replicates) are presented in Table 3. The results of pair-comparisons are shown in Table 4 (comparisons of mean) and Table 5 (comparisons of S.D.).

The effect of the sample size was to increase the accuracy of model predictions of both habitat suitability and presence/absence. However, the results show that sample size has negligible impact on the ranking of the four strategies: for habitat suitability prediction (accuracy measured by the Pearson's coefficient), the best strategy tends to be 'Equal-stratified', followed by 'Regular', 'Random' and 'Proportional-stratified'; for presence/absence prediction (accuracy measured by κ coefficient), the 'Regular' and 'Equal-stratified' strategies are the best, followed by 'proportional stratified' and 'Random'.

Increasing the sample size had also the effect of reducing the dispersion of model accuracy. The 'Regular' and 'Equal-stratified' samplings tended to be more robust (narrower dispersion) than the 'Random' and 'Proportional-stratified', both in habitat suitability and presence/absence prediction.

Even though statistically significant improvements could be obtained by using one sampling strategy in place of another, the average difference—with a similar sample size—was never larger than a few percent. In contrast, robustness—measured by replicate standard deviation—could often be improved by 30%.

4. Discussion

The goal of this study was to investigate (1) how prediction accuracy depended upon the sampling strategy used and (2) how sensitive it was to sample size. The results show that the most critical parameter of a sampling design is still sample size. Choosing the right sampling strategy can improve the result by a few percent and, more interestingly, reduce the risk of making a poor prediction. This is particularly apparent for presence/absence prediction. The maximum accuracy obtained amongst
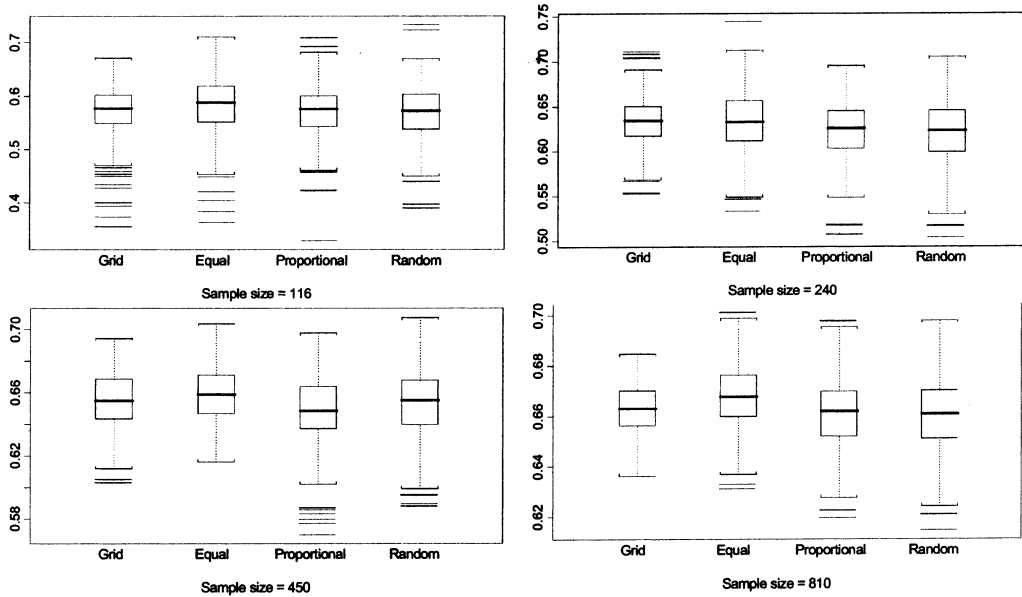
Fig. 2. Effects of the four sampling strategies and four sample sizes on the accuracy of habitat suitability prediction, as measured by Pearson's correlation coefficient. Each box-plot summarises 300 replications of the sampling-modelling procedure. As expected, the larger the sample size, the better the fit of the model. The 'equal-stratified' sampling strategy gives the best results.

all resampling runs is similar for all four strategies, but the 'regular' and 'equal-stratified' strategy never produce such low values as the 'proportional-stratified' and 'random' strategy do in a

