Using Niche-Based Models to Improve the Sampling of Rare Species

ANTOINE GUISAN,*†† OLIVIER BROENNIMANN,* ROBIN ENGLER,* MATHIAS VUST,* NIGEL G. YOCCOZ,† ANTHONY LEHMANN,§ AND NIKLAUS E. ZIMMERMANN‡

*University of Lausanne, Department of Ecology and Evolution (DEE), Laboratory for Conservation Biology (LBC), Biology Building, CH-1015 Lausanne, Switzerland
†Institute of Biology, University of Tromsø, 9037 Tromsø, Norway
‡Swiss Federal Research Institute WSL, Zürcherstrasse 111, CH-8903 Birmensdorf, Switzerland
§Swiss Center for Faunal Cartography, Terreaux 14, CH-2000 Neuchâtel, Switzerland

Abstract: *Because data on rare species usually are sparse, it is important to have efficient ways to sample additional data. Traditional sampling approaches are of limited value for rare species because a very large proportion of randomly chosen sampling sites are unlikely to shelter the species. For these species, spatial predictions from niche-based distribution models can be used to stratify the sampling and increase sampling efficiency. New data sampled are then used to improve the initial model. Applying this approach repeatedly is an adaptive process that may allow increasing the number of new occurrences found. We illustrate the approach with a case study of a rare and endangered plant species in Switzerland and a simulation experiment. Our field survey confirmed that the method helps in the discovery of new populations of the target species in remote areas where the predicted habitat suitability is high. In our simulations the model-based approach provided a significant improvement (by a factor of 1.8 to 4 times, depending on the measure) over simple random sampling. In terms of cost this approach may save up to 70% of the time spent in the field.*

Key Words: efficiency, endangered species, *Eryngium alpinum*, habitat suitability maps, population discovery, predicted species distribution, prospective sampling

Utilización de Modelos Basados en Nichos para Mejorar el Muestreo de Especies Raras

Resumen: *Debido a que los datos sobre especies raras generalmente son escasos, es importante contar con formas eficientes para obtener datos adicionales. Los métodos de muestreo tradicionales tienen valor limitado para especies raras porque en una gran proporción de sitios de muestreo seleccionados al azar es poco probable que se encuentre a la especie. Para estas especies, las predicciones espaciales de modelos de distribución basados en nicho pueden ser utilizadas para estratificar el muestreo e incrementar la eficiencia de muestreo. Los nuevos datos obtenidos luego son utilizados para mejorar el modelo inicial. La aplicación repetida de este método es un proceso adaptativo que puede permitir el incremento del número de ocurrencias nuevas. Ilustramos el método con un estudio de caso de una especie de planta rara y en peligro en Suiza y un experimento de simulación., Nuestro trabajo de campo confirmó que el método ayuda al descubrimiento de nuevas poblaciones de la especie en áreas remotas en las que la adecuación pronosticada del hábitat es alta. En nuestras simulaciones, el método basado en modelos aportó un mejoramiento significativo (por un factor de 1.8 a 4 veces, dependiendo de la medida) del muestreo aleatorio simple. En términos de costos, este método puede ahorrar hasta 70% del tiempo de trabajo de campo.*

Palabras Clave: descubrimiento de población, eficiencia, distribución pronosticada de especies, *Eryngium alpinum*, especies en peligro, mapas de adecuación del hábitat, muestreo prospectivo

††*email antoine.guisan@unil.ch*
Paper submitted February 23, 2005; revised manuscript accepted June 8, 2005.

Conservation Biology Volume 20, No. 2, 501–511

DOI: 10.1111/j.1523-1739.2006.00354.x

Introduction

Monitoring populations of rare and endangered species has become a priority for most conservation agencies. It provides the major source of data for updating World Conservation Union and national red lists (Lamoreux et al. 2003), the main use of which is setting the long-term goals of conservation programs, such as helping identify biotopes or areas particularly in need of protection (Prendergast et al. 1993). Although precise quantitative and qualitative criteria are used to assign a particular status to a species (IUCN/SSC 2001), such decisions depend on the way the target species populations are monitored and on the sampling design used for data collection. A range of sampling schemes can be used to estimate the state of the system and its rate of change, the most efficient of which usually combine repeated (random) samples among existing sites with additional independent random samples at unvisited locations (Yoccoz et al. 2001; Stauffer et al. 2002).

Yet this random component is seldom considered in conservation surveys of rare species, although it is fundamental for ensuring unbiased population estimates and calculating unbiased estimates of distributional and population states. For example, remote populations may differ from accessible ones with respect to important environmental and management variables. In the context of distribution modeling, bias is thus likely to affect the quantification of the realized niche of a species and therefore the assessments that could be based on it. These include assessing the impact of climate change on species distribution and associated risks of extinction or searching for reintroduction sites for a species.

When species are rare, standard sampling methods such as simple or stratified random sampling, based on a simple combination of the main environmental gradients, can be highly inefficient (Rushton et al. 2004). For instance, in a sample of 550 plots surveyed in a random-stratified way based on the elevation, slope, and aspect of the plot during two consecutive summers in the Swiss Alps (704.2 km^2), not one occurrence of the rare and endangered plant species *Eryngium alpinum* L. was recorded. This was despite the species being easily detectable if present and independent records of the species existing in the area within similar vegetation types. Directing the sampling to ensure inclusion of units with a higher probability of hosting the species is thus a desirable approach for increasing survey efficiency and reducing sampling costs.

Predictions from niche-based models of species distribution (Guisan & Zimmermann 2000) are promising tools in this respect (Côté & Reynolds 2002; Edwards et al. 2005) as a way to improve the sampling of species of conservation interest. Although population viability analyses have long been used in rare species management (Brook et al. 2000), spatially explicit, predictive, habitat distribution models have only recently been used in conservation biology (Vaughan & Ormerod 2003; Rushton et al. 2004). The majority of predictive models published in the literature were developed for common plant and animal species or for biodiversity. To date, relatively few successful applications of this approach have been published for rare and endangered plant species (but see Miller 1986; Elith & Burgman 2002; Engler et al. 2004), although reliable spatial predictions are essential for species of great conservation interest. Paucity of data, spatial inaccuracy, and lack of valid absences are the main reasons identified for this shortcoming (Engler et al. 2004).

