Ecology, 83(7), 2002, pp. 2027–2036

ECOLOGICAL-NICHE FACTOR ANALYSIS: HOW TO COMPUTE HABITAT-SUITABILITY MAPS WITHOUT ABSENCE DATA?

A. H. Hirzel,[1] J. Hausser,[1] D. Chessel,[2] and N. Perrin[1,3]

[1]*Laboratory for Conservation Biology, Institute of Ecology, University of Lausanne, CH-1015 Lausanne, Switzerland*
[2]*UMR CNRS 5023, Laboratoire de Biométrie et Biologie Evolutive, Université Lyon I, 69622 Villeurbanne Cedex, France*

Abstract. We propose a multivariate approach to the study of geographic species distribution which does not require absence data. Building on Hutchinson's concept of the ecological niche, this factor analysis compares, in the multidimensional space of ecological variables, the distribution of the localities where the focal species was observed to a reference set describing the whole study area. The first factor extracted maximizes the marginality of the focal species, defined as the ecological distance between the species optimum and the mean habitat within the reference area. The other factors maximize the specialization of this focal species, defined as the ratio of the ecological variance in mean habitat to that observed for the focal species. Eigenvectors and eigenvalues are readily interpreted and can be used to build habitat-suitability maps. This approach is recommended in situations where absence data are not available (many data banks), unreliable (most cryptic or rare species), or meaningless (invaders). We provide an illustration and validation of the method for the alpine ibex, a species reintroduced in Switzerland which presumably has not yet recolonized its entire range.

Key words: Capra ibex; *ecological niche; GIS; habitat suitability; marginality; multivariate analysis; presence–absence data; specialization; species distribution; Switzerland.*

Introduction

Conservation ecology nowadays crucially relies on multivariate, spatially explicit models in all research areas requiring some level of ecological realism. This includes population viability analyses (Akçakaya et al. 1995, Akçakaya and Atwood 1997, Roloff and Haufler 1997), biodiversity-loss risk assessment (Akçakaya and Raphael 1998), landscape management for endangered species (Livingston et al. 1990, Sanchez-Zapata and Calvo 1999), ecosystem restoration (Mladenoff et al. 1995, 1997), and alien-invaders expansions (Higgins et al. 1999). Such studies often conjugate the power of Geographical Information Systems (GIS) with multivariate statistical tools to formalize the link between the species and their habitat, in particular to quantify the parameters of habitat-suitability models.

Most frequently used among multivariate analyses are logistic regressions (Jongman et al. 1987, Peeters and Gardeniers 1998, Higgins et al. 1999, Manel et al. 1999, Palma et al. 1999), Gaussian logistic regressions (ter Braak and Looman 1987, Legendre and Legendre 1998), discriminant analyses (Legendre and Legendre 1998, Livingston et al. 1990, Manel et al. 1999), Mahalanobis distances (Clark et al. 1993), and artificial neural networks (Manel et al. 1999, Özesmi and Özesmi 1999, Spitz and Lek 1999). All these methods share largely similar principles:

1) The study area is modeled as a raster map composed of N adjacent isometric cells.

2) The dependent variable is in the form of presence/absence data of the focal species in a set of sampled locations.

3) Independent ecogeographical variables (EGV) describe quantitatively some characteristics for each cell. These may express topographical features (e.g., altitude, slope), ecological data (e.g., frequency of forests, nitrate concentration), or human superstructures (e.g., distance to the nearest town, road density).

4) A function of the EGV is then calibrated so as to classify the cells as correctly as possible as suitable or unsuitable for the species. The details of the function and of its calibration depend on the analysis.

Sampling the presence/absence data is a crucial part of the process. The sample must be unbiased to be representative of the whole population. Absence data in particular are often difficult to obtain accurately. A given location may be classified in the ''absence'' set because (1) the species could not be detected even though it was present (McArdle 1990, Solow 1993; for example, Kéry [2000] found that 34 unsuccessful visits were needed before one can assume with 95% confidence that the snake *Coronella austriaca* was absent from a given site), (2) for historical reasons the species is absent even though the habitat is suitable, or (3) the habitat is truly unsuitable for the species. Only the last cause is relevant for predictions, but ''false absences'' may considerably bias analyses.

Here we propose a new approach specifically designed to circumvent this difficulty. Requiring only

Manuscript received 27 March 2000; revised 4 March 2001; accepted 26 July 2001; final version received 12 November 2001.
[3] Corresponding author.
E-mail: Nicolas.Perrin@ie-zea.unil.ch

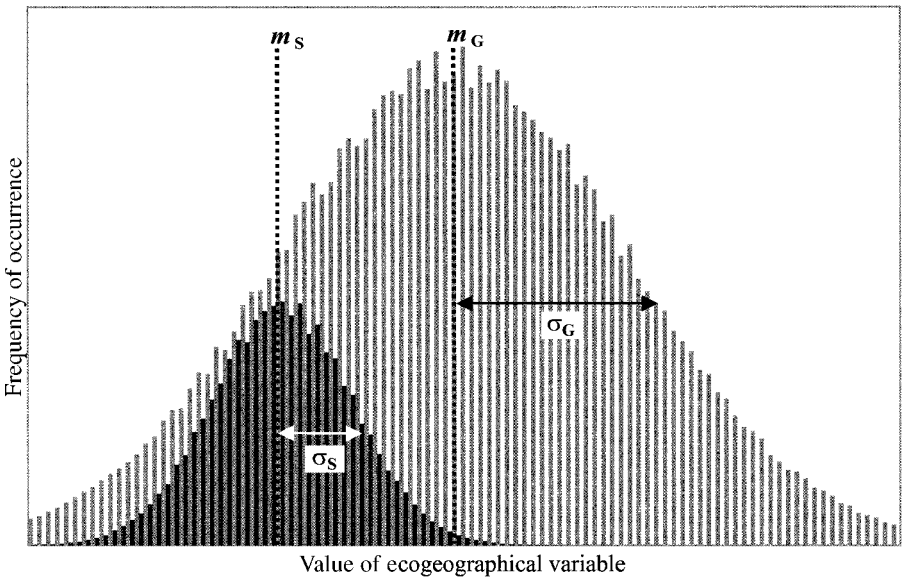

FIG. 1. The distribution of the focal species on any ecogeographical variable (black bars) may differ from that of the whole set of cells (gray bars) with respect to its mean ($m_S \neq m_G$), thus allowing marginality to be defined. It may also differ with respect to standard deviations ($\sigma_S \neq \sigma_G$), thus allowing specialization to be defined.

presence data as input, the Ecological-Niche Factor Analysis (ENFA) computes suitability functions by comparing the species distribution in the EGV space with that of the whole set of cells. In the present paper, we expose the concepts behind the ENFA, develop the mathematical procedures required (and implemented in the software Biomapper), and illustrate this approach through an habitat-suitability analysis of the Alpine ibex (*Capra ibex*).

