ELSEVIER

Ecological Modelling 145 (2001) 111–121

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

Assessing habitat-suitability models with a virtual species

A.H. Hirzel *, V. Helfer, F. Metral

Laboratory for Conservation Biology, Institute of Ecology, University of Lausanne, CH-1015 Lausanne, Switzerland

Received 22 November 2000; received in revised form 9 May 2001; accepted 22 May 2001

Abstract

This paper compares two habitat-suitability assessing methods, the Ecological Niche Factor Analysis (ENFA) and the Generalised Linear Model (GLM), to see how well they cope with three different scenarios. The main difference between these two analyses is that GLM is based on species presence/absence data while ENFA on presence data only. A virtual species was created and then dispatched in a geographic information system model of a real landscape following three historic scenarios: (1) spreading, (2) at equilibrium, and (3) overabundant species. In each situation, the virtual species was sampled and these simulated data sets were used as input for the ENFA and GLM to reconstruct the habitat suitability model. The results showed that ENFA is very robust to the quality and quantity of the data, giving good results in the three scenarios. GLM was badly affected in the case of the spreading species but produced slightly better results than ENFA when the species was overabundant; at equilibrium, both methods produced equivalent results. The use of a virtual species proved to be a very efficient method, allowing one to fully control the quality of the input data as well as to accurately evaluate the predictive power of both analyses. © 2001 Published by Elsevier Science B.V.

Keywords: Habitat suitability model; Ecological Niche Factor Analysis; Generalised Linear Model; Simulated data; Geographic information system; False absences; Model comparison

1. Introduction

Prediction of species distribution is an important element of conservation biology. Management for endangered species (Palma et al., 1999; Sanchez-Zapata and Calvo, 1999), ecosystem restoration (Mladenoff et al., 1997), species re-introductions (Breitenmoser et al., 1999), population viability analyses (Akçakaya et al., 1995; Akçakaya and Atwood, 1997) and human–wildlife conflicts (Le Lay et al., 2001) often rely on habitat-suitability modelling. Multivariate models are commonly used to define habitat suitability and, combined with geographical information systems (GIS), allow one to create potential distribution maps (Guisan and Zimmermann, 2000).

Numerous multivariate analyses were developed for building habitat suitability or abundance models, but very few studies compare their predictive power (for example, Lek et al., 1996; Paruelo and Tomasel, 1997; Guisan et al., 1999; Manel et

* Corresponding author. Tel.: +41-21-692-4176; Fax: +41-21-692-4105.

E-mail address: alexandre.hirzel@ie-zea.unil.ch (A.H. Hirzel).

0304-3800/01/$ - see front matter © 2001 Published by Elsevier Science B.V.

PII: S0304-3800(01)00396-9

al., 1999; Özesmi and Özesmi, 1999). In this paper, we compare a common method, the Generalised Linear Model (GLM) (for example, Austin et al., 1984; Augustin et al., 1996; Guisan et al., 1998), with the Ecological Niche Factor Analysis (ENFA), a new multivariate analysis (Hirzel et al., in press).

GLM is a generalisation of multiple regression analysis with a binomial distribution and logistic link that may fit polynomials of higher degree than linear. The dependent variable (presence/absence of the species) is explained by a sum of weighted ecogeographical predictors. The weights are tuned in order to generate the best fit between the model and the calibration data set (Jongman et al., 1987; Nicholls, 1989).

ENFA compares the ecogeographical predictor distribution for a presence data set consisting of locations where the species has been detected with the predictor distribution of the whole area. Like the Principal Component Analysis, ENFA summarises all predictors into a few uncorrelated factors retaining most of the information. But in this case, the factors have an ecological meaning: the first factor is the 'marginality', and reflects the direction in which the species niche mostly differs from the available conditions in the global area. Subsequent factors represent the 'specialisation'. They are extracted successively by computing the direction that maximises the ratio of the variance of the global distribution to that of the species distribution. A large part of the information is accounted for by a few of the first factors. The species distribution on these factors is used to compute a habitat suitability index for any set of descriptor values (Hirzel et al., in press).