Recent methodological progress (Guisan & Thuiller 2005), such as improved predictive algorithms and more causal environmental predictors at a better spatial resolution, has made predictive models more reliable and applicable to the sampling of rare and endangered species (Ferrier 2002). Here, we propose a procedure for using niche-based models of species distribution to direct a random-stratified prospect of the study area and improve the sampling of rare and endangered species. We illustrate the approach with the rare species *E. alpinum* by using data taken from a modeling study of endangered plant species in Switzerland (O.B., unpublished). We also tested the approach through simulations. Finally, we discuss the conditions for the successful application of this approach, highlight some of its main limitations, and make suggestions for future research.

Methods

A General Procedure for Model-Based Sampling of Species Distribution

The general procedure for model-based sampling of species distribution is pictured in Fig. 1. A first model is

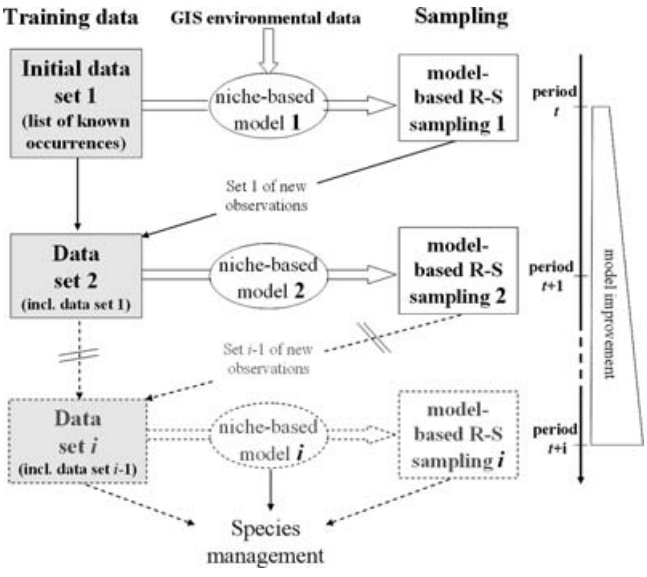

Figure 1. Analytical procedure illustrating the iterative model-based sampling process.

fitted with available data. These data include information on monitored populations or records of species that are typically extracted from a computer data bank, such as those developed in botanical conservatories, museums, or biological data centers (Graham et al. 2004). Spatial predictions derived from this initial model are then used to stratify the field sampling. Data gathered during this first sampling campaign serve to update the training data set and fit improved models that are used to direct the next sampling stage, and so on iteratively over several field seasons (Fig. 1). The higher prediction strata are expected to reveal those occurrences of suitable environmental conditions for the species where undetected or newly established populations, not recorded in the initial data set, might prevail. This iterative process can be repeated during the same field season or over consecutive seasons, and its success is mainly measured by the number of new occurrences found at each stage.

Study Area and Species Data

The study area is the same as in Engler et al. (2004). We illustrate the proposed approach with *E. alpinum*, commonly called alpine eryngo, an emblematic species endangered in most parts of the calcareous European Alps. It grows mainly on deep and moist soils with intermediate to high level of nutrients, preferably on steep slopes where screes of large blocks occur and solar radiation is rather high. The exact reasons for its recent decline are still unknown, but possible threats could be picking and changes in pastoral practices (Engler et al. 2004).

Original species data on *E. alpinum* were provided by the Swiss Floristic Network (CRSF) in Geneva (data can be ordered from http://www.crsf.ch). These records are usually provided by various volunteers and professionals, and thus they share the same idiosyncrasies as those in other natural history collections (Graham et al. 2004) except that, in our case, information on the positional accuracy of each record is also provided. Only data with a minimum horizontal accuracy of 25 m were retained for the analyses, to match the resolution of environmental data layers used. Remaining data were used for visual evaluation of the prediction maps. Because only presence observations were initially available in the database, fitting the initial model required the generation of pseudoabsences. This was based on occurrences of 11 other rare species. With easily detectable and well-known rare species, like *E. alpinum*, this method of generating pseudoabsences works well because providers of rare species observations usually have a good knowledge of most other rare species. Thus, for each single observation of a rare species one can confidently assume the absence of all other possible rare species that usually share the same type of habitat.

A total of 92 presences and 2380 absences were used to develop the initial predictive model. Following the recommendations of Manel et al. (2001), we weighted the absences in the generalized linear models (GLMs) to ensure an equal prevalence (0.5) between presences and absences. This was achieved by providing a vector of weights with "1" for presences and "92/2380" for absences; so the total weight of absences was equal to the total weight (sum) of presences.

Environmental Predictors

We selected quantitative predictors to reflect the main biophysical gradients with a recognized, physiological influence on plants. We generated maps of monthly precipitation sums and monthly temperature averages from the Swiss meteorological network (normals 1961–1990) and a 25-m digital elevation model (DEM) and used them to derive predictors with a more causal effect on the fitness and survival of species. Topographic positions and slope angle were additionally derived from the DEM to express microclimatic corrections and disturbance effects. Table 1 lists all the predictors that were used to fit the models. Details on these predictors can be found in Zimmermann and Kienast (1999).

Table 1. Quantitative geographic information system predictor variables used to fit the models for *Eryngium alpinum*.