MARGINALITY, SPECIALIZATION, AND THE ECOLOGICAL NICHE

Species are expected to be nonrandomly distributed regarding ecogeographical variables. A species with an optimum temperature, for instance, is expected to occur preferentially in cells lying within its optimal range. This may be quantified by comparing the temperature distribution of the cells in which the species was observed with that of the whole set of cells. These distributions may differ with respect to their mean and their variances (Fig. 1). The focal species may show some marginality (expressed by the fact that the species mean differs from the global mean) and some specialization (expressed by the fact that the species variance is lower than the global variance).

Formally, we define the marginality (M) as the absolute difference between global mean (m_G) and species mean (m_S), divided by 1.96 standard deviations (σ_G) of the global distribution

$$M = \frac{|m_G - m_S|}{1.96\sigma_G}. \quad (1)$$

Division by σ_G is needed to remove any bias introduced by the variance of the global distribution: a cell randomly chosen from a distribution is a priori expected to lie that much further from the mean as the variance in distribution is large. The coefficient weighting σ_G (1.96) ensures that marginality will be most often be between zero and one. Namely, if the global distribution is normal, the marginality of a randomly chosen cell has only a 5% chance of exceeding unity. A large value (close to one) means that the species lives in a very particular habitat relative to the reference set. Note that equation (1) is given here mainly to explain the principle of the method; the operational definition of marginality implemented in our software is provided by equation (10), which is a multivariate extension of (1).

Similarly, we define the specialization (S) as the ratio of the standard deviation of the global distribution (σ_G) to that of the focal species (σ_S),

$$S = \frac{\sigma_G}{\sigma_S}. \quad (2)$$

A randomly chosen set of cells is expected to have a specialization of one, and any value exceeding unity indicates some form of specialization. We reemphasize that specific values for these indexes are bound to depend on the global set chosen as reference, so that a species might appear extremely marginal or specialized on the scale of a whole country, but much less so on a subset of it.

Extending these statistics to a larger set of variables directly leads to Hutchinson's (1957) concept of the ecological niche, defined as a hyper-volume in the multidimensional space of ecological variables within which a species can maintain a viable population (Hutchinson 1957, Begon et al. 1996). The concept is

used here exactly in the same sense: by ecological niche we refer to the subset of cells in the ecogeographical space where the focal species has a reasonable probability to occur. This multivariate niche can be quantified on any of its axes by an index of marginality and specialization.

Some of these axes are obviously more interesting than others, and this is why a factor analysis is introduced. The reasons are actually double. First, ecological variables are not independent. As more and more are introduced in the description, multicollinearity and redundancy arise. One aim of factor analyses is to transform V correlated variables into the same number of uncorrelated factors. As these factors explain the same amount of total variance, subsequent analyses may be restricted to the few important factors (e.g., those explaining the largest part of the variance) without losing too much information.

Second, specialization is expected to depend on interactions among variables. For instance, the temperature one species prefers might vary with humidity. Species may thus specialize on a combination of variables, rather than on every variable independently. A factor analysis may allow extraction of the linear combinations of original variables on which the focal species shows most of its marginality and specialization. In Principal Component Analyses (Cooley and Lohnes 1971, Legendre and Legendre 1998), axes are chosen so as to maximize the variance of the distribution. In ENFA, by contrast, the first axis is chosen so as to account for all the marginality of the species, and the following axes so as to maximize specialization, i.e., the ratio of the variance in the global distribution to that in the species distribution.

Factor Extraction

Outline of the principles

We use raster maps, which are grids of N isometric cells covering the whole study area. Each cell of a map contains the value of one variable. Ecogeographical maps contain continuous values, measured for each of the V descriptive variables. Species maps contain boolean values (0 or 1), a value of 1 meaning that the presence of the focal species was proved on this cell. A value of zero simply means absence of proof.

Each cell is thus associated to a vector whose components are the values of the EGV in the underlying area, and can be represented by a point in the multidimensional space of the EGVs. If distributions are multinormal, the scatterplot will have the shape of a hyper-ellipsoid (Fig. 2). The cells where the focal species was observed constitute a subset of the global distribution and are plotted as a smaller hyper-ellipsoid within the global one. The first factor, or marginality factor, is the straight line passing through the centroids of the two ellipsoids. The species marginality is the distance between these centroids, standardized as in Eq. 1. Fig. 2 plots this step for a three-dimensional initial set.

To obtain the specialization factors, the reference system is changed in order to transform the species ellipsoid into a sphere, the variance of which equals unity in each direction. In this new metrics, the first specialization factor is the one that maximizes the variance of the global distribution (while orthogonal to the marginality factor). The other specialization factors are then extracted in turn, each step removing one dimension from the space, until all V factors are extracted. All specialization factors are orthogonal in the sense that the distribution of the species subset on any factor is uncorrelated with its distribution on the others. As factors are sorted by decreasing order of specialization, the first few (F) will thus generally contain most of the relevant information. Their small number and independence make them easier to use than the original EGVs, so that all following operations will be restricted to them. In particular, the suitability of any cell for the focal species (be it classified as 0 or 1 for observation data) will be calculated according to its position in the F-dimensional space.