Practically, the main difference between these analyses is the quality of input data: GLM needs presence/absence data, whereas ENFA only needs presence data. The latter is thus much less demanding than the former and it is interesting to compare their predictive power. Obviously, this power depends on the situation: for example, when absence data are reliable GLM could get extra power by using this information, but in other situations it could be misled by false absences (McArdle, 1990; Solow, 1993; see also 'stochastic zeros' in Welsh et al., 1996).

The goal of this paper is thus to circumscribe the domain of application of both methods from the point of view of absence data quality. It is more complex a task than simply comparing analyses on the same data set. Indeed, measuring their sensitivity to various data qualities entails exploring several distribution patterns of ecologically identical species in a common landscape. But as such species could not live simultaneously in a same place, it is impossible to find such data in the real world; it is therefore necessary to generate simulated species distribution data. Moreover, this method presents the following advantages: (1) the input data set can be fully controlled, qualitatively as well as quantitatively; and (2) the 'reality' being perfectly known, model accuracy assessment is straightforward and certain. Nevertheless, in order to track reality as closely as possible, the environmental predictors were taken from a real area in the Swiss Alps.

2. Methods

This study implied to build a virtual species completely characterised by its ecological niche, which would be modelled by a 'truth' habitat suitability map. Three data sets were then generated, simulating three different scenarios. These data sets, in conjunction with environmental variables, were fed into the GLM and ENFA analyses, which produced 'predicted' habitat suitability maps. Finally, resulting models were evaluated by statistically comparing each 'predicted' map with the 'truth' map. These steps (summarised in Fig. 1) will now be developed in full detail.

2.1. Ecogeographical variables

Although a virtual species is used, environmental data are real and issued from a square region of 25.6×25.6 km^2 located in the Swiss Alps (see Fig. 4) numerically modelled by 17 GIS raster maps of 256×256 cells, representing 17 ecogeographical variables. These predictors are topographical, ecological or related to human activities (see Table 1). They were derived from land-cover, topography, hydrography and road/

rail network GIS databases. Boolean variables were transformed into continuous ones by computing a new map, storing either the distance to the nearest cell of this category or the proportion of those cells within a circular moving-window of 1200-m radius. We used IDRISI 2.0 (Eastman, 1997) and BIOMAPPER 1.0 (Hirzel et al., 2001) to achieve these operations and to deal with the predictor maps.

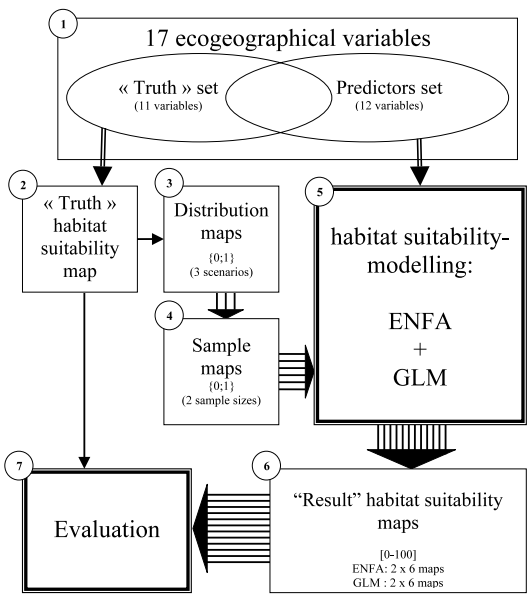

Fig. 1. Flow chart summarizing the steps involved in the study. (1) Seventeen ecogeographical variable (predictor) maps are prepared. (2) A 'truth' set of predictors is used to generate a 'truth' habitat suitability map. (3) On this basis, three distribution maps are generated corresponding to three distribution scenarios (see text). (4) The distribution maps are sampled with two sample sizes (300 and 1200 points), generating sample maps. (5) A 'predictors' set (partially overlapping the 'truth set'; see text) of the predictors is used in conjunction with the sample maps to compute predicting models with both GLM and ENFA methods. (5) These models are used to produce 'result' habitat suitability maps (6). (7) The 'result' maps are statistically compared with the 'truth' map to assess the predicting power of each analysis in each scenario. Single-framed boxes symbolise sets of maps; the type of data is indicated. Double-framed boxes symbolise statistical processes. Arrows symbolise the data flow, each shaft accounting for one map. See text for further explanations.