Variables	*Description*
Seasonal indices of precipitation	obtained by a principal component analysis on monthly precipitation maps; axis 1 is the yearly sum of precipitation, whereas axes 2 and 3 express two seasonal variations trends
Monthly average of potential daily global radiation in March and July	calculated from the digital elevation model (DEM) as the sum of direct and diffuse radiation
Number of days of precipitation per growing season	calculated from a regionalized regression model of the annual number of days with rainfall of more than 1 mm on elevation
Degree-days of growing season	calculated as the sum of interpolated temperatures above the threshold of 3° C
Number of frost days	defined as a sudden drop of the daily minimum temperature below 2° C preceded by a period of at least 1 day above 3° C; expresses the core 90% of the vegetation period
Topographic position	relative topographic exposure of a pixel compared with its local neighborhood; indicates ridge tops, regular slopes, or valley bottoms
Site water balance	estimate of the water available to plants during a water year, based on precipitation, evapotranspiration, and the soil bucket size derived from the soil suitability map of Switzerland and topographic position
Slope angle	derived from the DEM

Qualitative predictors were broken down into disjunctive classes, which were then used to filter the predictions made by the models. For qualitative filters we used simplified maps of geology and land cover types.

Niche-Based Statistical Models

We used generalized additive models (GAMs; Hastie & Tibshirani 1986) with a binomial distribution and logistic link function to fit the models based on topographic and climatic predictors (hereafter topoclimatic models). Model selection followed a stepwise algorithm based on Akaike's information criterion (AIC) that operates both backward and forward from an initial full model. We used splines as smoothers, with up to four degrees of freedom allowed in the stepwise procedure, as set by default in the program GRASP (Generalized Regression Analysis and Spatial Prediction; Lehmann et al. 2002*b*).

Each topoclimatic model was filtered by the binary variables derived from the qualitative predictors. To be a filter, the species must never occur on any pixel of the corresponding disjunctive class. Typical filters for *E. alpinum* include forested areas and all classes of siliceous bedrock. Although this might have the potential to bias the future sampling, for instance if the original data are biased against some rock or vegetation classes, such a risk is mostly reduced by also sampling outside the suitable areas (see below). Care should thus be taken when using such filters in other situations.

The agreement between predictions and observations was assessed using the standard area-under-the-curve (AUC) measure of a receiver-operating characteristic (ROC) plot (Fielding & Bell 1997) and its cross-validated version (AUC_{CV}). Values of AUC vary between 0.5 for an uninformative, random model and 1 for a model with perfect discrimination. Cross-validation was performed by splitting the data set into five partitions (5-fold CV) and reestimating the model coefficients at each loop.

The minimal predicted area (MPA; Engler et al. 2004) was calculated to complement the other evaluation measures and compare the predicted maps obtained after each loop of the reiterative modeling/field testing process (Fig. 1). It also constituted the primary stratification factor for the field sampling. The MPA is the surface obtained by considering all pixels with predicted values above the probability threshold that still encompasses 100% of the species occurrences (rule of parsimony). The resulting MPA map is binary, with 0 for cells where the habitat is predicted unsuitable and 1 for cells where it is predicted suitable.

To make spatial predictions of species distributions on a large data set possible (64 million pixels at 25-m resolution for Switzerland), we used lookup tables generated from the GAM models. A custom script was used in the ArcView geographic information software (ESRI, Redland, California) to implement the models and calculate the final potential maps.

The entire procedure, except the use of filters, was automated with the GRASP library of S-PLUS (Insightful Corp., Seattle, Washington) functions (Lehmann et al. 2002*b*; GRASP and the ArcView script are available online from http://www.cscf.ch/grasp).

Implementing the Model-Based, Random-Stratified Sampling

The model-based sampling was designed as follows. First, we fitted a model for the species based on topographic and climatic predictors, filtered by binary classes of qualitative predictors and used to predict the species distribution over the study area. Second, we calculated the MPA. Third, we superimposed the MPA map on a relief map and distinguished subareas of similar geological substratum belonging to the same macrotopographic unit (e.g., large massifs) (Fig. 2a). Because of limited sampling resources, only subareas including at least one patch of cells with high habitat suitability for the species (i.e., above the MPA threshold) were considered further for the sampling (Fig. 2a). We visited a random selection of these subareas in the field, according to a standardized field methodology. Hence we did not formally use a probability sample of the whole area, and the results should not be interpreted outside the range of the candidate subareas that define our target population.

We conducted survey walks in visited subareas, following a random trajectory based on the MPA map and designed to cross approximately an equal number of suitable and unsuitable pixels. In an attempt to reduce costs and increase survey efficiency, we opted for survey walks across the probability gradient rather than a strict random-stratified procedure, based on sampling individual pixels randomly in both strata (within and outside MPA), because surveying all of Switzerland was not possible. The full trajectory was recorded using a GPS system (Garmin eTrex Summit, Olathe, Kansas, locational accuracy <10 m). New occurrences, when found, were recorded and georeferenced using the GPS. A series of points without species observations were also sampled at regular intervals (at least 500 m, to avoid spatial autocorrelation) along the recorded trajectory to generate likely absences for the species.

All new and updated presences and absences were used to fit a second, improved model to be used as a next step to redirect the sampling effort (Fig. 1). New observed absences force the model to be more discriminating at locations where the prior model was predicting high suitability values. On the other hand new observations of presences help remove bias from predictions and narrow the species realized niche. A bias may for instance be introduced in the original model if the species observations offered only a partial view of the species niche, which