Mathematical procedures

Ecogeographical variables are first normalized as far as possible, e.g., through Box-Cox transformation (Sokal and Rohlf 1981). Though multinormality is theoretically needed for factor extraction through eigensystem computation (Legendre and Legendre 1998), this method seems quite robust to deviations from normality (Glass and Hopkins 1984). EGVs are then standardized by retrieving means and dividing by standard deviations:

$$z_{ij} = \frac{x_{ij} - \bar{x_j}}{\sigma_{x_j}} \tag{3}$$

where x_{ij} is the value of the variable x_j in cell i, $\bar{x_j}$ the mean of this variable over all cells, and σ_{x_j} its standard deviation. Let $\mathbf{Z}$ be the $N \times V$ matrix of standardized measurements z_{ij}. The $V \times V$ covariance matrix among standardized variables is then computed as

$$\mathbf{R}_G = \frac{1}{N}\mathbf{Z}^T\mathbf{Z} \tag{4}$$

where $\mathbf{Z}^T$ is the transposed matrix of $\mathbf{Z}$. Because of standardization (Eq. 3), $\mathbf{R}_G$ is also a correlation matrix.

The N_S lines of $\mathbf{Z}$ corresponding to the N_S cells where the focal species was detected are then stored in a new $N_S \times V$ matrix (say $\mathbf{S}$), from which the $V \times V$ species covariance matrix is calculated:

$$\mathbf{R}_S = \frac{1}{N_S - 1}\mathbf{S}^T\mathbf{S}. \tag{5}$$

Note that in contrast to $\mathbf{R}_G$, $\mathbf{R}_S$ is not a correlation matrix, since standardization was performed on the global data set, not on the species subset.

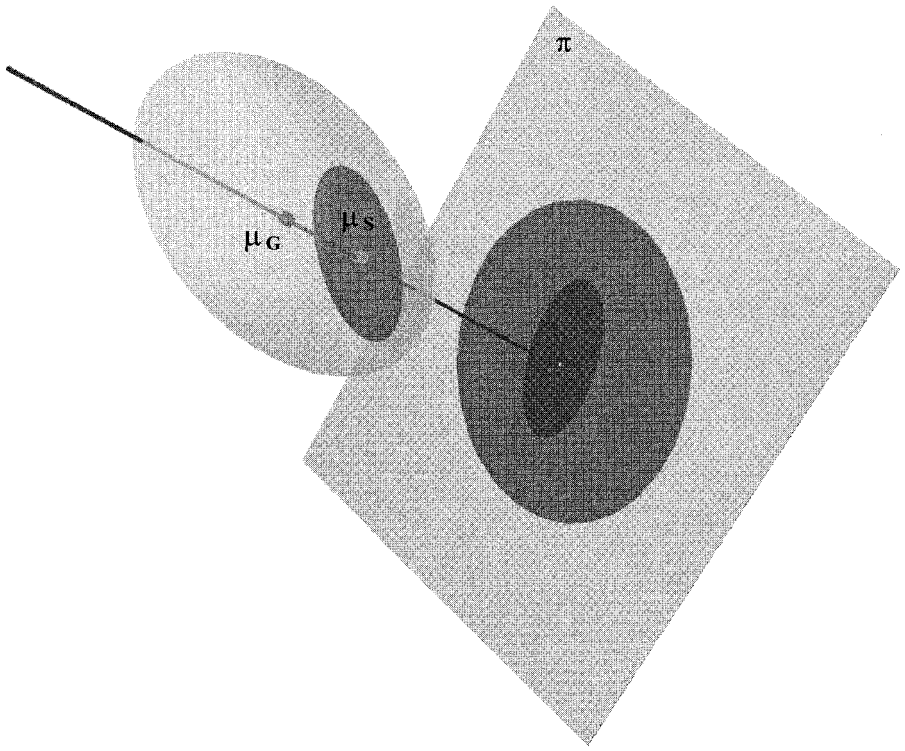

FIG. 2. Geometrical interpretation of the Ecological-Niche Factor Analysis. Square cells of the study area are represented in a three-EGV space. The larger, lighter balloon represents the global cloud of cells, while the smaller, darker balloon represents the subset of cells where the focal species was observed. The straight line passing through their centroids (μ_G and μ_S) is the marginality factor. In order to extract the variance associated with this factor, the cell coordinates are projected on a plane π perpendicular to it, thereby producing the two ellipses. In reality, those operations are typically conducted with 20–30 EGVs.

Let $\mathbf{u}$ be a normed vector of the EGV space. The variance of the global distribution on this vector is $\mathbf{u}^T\mathbf{R}_G\mathbf{u}$, while that of the species distribution is $\mathbf{u}^T\mathbf{R}_S\mathbf{u}$. The first specialization factor should thus maximize the ratio $\Theta(\mathbf{u}) = \mathbf{u}^T\mathbf{R}_G\mathbf{u}/\mathbf{u}^T\mathbf{R}_S\mathbf{u}$. However, this vector must also be orthogonal to the marginality factor $\mathbf{m}$, given as the vector of means over the V columns of $\mathbf{S}$:

$$\mathbf{m} = \left\{\frac{1}{N_S}\sum_{i=1}^{N_S} z_{ij}\right\}. \tag{6}$$

The problem therefore becomes that of finding the vector $\mathbf{u}$ that maximizes $\Theta(\mathbf{u})$ under the constraint $\mathbf{m}^T\mathbf{u} = 0$. This is equivalent to finding $\mathbf{u}$, such that

$$\begin{cases} \mathbf{u}^T\mathbf{R}_S\mathbf{u} = 1 \\ \mathbf{u}^T\mathbf{m} = 0 \\ \mathbf{u}^T\mathbf{R}_G\mathbf{u} \text{ max.} \end{cases} \tag{7}$$

A change in variables allows us to rewrite the problem

$$\begin{cases} \mathbf{v}^T\mathbf{v} = 1 \\ \mathbf{v}^T\mathbf{y} = 0 \\ \mathbf{v}^T\mathbf{W}\mathbf{v} \text{ max} \end{cases} \tag{8}$$

where $\mathbf{v} = \mathbf{R}_S^{1/2}\mathbf{u}$, $\mathbf{y} = \mathbf{z}/\sqrt{\mathbf{z}^T\mathbf{z}}$, and $\mathbf{z} = \mathbf{R}_S^{-1/2}\mathbf{m}$, $\mathbf{W} = \mathbf{R}_S^{-1/2}\mathbf{R}_G\mathbf{R}_S^{-1/2}$ is a symmetric matrix. It can be shown that the solution is given by the first eigenvector of

$$\mathbf{H} = (\mathbf{I}_V - \mathbf{y}\mathbf{y}^T)\mathbf{W}(\mathbf{I}_V - \mathbf{y}\mathbf{y}^T). \tag{9}$$