Table 1
Ecogeographical variables (predictor) used to generate the virtual habitat suitability map[a]

Predictor	Niche function	Weight (w_i)
Forest frequency	Linear (increasing)	6
Elevation	Gaussian	5
Southern aspect frequency	Linear (increasing)	2
Distance to towns	Truncated linear	2
Distance to forests	Linear (decreasing)	1
Slope $>30°$ frequency	Linear (decreasing)	1
Distance to waters	Linear (decreasing)	1
Distance to villages	Gaussian	1
Distance to primary roads	Truncated linear	1
Distance to secondary roads	Truncated linear	1
Distance to railways	Truncated linear	1

[a] Following Eq. (1), the global habitat suitability value is computed by a weighted average of the partial niche coefficients, which are themselves computed from the predictors by niche functions (cf. Fig. 1). This table indicates for each predictor (first column) which type of niche function was used to computes its partial niche coefficient (second column) and which was its weight for the global habitat suitability computation (third column).

2.2. Virtual ecological niche: the 'truth' habitat suitability map

On this spatial canvas, the virtual species was generated by creating a simulated ecological niche in an n-dimensional space, sensu Hutchinson (1957). It was modelled by a niche coefficient H ($H \in [0,1]$), which can be viewed as a probability of each cell to belong to the niche; note that H is de facto a habitat suitability index. This value was built as summarised in Eq. (1).

$$H = \frac{1}{\sum w_i} \sum w_i H_i + \varepsilon \quad (1)$$

where H is the habitat suitability of the focal cell, H_i is the value of the ith partial niche coefficient, w_i is the weight assigned to the ith partial niche coefficient, and ε is a random value.

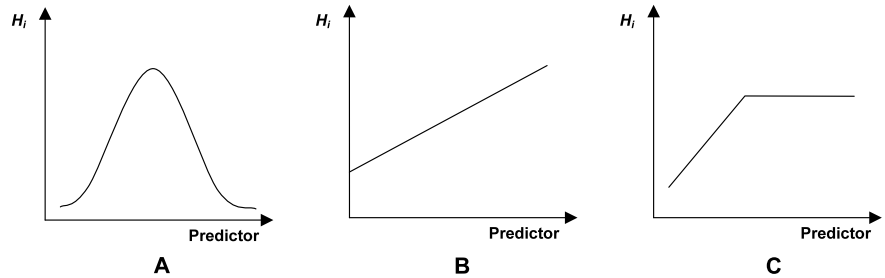

Fig. 2. The partial niche coefficient H_i ($H_i \in [0,1]$) is a function of each ecogeographical variable (predictor). Three types of function model three types of niche optimum: (A) the optimal value of the predictor lies somewhere in the middle of the available range and decreases 'gaussianly' in either direction. (B) and (C) are typically used to model distance related variables, either to disturbance or food sources. (B) The habitat suitability is linearly increasing (or decreasing) as the location goes farther from the source. (C) The truncated linear type shows a buffer zone effect, the influence of the source becoming null above some given range.

Global habitat suitability is composed of a weighted average of partial niche coefficients (H_i) and a stochastic coefficient (ε). The partial niche coefficients are the habitat suitability engendered by each predictor value; they were computed from 11 predictors (playing the role of Hutchinsonian environmental-space dimensions), picked out of the 17 available predictors, by 11 niche functions (Table 1). Three types of functions were used to model three types of environmental optimum: (1) a gaussian function modelled a median optimum, (2) a linear function modelled an extreme optimum, and (3) a truncated linear function modelled a buffer zone effect (see Fig. 2).

Each of these H_i values was then weighted by a w_i factor and the global niche coefficient calculated as their weighted average. Finally, a random term ε, generated from a uniform distribution in the range $[-0.05, 0.05]$, was added.