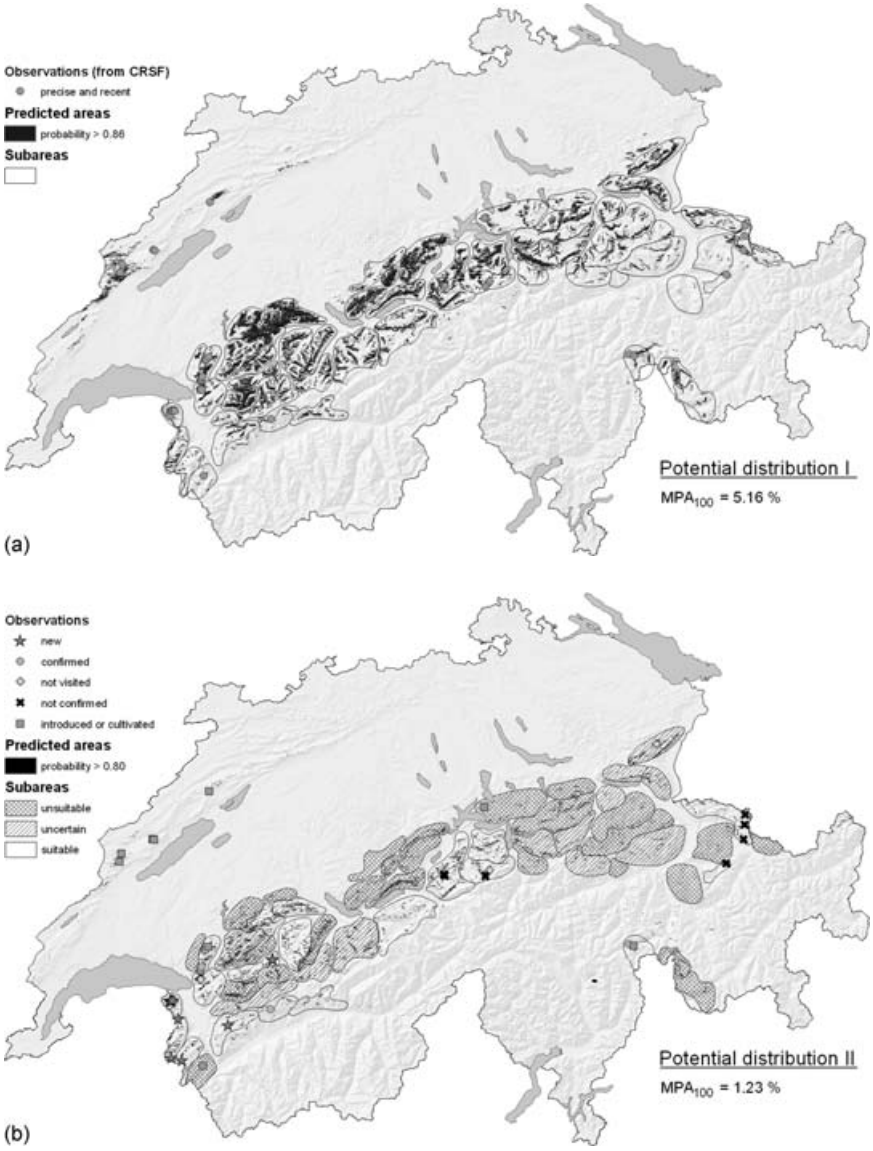

Figure 2. Potential distribution maps for Eryngium alpinum *in Switzerland. (a) Model 1, used to stratify the sampling. Subareas used for sampling are also shown (polygons; see text). Observations correspond to recent (1995 or later) and precise (25-m accuracy) observations. (b) Model 2, improved from model 1 with data sampled in the field. Updated knowledge on geological units important for the species, derived from the field campaign, is also shown (polygons with various shading). The predicted area appears in both maps as dark grey and corresponds to the minimal predicted area (MPA; see methods) containing 100% of observations (MPA_{100}).*

could result if only some geographic areas were subjectively sampled. Because visual detectability was high for *E. alpinum*, absences generated in this way are likely to be reliable. This is unlikely to hold for inconspicuous species; thus field sampling and modeling approaches, including detectability (Stauffer et al. 2002; Edwards et al. 2005; MacKenzie et al. 2005), should be used in these cases. The status of each binary class of qualitative predictor used as a filter was also revised after field sampling (Fig. 1) by checking that no new presence occurred in any of them.

Testing the Approach by Simulation

To test sampling improvement over several iterations, we conducted simulations with a virtual species in a real landscape (Hirzel & Guisan 2002). The distribution of a virtual species can be defined from the available environmental predictors by specifying response curves for each predictor and combining them to draw a habitat suitability map. The final species distribution map is then obtained by cutting the previous suitability map into a presence-absence map reflecting presence and absence of the virtual species in each pixel of the study area. The advantage here is that the truth is known; thus model predictions can always be compared with this original "true" distribution of the species.

In our case, the true distribution of the virtual species in the study area was artificially derived from the first model for *E. alpinum* so that it remained as close as possible to a real situation (same type of distribution, same study area). To turn the probabilistic map into a binary presence-absence map, we used a cut-off value of 0.8, yielding a distribution restricted to about 5% of the study area. This map reflected the true distribution of the virtual species. This way, field iterations could be simulated

simply by checking the presence or absence of the virtual species at each sampling site on this map. After each iteration, the predicted distribution map was reclassified using the MPA threshold—as required by the model-based stratification of the sampling—and the resulting binary predictions were compared with the virtual distribution with the kappa coefficient (Cohen 1960). We used kappa here instead of the AUC because the comparisons involved directly two binary variables. Hence it provided a reliable measure of model improvement over time (here, the successive iterations), reflecting a convergence, or no convergence, toward the true distribution. For computational reasons the simulations were not run over the entire Swiss landscape but on a subset of 500,000 pixels selected in a random-stratified way across the main environmental gradients. This subset of pixels accurately reflected the prevalence of the different habitat suitability classes of the virtual species.