Indeed,

1) $\mathbf{y}$ is an eigenvector of $\mathbf{H}$ because $\mathbf{H}\mathbf{y} = (\mathbf{I}_V - \mathbf{y}\mathbf{y}^T)\mathbf{W}(\mathbf{I}_V - \mathbf{y}\mathbf{y}^T)\mathbf{y} = 0$;

2) $\mathbf{H}$ is symmetrical and thus admits a base of orthonormed eigenvectors so that $\mathbf{H}\mathbf{v} = \lambda\mathbf{v} \Rightarrow \mathbf{v}^T\mathbf{y} = 0$; and,

3) $\mathbf{v}^T\mathbf{H}\mathbf{v}$ is maximum for the first eigenvector, which also maximizes $\mathbf{v}^T\mathbf{W}\mathbf{v}$ since $\mathbf{v}^T\mathbf{y} = 0 \Rightarrow \mathbf{v}^T\mathbf{H}\mathbf{v} = \mathbf{v}^T(\mathbf{I}_V - \mathbf{y}\mathbf{y}^T)\mathbf{W}(\mathbf{I}_V - \mathbf{y}\mathbf{y}^T)\mathbf{v} = \mathbf{v}^T\mathbf{W}\mathbf{v}$.

The V eigenvectors of $\mathbf{H}$ are then back transformed, and the new eigenvectors ($\mathbf{u} = \mathbf{R}_S^{-1/2}\mathbf{v}$) are stored in a matrix $\mathbf{U}$. These vectors are $\mathbf{R}_S$-orthogonal (all $\mathbf{S}\mathbf{u}$ distributions have variance 1 and are uncorrelated). Furthermore, due to the constraint that $\mathbf{u}$ be orthogonal to $\mathbf{m}$, this system has one null eigenvalue. The corresponding eigenvector is thus deleted from $\mathbf{U}$, and $\mathbf{m}$ is substituted instead as the first column. It should be noted that, although all marginality is accounted for by the first factor, this factor is not ''pure,'' in that the niche of the focal species may also display some restriction on it, in addition to its departure from the

mean. The amount of specialization on this first axis is provided by the difference between the traces (sum of all eigenvalues) of **W** and **H**.

All these procedures are implemented in the software Biomapper (A. H. Hirzel, J. Hausser, and N. Perrin, University of Lausanne, Lausanne, Switzerland).

Interpretation of the factors

The coefficients m_i of the marginality factor express the marginality of the focal species on each EGV, in units of standards deviations of the global distribution. The higher the absolute value of a coefficient, the further the species departs from the mean available habitat regarding the corresponding variable. Negative coefficients indicate that the focal species prefers values that are lower than the mean with respect to the study area, while positive coefficients indicate preference for higher-than-mean values. An overall marginality M can be computed over all EGV as

$$M = \frac{\sqrt{\sum_{i=1}^{V} m_i^2}}{1.96} \tag{10}$$

so that the marginalities of different species within a given area can be directly compared.

The coefficients of the next factors receive a different interpretation: the higher the absolute value, the more restricted is the range of the focal species on the corresponding variable. Note that only absolute values matter here, since signs are arbitrary. The eigenvalue λ_i associated to any factor expresses the amount of specialization it accounts for, i.e., the ratio of the variance of the global distribution to that of the species distribution on this axis. Eigenvalues usually rapidly decrease from the second factor to the last one, so that only the first four or five axes are useful to compute habitat suitability. Different criteria may be used for the selection process, such as direct comparison with the broken-stick distribution, or threshold value for cumulative variance.

A global specialization index can be computed as

$$S = \frac{\sqrt{\sum_{i=1}^{V} \lambda_i}}{V} \tag{11}$$

and can be used for among-species comparisons, provided the same area is used as reference.

Habitat-Suitability Maps

A variety of methods can be envisaged to compute the suitability for the focal species of any cell from the study area. Among the several alternatives tested, the following approach turned out to be quite robust and was implemented in Biomapper. It builds on a count of all cells from the species distribution that lay as far or farther apart from the median than the focal cell on a factor axis. This count is normalized in such a way that the suitability index ranges from zero to one.

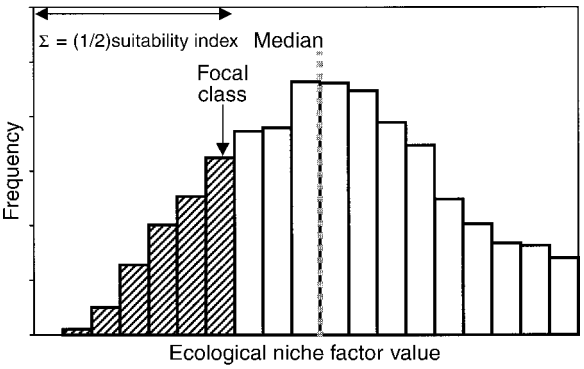

FIG. 3. The suitability of any cell from the global distribution is calculated from its situation (arrow) relative to the species distribution (histogram) on all selected niche factors. Specifically, it is calculated as twice the dashed area (sum of all cells from the species distribution that lie as far or farther from the median dashed vertical line) divided by the total number of cells from the species distribution (surface of the histogram).

Practically, this is performed by dividing the species range on each selected factor in a series of classes, in such a way that the median would exactly separate two classes (Fig. 3). For every cell from the global distribution, we count the number of cells from the species distribution that lay either in the same class or in any class farther apart from the median on the same side (Fig. 3). Normalization is achieved by dividing twice this number by the total number of cells in the species distribution. Thus, a cell laying in one of the two classes directly adjacent to the median would score one, and a cell laying outside the species distribution would score zero.

An overall suitability index of the focal cell can then be computed from a combination of its scores on each factor. In order to account for the differential ecological importance of the factors, we attribute equal weight to marginality and specialization, but, while all the marginality component goes to the first factor, the specialization component is apportioned among all factors proportionally to their eigenvalue (the marginality factor may thus take more than half of the weight if it also accounts for some specialization).