significant number of runs. To put it in another way, the range of possible accuracy values (i.e. variability around the mean) increases toward lower values from regular to equal-stratified to
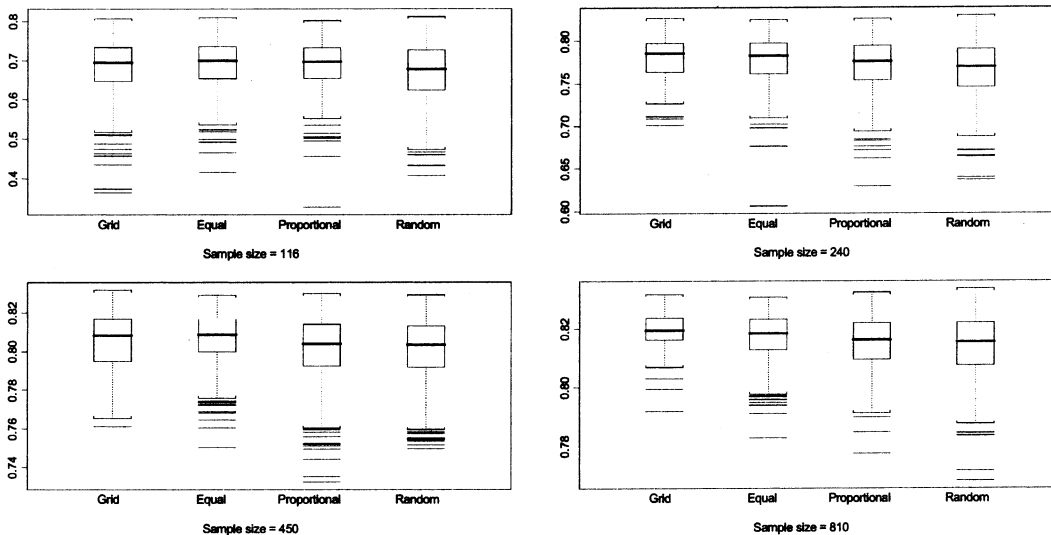

Fig. 3. Effects of the four sampling strategies and four sample sizes on the accuracy of presence/absence prediction, as measured by Cohen's κ agreement coefficient. Each box-plot summarises 300 replications of the sampling-modelling procedure. As expected, the larger the sample size, the better the fit of the model. The 'regular' sampling strategies give the best results.

Table 4
Pair-comparisons of the sampling strategies

Pairs of sampling strategies	Sample size			
	116	240	450	810
Prediction accuracy measure				
Pearson				
Regular/equal	0.012 NS	0.247 NS	0.009*	0.000*
Regular/proportional	0.500 NS	0.000*	0.000*	0.101 NS
Regular/random	0.142 NS	0.000*	0.054 NS	0.016*
Equal/proportional	0.025 NS	0.000*	0.000*	0.000*
Equal/random	0.003*	0.000*	0.000*	0.000*
Proportional/random	0.189 NS	0.144 NS	0.045 NS	0.143 NS
κ				
Regular/equal	0.208 NS	0.083 NS	0.132 NS	0.028 NS
Regular/proportional	0.360 NS	0.000*	0.005*	0.000*
Regular/random	0.000*	0.000*	0.005*	0.000*
Equal/proportional	0.352 NS	0.005*	0.000*	0.000*
Equal/random	0.001*	0.000*	0.000*	0.000*
Proportional/random	0.001*	0.008*	0.438 NS	0.103 NS

The average prediction accuracy values are tested for equality by mean of a 2000-replicates bootstrap. Probability values are then checked for significance with the sequential Bonferroni method with a 0.05 threshold.

random to proportional-stratified sampling (Fig. 3).

This study did not address the problems caused by biased sampling designs (e.g. accessibility, detectability biases), as all four strategies under review were unbiased. The main question thus centered on how the sample points should be distributed along the various environmental

Table 5
Pair-comparisons of the strategies

Pairs of sampling strategies	Sample size			
	116	240	450	810
Prediction accuracy measure				
Pearson				
Regular/equal	0.003*	0.002*	0.282 NS	0.016*
Regular/proportional	0.033 NS	0.014 NS	0.012*	0.000*
Regular/random	0.055 NS	0.000*	0.000*	0.000*
Equal/proportional	0.237 NS	0.214 NS	0.024 NS	0.012*
Equal/random	0.121 NS	0.235 NS	0.000*	0.011*
Proportional/random	0.352 NS	0.074 NS	0.114 NS	0.457 NS
κ				
Regular/equal	0.380 NS	0.014 NS	0.044 NS	0.037 NS
Regular/proportional	0.278 NS	0.000*	0.130 NS	0.001*
Regular/random	0.139 NS	0.000 *	0.149 NS	0.000*
Equal/proportional	0.366 NS	0.109 NS	0.002 *	0.040 NS
Equal/random	0.099 NS	0.014 NS	0.002*	0.000*
Proportional/random	0.073 NS	0.135 NS	0.376 NS	0.006*

The S.D. of prediction accuracy values is tested for equality by means of a 2000-replicates bootstrap. Probability values are then checked for significance with the sequential Bonferroni method with a 0.05 threshold.

gradients—how the 'probes' should be positioned in the landscape—in order to best reflect the area. The investigation of bias impact on modelling accuracy could be the goal of a further analysis.

These results give an interesting insight into the characteristics of a good sampling strategy. Clearly, the 'equal-stratified' and 'regular' designs produce better results than 'random' and 'proportional-stratified'. These four sampling strategies may be classified along two lines: the first is the amount of knowledge needed to implement the strategy; although the 'regular' and 'random' strategies are purely spatial techniques, both 'stratified' strategies need extra information about the environment. The second axis is the way the sample points are drawn; the 'random' and 'proportional-stratified' strategies obey a fairly loose rule of point selection, whereas the 'regular' and 'equal-stratified' strategies are more systematic and impose a restricted choice (Table 6). Actually, with a large sample size, the 'proportional-stratified' strategy tends toward pure randomness. What our results show is that the factors increasing sampling efficiency are, in decreasing order of priority: (1) to increase sample size, (2) to use regular sampling and (3) to use environmental information to stratify sampling.