The iterative, model-based, random-stratified sampling procedure was simulated as follows:

1. An initial data set of 30 presences and 30 absences was randomly selected among the pools of 92 presences and 2380 absences used to calibrate the "true distribution" of our virtual species. Hereafter we refer to this data set as the training data set.
2. Using the training data set, a GAM was fitted and predictions were made across the entire study area (i.e., the 500,000 randomly chosen pixels). We used the MPA threshold to transform probabilistic predictions into binary values (presence or absence of the species).
3. The agreement between the binary predictions map and the known binary distribution of the virtual species was evaluated using the kappa coefficient.
4. An equal number of sample sites were selected randomly within (predicted presence) and outside (predicted absence) the MPA (i.e., the two sampling strata). To avoid spatial autocorrelation we did not allow the selection of new sample sites within a radius of 500 m of any already sampled site.
5. The fieldwork was simulated by checking, from the virtual distribution, for true presence or absence in each sample site. At this stage we recorded the number of new occurrence sites found at each iteration. All new and updated observations were then added to the training data set.
6. Steps 2 to 5 were repeated 10 times (e.g., simulating a 10-year survey). Because the selection of the new sample sites involved randomness, we repeated each loop 20 times to quantify the related uncertainty (Fig. 3).

We ran this simulation procedure for three different experiments: sample 30, sample 60, and sample 120. Numbers indicate the initial number of observations and the number of sites sampled at each iteration (e.g., 30 means 15 sites sampled within the "predicted presence" strata and 15 sites within the "predicted absence" strata).

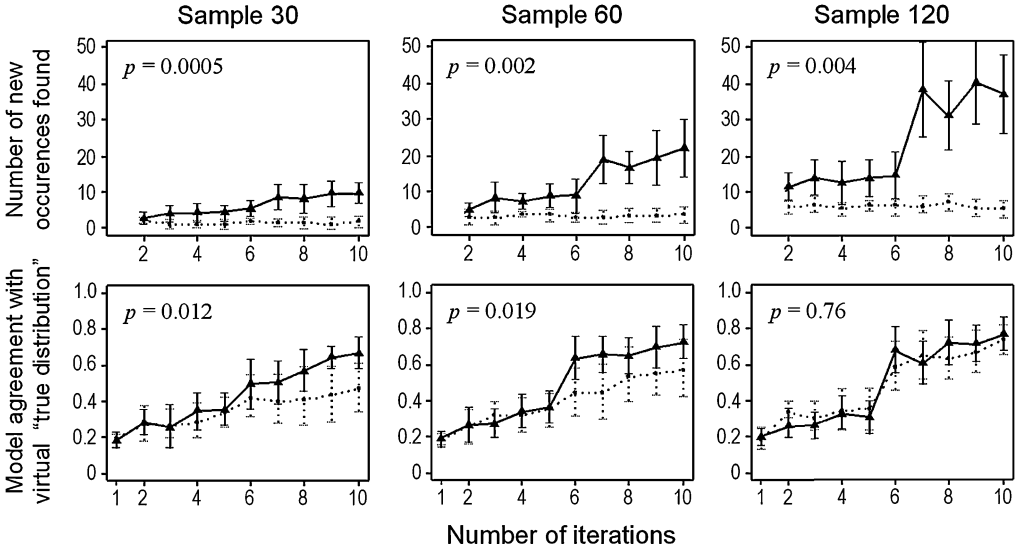

Figure 3. Average (±SE) number of new occurrences of Eryngium alpinum *at each iteration (upper graphs), and the kappa agreement between the predictive model built at each iteration and the true distribution of the virtual species (lower graphs). Solid lines represent simulation with model-based stratification when selecting the new locations to sample. Dashed lines are the control simulation, where the new points to be sampled were chosen completely at random. The horizontal series of graphs correspond to increasing number of sample points visited at each iteration (30, 60, 120). The* p *values are only indicative because they depend on the number of simulations. Each point on a graph was obtained as an average of 20 runs. The field work of iteration* i + 1 *is done using the model of iteration* i*. This explains why the "number of new occurrences found" starts at iteration 2.*

Additionally, control simulations in which the new sample sites were chosen completely at random in the landscape, rather than on a model-based stratification basis, were run for each scenario (Fig. 3). We repeated these control simulations 20 times. Differences between model-based and full-random-sampling simulations were assessed with paired t tests or Wilcoxon signed rank tests when normality was not satisfied. When values were $>$ 0.05, the difference between the two curves was not significant.

Results

A New Picture of the Species' Distribution

Hardly any of the records from the eastern part of the Swiss Alps were confirmed by revisiting existing occurrence localities in 2003, thus offering a very different view of the species range at the end of the field campaign from that suggested by the historical data records. Based on confirmed records only, the status of the species in the Swiss red list would qualify for a downgrade from vulnerable to endangered. Interestingly, most of the unconfirmed records seemed to be specimens that had escaped from nearby gardens, where the plant was cultivated artificially, and in most cases in areas that do not appear naturally suitable to the species long-term survival (when compared with sites where the species was observed).

Model-Based, Random-Stratified Sampling for *E. alpinum*

The first predictive model for *E. alpinum* yielded values of simple evaluation and cross-validation AUC of 0.97 and 0.96, respectively, and the corresponding map was filtered using several qualitative variables (Fig. 2a). Positive filter classes for the species were land use (1, bushes in a 200-m buffer zone from meadow/pasture; 2, rocks in a 200-m buffer zone from meadow/pasture; 3, drain stone; 4, meadow/pasture) and geology (rock and soil) (1, sand and silts; 2, coarse screes [fallen blocks]; 3, marl; 4, marly schist; 5, calcareous phyllites; 6, massive limestone; 7, sandy limestone). For any other qualitative class, predictions were reset to 0. The 100% threshold for this first map was at $p = 0.86$, and the MPA covered 5.1% of the entire Swiss territory. The resulting binary map was superimposed on the shaded relief and divided into subareas (Fig. 2a). A random selection of nearly half the subareas (45.7%, covering 1438 km^2) was visited during summer 2003. Seventy-nine survey walks adequately sampled the two sampling strata, with 56.9% of the paths in areas outside MPA ($p < 0.86$) and the remaining 43.1% of paths within MPA areas ($p \geq 0.86$).

Overall, seven new populations were discovered, which were not previously recorded in the database or the literature. All the new populations occurred in western Switzerland, within highly suitable landscape patches (probability of presence $>$ 91%), and were subsequently used to fit an improved model for the species (Fig. 1).