Repeating this procedure for each cell allows to produce a habitat-suitability map, where suitability values range from 0 to 1. To convert this quantitative (or semi-quantitative) map into a presence/absence one, a threshold value may be chosen, above which the cell will be considered as suitable. A validation data set can be used to find out the best threshold value, e.g., through the ROC plot method (Zweig and Campbell 1993, Fielding and Bell 1997), given some a priori cost values to each type of inferential error. Typically, since our method builds on presence data only, lower costs

TABLE 1. Nature and source of the 34 ecogeographical variables used in the Ecological-Niche Factor Analysis (ENFA) of ibex distribution.

Official database	Topic	Source†	Derived EGV
AS85R	cover use	OFS	frequency and proximity of rock, snow, forests, meadows, etc.
DHM	topography	OFS	altitude, slope, aspect, SD of altitude
GWN	hydrography	OFS	proximity of rivers and lakes
Vector 200	land map	OFT	proximity of villages, towns, railways, roads, etc.
SB	ibex colonies	OFEFP	calibrating and validation of presence data sets

†Abbreviations: OFT, Swiss Office of Topography; OFS, Swiss Office of Statistics; and OFEFP, Swiss Federal Office of Environment.

should be attributed to cells wrongly considered as suitable than to cells wrongly considered as unsuitable.

An Application to the Alpine Ibex

The alpine ibex (*Capra ibex*) was exterminated from the Swiss Alps in the last century due to excessive hunting. Reintroduction attempts starting in 1911 were highly successful, and colonies have since grown rapidly throughout the Swiss Alps. As a legally protected species, ibex populations were carefully monitored since their reintroduction, so that presence data are highly reliable. However, as their expansion is still hindered by the patchiness of their habitat, they presumably do not occupy yet all suitable regions. Absence data therefore do not necessarily reflect poor-quality habitat, a point which strongly advocates for the use of an analysis relying on presence data only.

The whole of Switzerland was chosen as reference area, and modeled as a raster map based on the Swiss Coordinate System (plane projection), comprising 4 145 530 square cells of 1 ha (100 × 100 m) each. We used 34 ecogeographical variables derived from governmental data bases (Table 1). Topographical data (altitude, slope, and aspect) were directly obtained as quantitative variables. Frequency and distance data were derived from boolean variables describing soil occupancy, as official data bases attribute each cell to one category only (snow, rocks, meadow, forest, building, etc), according to a regular sampling. Distance data express the distance between the focal cell and the closest cell belonging to a given category. Frequency describes the proportion of cells from a given category within a circle of 1200 m radius around the focal cell. Circle surface is $\sim$5 km^2, which corresponds to the mean area explored daily by individual ibexes (Abderhalden and Buchli 1997).

The presence database was a digitized map of ibex home ranges. The polygons drawn by fauna managers were converted in raster format at the same resolution as EGV maps. This raster was then randomly partitioned into two data sets, every cell having a 0.5 chance of belonging to each set. The first set (101 564 cells) was used to calibrate the model, and the other set (101 550 cells) to validate it. Application of the ENFA method to the calibration set provided an overall marginality of $M = 1.1$ and an overall specialization value of $S = 2.2$, showing that ibex's habitat differs drastically from the mean conditions in Switzerland, and that

TABLE 2. Variance explained by the first five (out of 34) ecological factors, and coefficient values for the 13 most important initial variables.

EGV	Marginality (46%)	Spec. 1 (11%)	Spec. 2 (8%)	Spec. 3 (6%)	Spec. 4 (4%)
Rock frequency	0.350	−0.105	−0.132	−0.186	−0.163
Grass frequency	0.321	−0.021	−0.044	−0.043	−0.039
Altitude	0.269	−0.365	−0.561	0.005	0.111
Distance to grass	−0.245	0.021	−0.096	−0.121	−0.073
Distance to agricultural meadows	0.231	−0.062	0.017	0.588	−0.784
Frequency of >30° slope	0.230	−0.035	0.074	0.011	0.004
Frequency of dense forest	−0.228	0.025	−0.727	0.037	−0.001
Distance to secondary roads	0.222	−0.012	−0.049	−0.129	0.115
Frequency of agricultural meadows	−0.212	−0.907	−0.011	0.236	−0.383
Distance to towns	0.209	0.012	0.049	−0.021	0.002
1 SD of altitude	0.204	0.000	−0.005	−0.032	−0.044
Distance to forests	0.203	0.015	0.165	−0.068	0.108
Distance to villages	0.200	0.003	0.017	−0.097	0.320

Notes: EGVs are sorted by decreasing absolute value of coefficients on the marginality factor. Positive values on this factor mean that ibex prefer locations with higher values on the corresponding EGV than the mean location in Switzerland. Signs of coefficient have no meaning on the specialization factors. The amount of specialization accounted for is given in parentheses in each column heading.