As sampling is a money- and time-consuming task, prioritisation of requirements is of great importance. In this sense, these results show that stratifying the sampling design—a method requiring a certain amount of supplementary information—does not significantly improve the model accuracy attained by regular grid designs, which are far simpler to implement, at least when the aim is the prediction of presence/absence. However, further investigations need to be conducted to confirm our results, since these were obtained for one particular area and a single

virtual species. In particular, one might consider defining the ecological requirements of the latter in another way.

Acknowledgements

References

Aspinall, R.J., Lees, B.G., 1994. Sampling and analysis of spatial environmental data. In: Waugh, T.C., Healey, R.G. (Eds.), Advances in GIS research. Taylor & Francis, Southhampton, pp. 1086–1098.

Austin, M.P., Heyligers, P.C., 1989. Vegetation survey design for conservation: gradsect sampling of forests in north-east New South Wales. Biol. Conserv. 50, 13–32.

Austin, M.P., Adomeit, E.M., 1991. Sampling strategies costed by simulation. In: Margules, C.R., Austin, M.P. (Eds.), Nature Conservation: Cost Effective Biological Surveys and Data Analysis. CSIRO, Australia, pp. 167–175.

Austin, M.P., Heyligers, P.C., 1991. New approach to vegetation survey design: gradsect sampling. In: Margules, C.R., Austin, M.P. (Eds.), Nature Conservation: Cost Effective Biological Surveys and Data Analysis. CSIRO, Australia, pp. 31–36.

Austin, M.P., Meyers, J.A., Belbin, L. Doherty, M.D., 1995. Modelling of landscape patterns and processes using biological data. Subproject 5: Simulated data case study. Division of Wildlife and Ecology, Commonwealth Scientific and Industrial Research Organisation (CSIRO).

Cohen, J., 1960. A coefficient of agreement of nominal scales. Educ. Psychol. Measure. 20, 37–46.

Eastman, J.R., 1999. IDRISI 32. Worcester, Clark University. URL: http://www.clarklabs.org.

Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 24, 38–49.

Franklin, J., 1998. Predicting the distribution of shrub species in southern California from climate and terrain-derived variables. J. Veg. Sci. 9, 733–748.

Goedickemeier, I., Wildi, O., Kienast, F., 1997. Sampling for vegetation survey: some properties of a GIS-based stratification compared to other statistical sampling methods. Coenoses 12, 43–50.

Guisan, A., Theurillat, J.-P., 2000. Equilibrium modelling of alpine plant distribution and climate change: how far can we go. Phytocoenologia 30, 353–384.

Table 6
Classification of the sampling strategies

	No extra knowledge needed	Ecological knowledge needed
Random	'Random'	'Proportional-stratified'
Systematic	'Regular'	'Equal-stratified'

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135 (2–3), 147–186.

Guisan, A., Weiss, S.B., Weiss, A.D., 1999. GLM versus CCA spatial modelling of plant species distribution. Plant Ecol. 143, 107–122.

Guisan, A., Theurillat, J.-P., Kienast, F., 1998. Predicting the potential distribution of plant species in an alpine environment. J. Veg. Sci. 9, 65–74.

Hirzel, A., Helfer, V., Métral, F., 2001. Assessing habitat-suitability models with a virtual species. Ecol. Model. 145, 111–121.

Jaberg, C., Guisan, A., 2001. Modelling the distribution of bats in relation to landscape structure in a temperate mountain environment. J. Appl. Ecol. 38 (6), 1169–1181.

Jongman, R.H.G., ter Braak, C.J.F., van Tongeren, O.F.R., 1995. Data Analysis in Community and Landscape Ecology. Cambridge University Press, Cambridge, UK.

Le Duc, M.G., Hill, M.O., Sparks, T.H., 1992. A method for predicting the probability of occurrence using data from systematic surveys. Watsonia 19, 97–105.

Maggini, R., Guisan, A., Cherix, D., 2002. A stratified approach for modelling the distribution of a threatened ant species in the Swiss National Park. Biodivers. Conserv., in press.

Margules, C.R., Austin, M.P., 1994. Biological models for monitoring species decline: the construction and use of data bases. Philos. Trans. R. Soc. London B 344, 69–75.

Mohler, C.L., 1983. Effect of sampling pattern on estimation of species distributions along gradients. Vegetation 54, 97–102.

Rice, W.R., 1989. Analysing tables of statistical tests. Evolution 43, 223–225.

Wessels, K.J., Van Jaarsveld, A.S., Grimbeek, J.D., Van der Linde, M.J., 1998. An evaluation of the gradsect biological survey method. Biodivers. Conserv. 7, 1093–1121.