The second model, improved from a field-based update of the initial data (new observations, update of historical records, and a set of field-checked absences), and revised filters (land use 1 and geology [1–3] removed; classes of slope and aspect added) provided a much narrower prediction of the species range (MPA of 1.2%) that better reflected its actual niche, distribution (local and patchy), and endangered status in the Swiss Alps (Fig. 2b). The same predictors were retained; thus only their coefficients were updated. Most subareas where the species was previously predicted by the primary model, but where no species occurrence could be found, no longer included any suitable patch in the map drawn from the improved model (Fig. 2b). This second model and related spatial predictions had simple validation and cross-validation AUC coefficients of 0.99 and 0.98, respectively.

Simulations

In all three scenarios the number of new presences found over 10 iterations was always much higher when using the model-based stratification approach (about four times in all simulations; Fig. 3). For instance, at a sample size of 30, only 6 iterations were needed with the model-based sampling to double the initial number of presences (i.e., to pass from 30 to 60), whereas nearly 20 iterations would have been needed with simple random sampling. In this case, the model-based approach was 3.3 times faster, corresponding to a net gain in time of 70%. At a sample size of 60, the model-based approach was still 2.3 times faster (considering the doubling case), allowing a 55% gain of time compared with random sampling. In all experiments, the cumulative number of new occurrences found by model-based sampling after 10 iterations was always about 4 times greater than those found by a simple random sampling.

Averages and standard errors of the adequacy between the predictions and the true distribution of the virtual species are also reported for all simulations (Fig. 3). Except for the sample120 scenario, where no significant difference was found, the model-based sampling approach led, on average, to a predicted distribution that had higher agreement with the true distribution of the virtual species than that obtained with the random sampling.

Discussion

We have presented and illustrated the use of predictions obtained from niche-based species distribution models for stratifying sampling and improving data sets on rare and endangered species. Similar approaches based on simple models (e.g., bioclimatic envelope models) have

been published independently (Parris 2002; Poon & Margules 2004). The procedure has the advantage of fitting an existing sampling design in ecology and conservation biology (random-stratified approach), and it adds a new model-based sampling scheme that is expected to increase sampling efficiency, particularly the chance to discover new populations of rare species. The approach was successfully verified through field checking—with seven new populations discovered for the severely endangered species *E. alpinum*—and in computer simulations.

Robust Modeling Approaches Needed for Stratification

To be efficient, the stratification ideally has to be based on as robust an initial model as possible to reduce costs by reiterating only a few sampling stages. Thus the modeling strategy itself is important. Four elements can contribute to a successful modeling effort when species data from such heterogeneous data banks are used. First, species occurrences need to be carefully selected before conducting the statistical analyses by considering the positional accuracy of site location (when recorded in the database). In our study we dropped several occurrences from the original data set, keeping only those corresponding to the grain size and positional accuracy used for the modeling. Such operations limit measurement errors significantly (Engler et al. 2004).

Second, close attention needs to be given to preparing and selecting predictors, including only those expected to have a causal physiological effect on the species (Austin 2002).

Third, using an appropriate modeling technique for the data at hand (e.g., when sufficient occurrences are available, use a flexible semiparametric modeling approach such as GAM) allows for the calculation of response curves that more closely fit the data (Guisan et al. 2002; Lehmann et al. 2002*a*). This is particularly appealing when the predictions are to be made in the same area that was used to calibrate the model. Response plots in GAMs may additionally allow one to check plot density along each environmental predictor and improve the sampling of unevenly sampled predictors during the next field campaign. When fewer data are available, more simple approaches such as climatic envelopes (e.g., BIOCLIM) can be used instead (Guisan & Thuiller 2005). Finally, fourth, when absences are not available from a systematic survey, generating pseudoabsences from presence sites of other rare species seems a promising approach (A.L., unpublished).

Use of Survey Routes

Survey routes recorded on GPS may be a valuable way to explore sampling frameworks further, as a means to lower survey costs and provide additional absences for improving the models. As our results show, generating absences along survey routes can be useful to increase the discriminatory power of the models. Some important limitations exist, however, and should be highlighted before this approach can be further promoted.

Although our species was easily detectable, which was convenient for the first test of the approach, many species will indeed be much harder to detect in the field. Species detectability (MacKenzie et al. 2005) is a major component to consider because it can affect the whole model-based approach in two ways: (1) as additional cost during the field sampling (e.g., more time might be needed to detect cryptic than conspicuous species) and (2) in the way absences are derived from survey routes because these absences might not have the same reliability for different species. The latter aspect is of prime importance when presence-only data are the sole initial source of species data available at the start of the study, as is the case with data from natural history collections (Graham et al. 2004; Rushton et al. 2004).

The use of survey routes across stratifying gradients (in our case, predicted presence or absence of the species) is specific to cases where survey costs need to be reduced. The approach is similar to the cost-reduction "gradsec" sampling technique described by Austin and Heyligers (1989), although applied to a model-based stratified sampling situation rather than to an environmentally based stratification of sampling. Refinements might come here from such guided transects across the stratifying gradients (Ståhl et al. 2000; Thompson 2004) rather than survey routes.

In cases where there is no need to reduce survey costs, a strict model-based, random-stratified approach (i.e., no survey route, no gradsec) in which target cells are randomly chosen in each prediction strata throughout the whole study area should be preferred. The latter is the approach that was used in our simulations.

Confirmation by Simulations

Our simulation results showed that using a model-based rather than a simple random sampling (control) yields both a greater number of presences and a higher adequacy of model predictions. Obtaining such improvement over a random sampling is not surprising because no one would sample rare species randomly in the field, but it provides a standardized and objective measure of improvement. This is, however, mostly true below a certain sample size. With 120 sites (or more) sampled at each iteration, the model-based approach did not converge significantly faster toward the true distribution than the simple random sampling, but the number of new presences found remained much greater (nearly four times). With a sample size of 30 and 60 sites visited at each iteration, the improvement was very apparent, with a significantly greater number of presences found. Thus, the fewer the number of sites sampled at each iteration, the more valuable the use of

the model-based approach. This has direct implications for reducing costs of surveys in nature conservation.