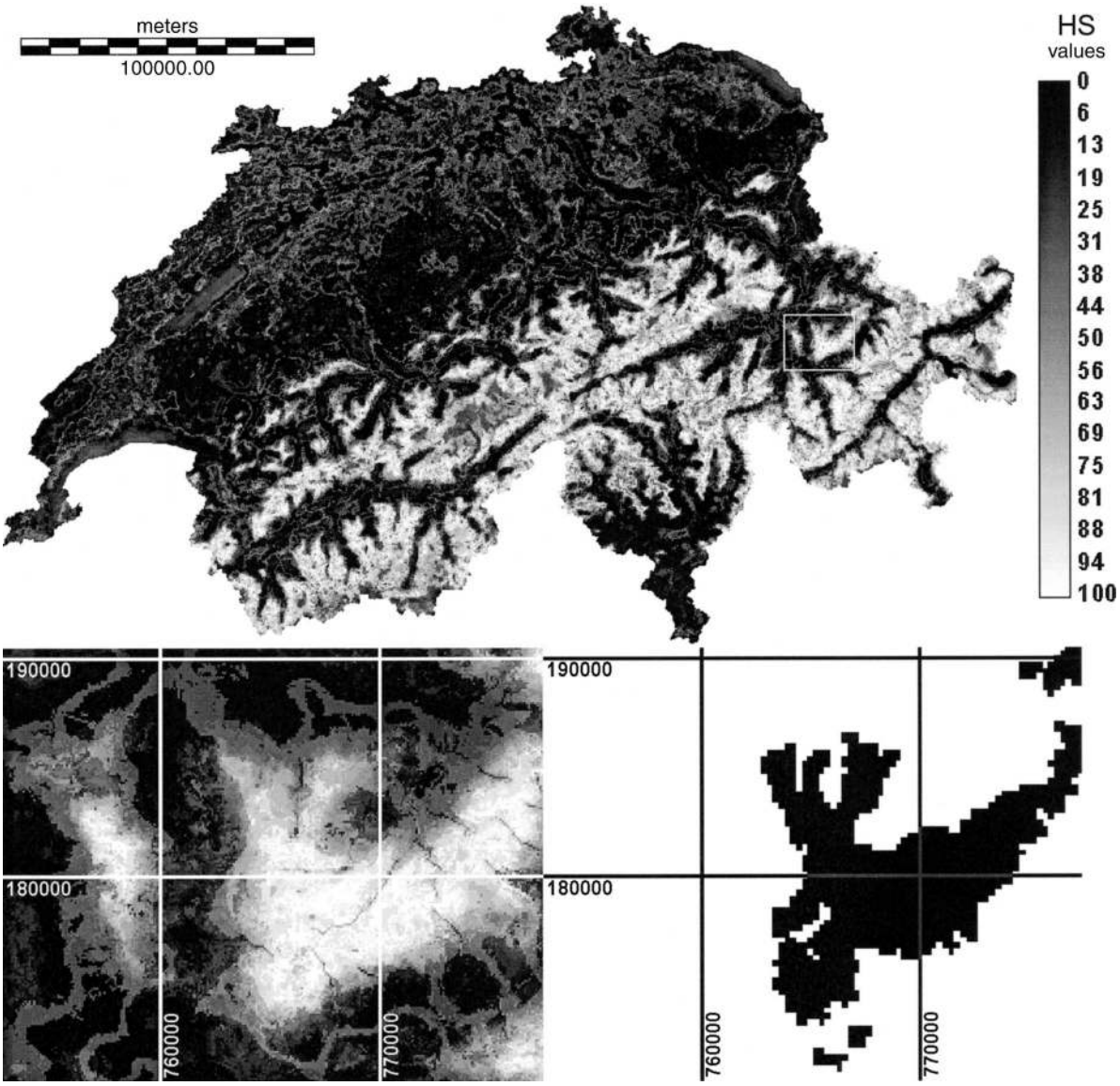

FIG. 4. Habitat-suitability map for alpine ibex in Switzerland, as computed from ENFA. The scale on the right, marked ''HS values,'' shows the habitat suitability values represented by each shade in the map. The inset is displayed at a larger scale at bottom left, with ibex presence data on the right. In the habitat-suitability map, light shading denotes areas more suitable for ibex, and dark shading denotes areas less suitable. In the ibex presence map at bottom right, dark shading denotes ibex presence. The largest suitable patch is indeed occupied, while the smaller one is not, being either too small or unreachable.

ibex are quite restrictive on the range of conditions they withstand. The five factors retained (out of the 34 computed) accounted for 74% of the total sum of eigenvalues (that is, 100% of the marginality and 74% of the specialization). The marginality factor alone accounted for 46% of this total specialization, a quite important value, meaning that ibex display a very restricted range on those conditions for which they mostly differ from background Switzerland conditions.

Marginality coefficients (Table 2) showed that ibexes are essentially linked to high-altitude, steep, and rocky slopes, rich in pastures (rock frequency = 0.35, altitude = 0.27, frequency of slopes $>30°$ = 0.23, grass frequency = 0.32, distance to grass = −0.24). By contrast, ibex tends to avoid forest (frequency = −0.23) and human activities (distance to secondary roads = 0.22, distance to agricultural meadows = 0.23). Aspect (northness, eastness) as well as snow and water (lakes, rivers) had only marginal effects. The very large eigenvalue (76.6) attributed to this first factor means that randomly chosen cells in Switzerland are ~80 times more dispersed on this axis than the cells were ibex was recorded. Or in other words, ibex are extremely sensitive to shifts from their optimal conditions on this

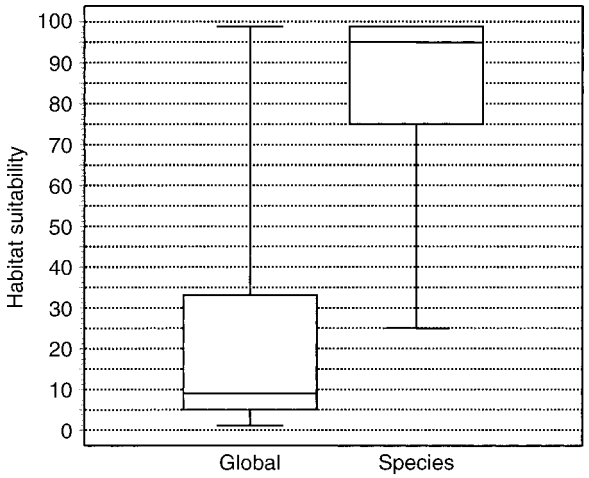

FIG. 5. Box plots presenting the distributions of the habitat-suitability values for the whole set of cells (left) and the validation subset (right). The latter is made of the 101 550 cells with ibex that were not included in the analysis. Boxes delimit the interquartile range, the middle line indicating the median; whiskers encompass the 80% confidence interval. The two distributions obviously differ, as the global one is mostly confined to low values (suitability, 5–33%), while the validation set concentrates on high-suitability values (75–99%).

axis. The next factors account for some more specialization, mostly regarding agricultural meadow frequency and altitude (second factor) as well as forest frequency (third factor), showing some sensitivity to shifts away from their optimal values on these variables.