A Flexible, Adaptive Approach

One advantage of this approach is to let the model reflect the current state of a species distribution (or of a set of species). In this sense, it is adaptive. This is particularly important because some environmental conditions or the fitness of some populations might change during the course of the survey, possibly resulting in a temporary lower model evaluation. Such a temporary decrease in model performance may also happen under stable conditions, as observed by simulations (internal variation within the 20 replicates; see error bars in Fig. 3). In the latter cases, the model returned rapidly (usually within the next iteration) to a situation of higher agreement with the true geographic distribution. The same might be expected after an environmental change, which is thus only likely to delay model convergence by some number of iterations.

Which Method for Which Data?

In this study, we had recourse to GAMs to predict the distribution of a rare species for which 92 occurrences were initially available. GAMs, however, are relatively data-hungry modeling techniques. As such they need a sufficient number of occurrences to be fitted. In many cases fewer occurrences will be available at the start of the study, as in the case of many rare and endangered species in data-poor countries, which will prevent the use of such an advanced technique (and others of similar complexity). When too few species occurrences are available (say < 20), alternative approaches must be found. Alternatives can be (1) a simpler technique such as climatic envelopes (Poon & Margules 2004); (2) a modeling approach at the community level (Ferrier et al. 2002), where the data for more common species may help support the modeling of less frequent species; or (3) models for more common species that are frequently associated with the rare target species. In the last case, Edwards et al. (2005) show that a model-based stratified sampling based on common species can significantly improve detectability of rare species in coarse-scale surveys.

The resolution used in our study was also much finer than often available elsewhere. Although this approach could theoretically be conducted at any resolution, problems may well appear when the size of the sampling unit becomes too large to be properly surveyed in the field. Again, the choice of an appropriate resolution and survey design is likely to depend primarily on a species' detectability (screening a large sampling cell is easier for a conspicuous than for a cryptic species). Future tests of the method may thus need to take detection probability into account in the sampling design and related statistical inferences (MacKenzie et al. 2005).

Sampling Bias and Predictors Used

There might be a risk of converging toward a biased model, because of the influence of some of the initial parameters (e.g., initial sample size, number of sites sampled at each field iteration, initial environmental or geographic bias in the data). As long as the area is sampled probabilistically in both the suitable and unsuitable strata, however, the model iterations should converge rapidly toward unbiased estimates.

Besides an initial sampling bias, predictions may additionally be weakened—and thus with it the strength of the approach—if one or several proximal environmental predictors (Austin 2002) are not included in the initial model or if the model is severely overfitted. It is thus critical to decide which predictors should be included and how strict the limitation of the final number of predictors in the model should be. With only a few species occurrences, a limitation rule that is too strict (e.g., a stepwise procedure in a regression model, based on the Bayesian information criterion) may prevent the inclusion of some proximal predictors, whereas including too many predictors may cause the model to direct the sampling too narrowly toward some situations. In both cases the effect is likely to be a reduction in sampling efficiency. Hence a tradeoff has to be found, which still requires further investigation.

Future Research Directions

Where multiple species and overall biodiversity are the focus, distinct sampling designs could be used (Ferrier 2002; Edwards et al. 2005) or the same model-based, random-stratified design might be adapted to a multispecies situation. The use of alternative modeling methods such as community modeling (Ferrier 2002; Ferrier et al. 2002) should also be investigated, within a similar model-based sampling framework, when too few occurrences are available at the start of the study but more common species were jointly inventoried (which was not the situation in our case).

Improved field design may also be implemented within such a model-based approach, such as adaptive cluster sampling (Thompson & Seber 1996). In adaptive cluster sampling, a more intensive search of the species in neighbor plots is made every time the species is met at one of the plots included in the original design. The search is conducted within a neighborhood window of specified shape. Every time the species is found again within this window, the same procedure is repeated until no additional occurrence is found. This design is optimal when improved estimates of the whole population size in the area are requested and is particularly efficient for sampling rare, clustered species (Thompson & Seber 1996; Christman 2004). Hence it will be particularly suited for the case of rare species with large populations in the areas where they occur (i.e., the "urban" type described in

Collins et al. [1993]). Such adaptive cluster sampling was for instance successfully tested in a parallel study of an invasive species—*Heracleum mantegazzianum*—that has still a scattered distribution in the study area (A.G., unpublished results).

Model-based sampling of rare species, involving reiterative alternation of modeling and field sampling phases, shows great promise for strengthening and complementing conservation practices and reducing sampling costs. It would be optimal if applied together with the monitoring of those additional parameters required by population management (e.g., accessibility of sites as a surrogate for picking; agricultural practices as a surrogate for grazing). The approach would still deserve more thorough testing, however, especially in the field, before it can be applied more widely. Once more broadly validated, it may be seen as a valuable approach for nature conservation agencies.

Acknowledgments

We thank B. Baümler at the Swiss Floristical Center in Geneva for making the species data available; L. Rechsteiner, C. Trippi, and Y. Naciri-Graven for additional species data or for help with the field work; and M. Burgman and two anonymous reviewers for useful comments on an earlier version of the manuscript. We further thank the MAVA Foundation (Switzerland), the European Union through the FP6 project EVK3-2003-505373-IntraBioDiv, and the Biodiversity Programme of the Norwegian Research Council for financial support.

Literature Cited

Austin, M. P. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecological Modelling **157:**101–118.

Austin, M. P., and P. C. Heyligers. 1989. Vegetation survey design for conservation: gradsec sampling of forests in north-east New South Wales. Biological Conservation **50:**13–32.

Brook, B. W., J. J. O'Grady, A. P. Chapman, M. A. Burgman, H. R. Akcakaya, and R. Frankham. 2000. Predictive accuracy of population viability analysis in conservation biology. Nature **404:**385–387.

Christman, M. C. 2004. Sequential sampling for rare and geographically clustered populations. Pages 134–145 in W. L. Thompson, editor. Sampling rare or elusive species. Island Press, Washington, D.C.

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20:**37–46.

Collins, S. L., S. M. Glenn, and D. W. Roberts. 1993. The hierarchical continuum concept. Journal of Vegetation Science **4:**149–156.

Côté, I. M., and J. D. Reynolds. 2002. Predictive ecology to the rescue? Science **298:**1181–1182.

Edwards, T. C. J., R. Cutler, N. E. Zimmermann, L. Geiser, and J. Alegria. 2005. Model-based stratifications for enhancing the detection of rare ecological events: lichens as a case study. Ecology **86:**1081–1090.

Elith, J., and M. A. Burgman. 2002. Predictions and their validation: rare plants in the central highlands, Victoria, Australia. Pages 303–313 in J. M. Scott, P. J. Heglund, J. B. Haufler, M. Morrison, M. G. Raphael, W. B. Wall, and F. Samson, editors. Predicting species occurrences: issues of accuracy and scale. Island Press, Covelo, California.

Engler, R., A. Guisan, and L. Rechsteiner. 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. Journal of Applied Ecology **41:**263–274.

Ferrier, S. 2002. Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? Systematic Biology **51:**331–363.

Ferrier, S., M. Drielsma, G. Manion, and G. Watson. 2002. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. II. Community-level modelling. Biodiversity and Conservation **11:**2309–2338.

Fielding, A. H., and J. F. Bell. 1997. A review of methods for the assessment of prediction errors in conservation presence-absence models. Environmental Conservation **24:**38–49.

Graham, C. H., S. Ferrier, F. Huettman, C. Moritz, and A. T. Peterson. 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology & Evolution **19:**497–503.

Guisan, A., and W. Thuiller. 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters **8:**993–1009.

Guisan, A., and N. E. Zimmermann. 2000. Predictive habitat distribution models in ecology. Ecological Modelling **135:**147–186.

Guisan, A., T. C. Edwards, and T. Hastie. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modelling **157:**89–100.

Hastie, T., and R. Tibshirani. 1986. Generalized additive models. Statistical Sciences **1:**297–318.

Hirzel, A., and A. Guisan. 2002. Which is the optimal sampling strategy for habitat suitability modelling. Ecological Modelling **157:**331–341.

IUCN (World Conservation Union)/SSC (Species Survival Commission). 2001. IUCN Red List categories. IUCN, Gland, Switzerland.

Lamoreux, J., et al. 2003. Value of the IUCN Red List. Trends in Ecology & Evolution **18:**214–215.

Lehmann, A., J. M. Overton, and M. P. Austin. 2002*a*. Regression models for spatial prediction: their role for biodiversity and conservation. Biodiversity and Conservation **11:**2085–2092.

Lehmann, A., J. M. Overton, and J. R. Leathwick. 2002*b*. GRASP: generalized regression analysis and spatial prediction. Ecological Modelling **157:**189–207.

MacKenzie, D. I., J. D. Nichols, N. Sutton, K. Kawanishi, and L. L. Bailey. 2005. Improving inferences in population studies of rare species that are detected imperfectly. Ecology **86:**1101–1113.

Manel, S., H. C. Williams, and S. J. Ormerod. 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. Journal of Applied Ecology **38:**921–931.

Miller, R. I. 1986. Predicting rare plant distribution patterns in the southern Appalachians of the south-eastern U.S.A. Journal of Biogeography **13:**293–311.

Parris, K. M. 2002. The distribution and habitat requirements of the great barred frog (*Mixophyes fasciolatus*). Wildlife Research **29:**469–474.

Poon, E. L., and C. R. Margules. 2004. Searching for new populations of rare plant species in remote locations. Pages 189–207 in W. L. Thompson, editor. Sampling rare or elusive species. Island Press, Washington, D.C.

Prendergast, J. R., R. M. Quinn, J. H. Lawton, B. C. Eversham, and D. W. Gibbons. 1993. Rare species, the coincidence of diversity hotspots and conservation strategies. Nature **365:**335–337.

Rushton, S. P., S. J. Ormerod, and G. Kerby. 2004. New paradigms for modelling species distributions? Journal of Applied Ecology **41:**193–200.

Ståhl, G., A. Ringvall, and T. Lamas. 2000. Guided transect sampling for assessing sparse populations. Forest Science **46:**108–115.

Stauffer, H. B., C. J. Ralph, and S. L. Miller. 2002. Incorporating detection uncertainty into presence-absence surveys for Marblet Murrelet.

Pages 357–365 in J. M. Scott, P. J. Heglund, M. L. Morrison, M. G. Raphael, W. A. Wall, and F. B. Samson, editors. Predicting species occurrences: issues of accuracy and scale. Island Press, Covelo, California.

Thompson, S. K., and G. A. Seber. 1996. Adaptive sampling. Wiley, New York.

Thompson, W. L. 2004. Future directions in estimating abundance of rare or elusive species. Pages 389–399 in W. L. Thompson, editor. Sampling rare or elusive species. Island Press, Washington, D.C.

Vaughan, I. P., and S. J. Ormerod. 2003. Improving the quality of distribution models for conservation by addressing shortcomings in the field collection of training data. Conservation Biology **17:** 1601–1611.

Yoccoz, N. G., J. D. Nichols, and T. Boulinier. 2001. Monitoring of biological diversity in space and time. Trends in Ecology & Evolution **16:**446–453.

Zimmermann, N. E., and F. Kienast. 1999. Predictive mapping of alpine grasslands in Switzerland: species versus community approach. Journal of Vegetation Science **10:**469–482.