A suitability map was built from these five factors for the whole of Switzerland, which is plotted on Fig. 4. Enlargement of a small part of it shows one suitable patch where ibex was indeed recorded (on the right), as well as one smaller suitable patch (on the left) where ibex was not recorded. This patch presumably was not colonized because of its isolation, or too small to sustain a viable colony on the long term. As false positives like this give no indication about the quality of our model, standard quality estimators like the kappa index (Monserud and Leemans 1992), which attribute the same importance to false positives and false negatives, cannot be used to validate it. Instead, we evaluated the distribution of suitability values of cells from the validation set. As shown in Fig. 5, these cells differ drastically from the global distribution. Predicted suitability exceeds 0.5 in 83% of cells, which differs highly significantly ($P < 0.0001$, bootstrap test) from the value of 24% expected if cells were randomly chosen from the global distribution.

DISCUSSION

Niche factors and distribution maps

The originality of the present approach lies in the fact that it builds on the concept of ecological niche, which is central to the whole field of ecology. A basic tenet of the niche theory is that fitness (or habitat suitability) does not bear monotonic relationships with conditions or resources, but instead decreases from either side of an optimum. In this respect, our approach differs fundamentally from other techniques like discriminant functions or first-order regressions, where relationships are assumed linear and monotonic. Accordingly, our analysis directly provides two key measurements regarding the niche of the focal species, namely those of marginality and of specialization. Outputs thus have intuitive ecological meaning, and allow direct comparisons with the niche of different species. Application of the analysis to evaluate, e.g., species packaging or niche-overlap measurement among members of a guild, would be straightforward extensions of the present approach (e.g., Dolédec et al. 2000)

Our application to ibex data, for instance, provides quantitative estimates of marginality and specialization for this species which evidence its very peculiar ecological requirements. Furthermore, interpretation of the factors in terms of the EGVs turns out to be very consistent with the experience of field specialists. In particular, the EGVs that correlate with the marginality factor are precisely those most often cited as particularly relevant for ibex ecology (Rauch 1941, Nievergelt 1966, Nievergelt and Zingg 1986, Hausser 1995, Hindenlang and Nievergelt 1995).

These results obviously suffer from the same caveat as any inferential approach: a variable might turn out to correlate with one of the main axes not because of its intrinsic importance, but because it correlates strongly with another crucially important variable. ENFA is a purely descriptive method and cannot extract causality relations. Nonetheless, it provides (at worst) important cues about preferential conditions, and remains a powerful tool to draw potential habitat maps.

In this respect, a limitation of our software is that it does not yet include confidence intervals on distribution maps. Increasingly, conservation managers are demanding risk analyses that incorporate uncertainties in model predictions. These could clearly be obtained through the bootstrapping of presence data. Though not yet implemented in Biomapper, this procedure will certainly provide an important and useful extension.

A second limitation, less easy to deal with, is that ENFA only handles linear dependencies within the species niche. Multiplicative or nonlinear interactions cannot be accommodated in the present state, except through transformations or nonlinear combinations of the original ecogeographical variables.

A third limitation is that some EGVs may turn out to be constant in $\mathbf{S}$, or in linear combination with other EGVs, which makes $\mathbf{R}_S$ singular. This is likely to happen with coarsely measured data or small species data sets. Whenever this happens, Biomapper identifies the constant or correlated EGVs so that the user can remove (one of) them from the analysis. An alternative ap-

proach would obviously consist of improving the field sample, either by increasing the presence data set or by measuring EGVs on a finer scale.

Finally, a last important point to emphasize again is that our approach characterizes ecological niches relative to a reference area. Marginality and specialization are thus bound to depend on the geographic limits of the study area. Some species may turn out to occur at the very edge of their distribution, and may thus appear quite specialized in the reference set, however widespread they might be otherwise. Reciprocally, ibex would have appeared much less marginal and specialized had our sampling area been restricted to the Alps. The same distinction must be applied here as the one made between fundamental and realized niches (Hutchinson 1957). Our analysis does not investigate fundamental niches, but only their specific realization within a given geographical context.

ENFA vs. logistic regressions

With respect to more standard techniques, a crucial advantage of ENFA is that it does not require absence data. Presence data are compared instead with background environment. This of course implies that presence data should be unbiased samples of actual distributions, which we suspect might not be the case of many available database, since sampling efforts are frequently biased with respect to environment. However, though this problem is difficult to circumvent, the point must also be made that database often simply lack any absence data. And when available, these may turn out to be either unreliable (in the case of cryptic or poorly known species) or meaningless (in the case of invading species, or those living in fragmented habitats where some patches have become extinct). As many species enter one of these categories, our approach potentially has a wide application range. In particular, predictions about the expected expansion of invading species seem a promising test bed.

In the case of stable populations from well-known species, one might prefer more classical approaches such as logistic regressions, able to extract relevant information from absence or abundance data. This point deserves proper investigation, in order to localize the threshold where the benefits gained from incorporating absence data are compensated by the costs induced by their possible poor quality. Investigations are presently in progress to compare the power of ENFA to that of classical logistic regression analyses under different biological and sampling scenarios, in order to assess their respective advantages and inconveniences. Preliminary results show ENFA to be more robust than classical logistic regressions with respect to several habitat-occupancy scenarios (Hirzel et al. 2001).

Finally, the point must also be made that the procedures used by standard stepwise analyses to select significant EGVs from the original set turn out to be highly sensitive to the algorithms chosen, as well as to the input order. Consequences are that (1) many trials are needed in order to sort out the ''best'' model, and (2) variables that bear a causal relationship to the focal species' presence might well be lost in the process, if other EGVs present spurious correlations. This implies some subjective choices, and requires a good a priori knowledge of the focal species' ecology. In contrast, our factor analysis does not reject any input EGV, but only weights them. The subjective components and a priori knowledge required are thereby kept minimal, and correlations among variables and axes are immediately visible and interpretable.

Acknowledgments

This research was supported by the Swiss Federal Office for Environment, Forest and Landscape (OFEFP), grant 0310.3600.305. We warmly thank H. J. Blankenhorn for his enthusiastic endorsement of the Ibex Project, and all people who tested Biomapper, particularly P. Patthey for his thorough check and judicious suggestions.

Literature Cited

Abderhalden, W., and C. Buchli. 1997. Steinbockprojekt Albris/SNP, Schlussbericht. ARINAS and FORNAT, Zernez (Graubunden), Switzerland.

Akçakaya, H. R., and J. L. Atwood. 1997. A habitat-based metapopulation model of the California Gnatcatcher. Conservation Biology **11**:422–434.

Akçakaya, H. R., M. A. McCarthy, and J. L. Pearce. 1995. Linking landscape data with population viability analysis: management options for the helmeted honeyeater *Lichenostomus melanops cassidix*. Biological Conservation **73**: 169–176.

Akçakaya, H. R., and M. G. Raphael. 1998. Assessing human impact despite uncertainty viability of the northern spotted owl metapopulation in the northwestern USA. Biodiversity and Conservation **7**:875–894.

Begon, M., J. L. Harper, and C. R. Townsend. 1996. Ecology. Blackwell Science, Oxford, UK.

Clark, J. D., J. E. Dunn, and K. G. Smith. 1993. A multivariate model of female black bear habitat use for a geographic information system. Journal of Wildlife Management **57**:519–526.

Cooley, W. W., and P. R. Lohnes. 1971. Multivariate data analysis. John Wiley, New York, New York, USA.

Dolédec, S., D. Chessel, and C. Gimaret-Carpentier. 2000. Niche separation in community analysis: a new method. Ecology **81**:2914–2927.

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.

Glass, G. V., and K. D. Hopkins. 1984. Statistical methods in education and psychology. Second edition. Prentice Hall, London, UK.

Hausser, J. 1995. Mammifères de Suisse. Birkhäuser, Bâle, Switzerland.

Higgins, S. I., D. M. Richardson, M. C. Richard, and T. H. Trinder-Smith. 1999. Predicting the landscape-scale distribution of alien plants and their threat to plant diversity. Conservation Biology **13**:303–313.

Hindenlang, K., and B. Nievergelt. 1995. *Capra ibex* L., 1758. Pages 450–456 *in* J. Hausser, editor. Mammifères de Suisse. Birkhäuser, Basel, Switzerland.

Hirzel, H., V. Helfer, and F. Metral. 2001. Assessing habitat-suitability models with a virtual species. Ecological Modelling **145**:111–121.

Hutchinson, G. E. 1957. Concluding remarks. Cold Spring Harbour Symposium on Quantitative Biology **22**:415–427.
Jongman, R. H. G., C. J. F. ter Braak, and O. F. R. Van Tongeren. 1987. Data analysis in community and landscape ecology. Cambridge University Press, Cambridge, UK.
Kéry, M. 2000. Ecology of small populations. Dissertation. University of Zürich, Zürich, Switzerland.
Legendre, L., and P. Legendre. 1998. Numerical ecology. Second English edition. Elsevier Science BV, Amsterdam, The Netherlands.
Livingston, S. A., C. S. Todd, W. B. Krohn, and R. B. Owen. 1990. Habitat models for nesting bald eagles in Maine. Journal of Wildlife Management **54**:644–657.
Manel, S., J. M. Dias, S. T. Buckton, and S. J. Ormerod. 1999. Alternative methods for predicting species distribution: an illustration with Himalayan river birds. Journal of Applied Ecology **36**:734–747.
McArdle, B. H. 1990. When are rare species not there? Oikos **57**:276–277.
Mladenoff, D. J., R. C. Haight, T. A. Sickley, and A. P. Wydeven. 1997. Causes and implications of species restoration in altered ecosystems: a spatial landscape projection of wolf population recovery. Bioscience **47**:21–31.
Mladenoff, D. J., T. A. Sickley, R. G. Haight, and A. P. Wydevens. 1995. A regional landscape analysis and prediction of favorable Gray Wolf habitat in the northern Great Lakes region. Conservation Biology **9**:279–294.
Monserud, R. A., and R. Leemans. 1992. Comparing global vegetation maps with the Kappa statistic. Ecological Modelling **62**:275–293.
Nievergelt, B. 1966. Der Alpensteinbock (*Capra ibex L.*) in seinem Lebensraum: ein ökologischer Vergleich verschiedener Kolonien. Mammalia depicta, Hamburg, West Germany.
Nievergelt, B., and R. Zingg. 1986. *Capra ibex* Linnaeus, 1758—Steinbock. Pages 384–404 *in* J. Niethammer and F. Krapp, editors. Handbuch der Säugetiere Europas: Paarhufer. AULA-Verlag, Wiesbaden, Germany.
Özesmi, S. L., and U. Özesmi. 1999. An artificial neural network approach to spatial habitat modelling with interspecific interaction. Ecological Modelling **116**:15–31.
Palma, L., P. Beja, and M. Rodgrigues. 1999. The use of sighting data to analyse Iberian lynx habitat and distribution. Journal of Applied Ecology **36**:812–824.
Peeters, E. T. H., and J. J. P. Gardeniers. 1998. Logistic regression as a tool for defining habitat requirements of two common gammarids. Freshwater Biology **39**:605–615.
Rauch, A. 1941. Le bouquetin dans les Alpes. Payot, Paris, France.
Roloff, G. J., and J. B. Haufler. 1997. Establishing population viability planning objectives based on habitat potentials. Wildlife Society Journal **25**:895–904.
Sanchez-Zapata, J. A., and J. F. Calvo. 1999. Raptor distribution in relation to landscape composition in semi-arid Mediterranean habitats. Journal of Applied Ecology **36**: 254–262.
Sokal, R. R., and F. J. Rohlf. 1981. Biometry: the principles and practice of statistics in biological research. W.H. Freeman, New York, New York, USA.
Solow, A. R. 1993. Inferring extinction from sighting data. Ecology **74**:962–964.
Spitz, F., and S. Lek. 1999. Environmental impact prediction using neural network modelling: an example in wildlife damage. Journal of Applied Ecology **36**:317–326.
ter Braak, C. J. F., and C. W. N. Looman. 1987. Regression. Pages 29–77 *in* R. H. G. Jongman, C. J. F. ter Braak, and O. F. R. Van Tongeren, editors. Data analysis in community and landscape ecology. Cambridge University Press, Cambridge, UK.
Zweig, M. H., and G. Campbell. 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clinical Chemistry **39**:561–577.