The niche-function parameters and the weights were arbitrarily tuned in order to generate about 50% of cells with $H > 0.5$.

This produces the 'truth' habitat suitability map (Fig. 3), representing the 'real' intrinsic preferences of our virtual species. By 'truth' map, we are meaning that it represents the kind of information usually unreachable by ecologists, the information they are trying to reveal through field sampling and statistical analysis. The 'truth' map will be constantly used as a basis to generate data and as a reference to assess the accuracy of habitat suitability analyses. A 3–D view of the landscape is presented in Fig. 4 to provide a better understanding.

2.3. Distribution maps

Computed on the basis of the 'truth' map, the distribution maps give the 'truth' presence/absence of the virtual species, information usually unavailable to field ecologists.

Three distribution scenarios were addressed in order to determine the advantages and drawbacks

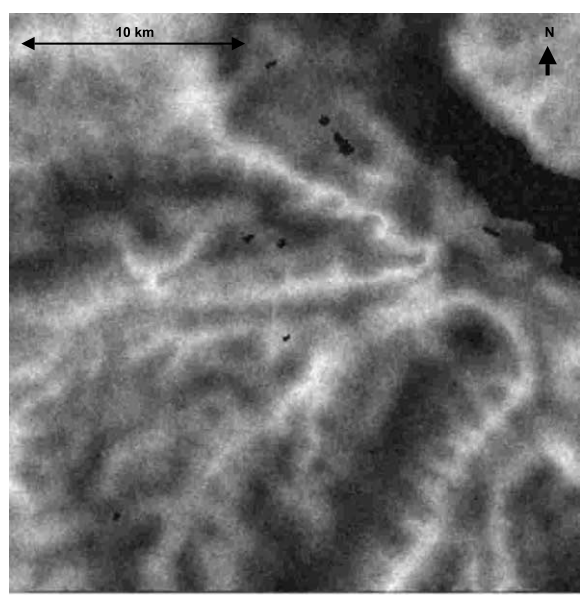

Fig. 3. The 'truth' habitat suitability map generated to model the ecological niche of the virtual species. High suitability areas are indicated by white pixels.

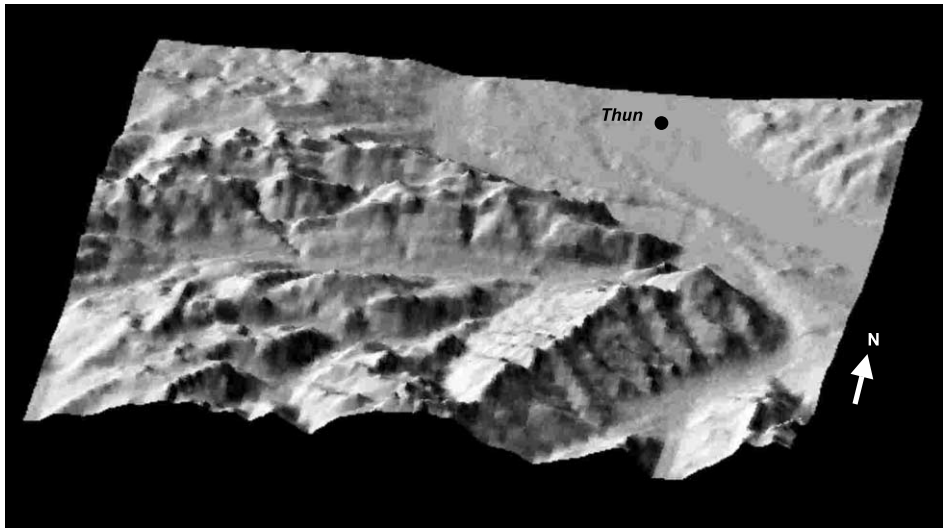

Fig. 4. Three-dimensional view of the studied region, a 25.6×25.6 km^2 area in the Bern Alps (Switzerland). The landscape is viewed from the south. In the north-east corner lie the Aare valley and Thun town; the flat area is the lake of Thun. In the middle of the landscape lies the west–east orientated Simmental valley. Elevations range from 551 to 2637 m above sea level.

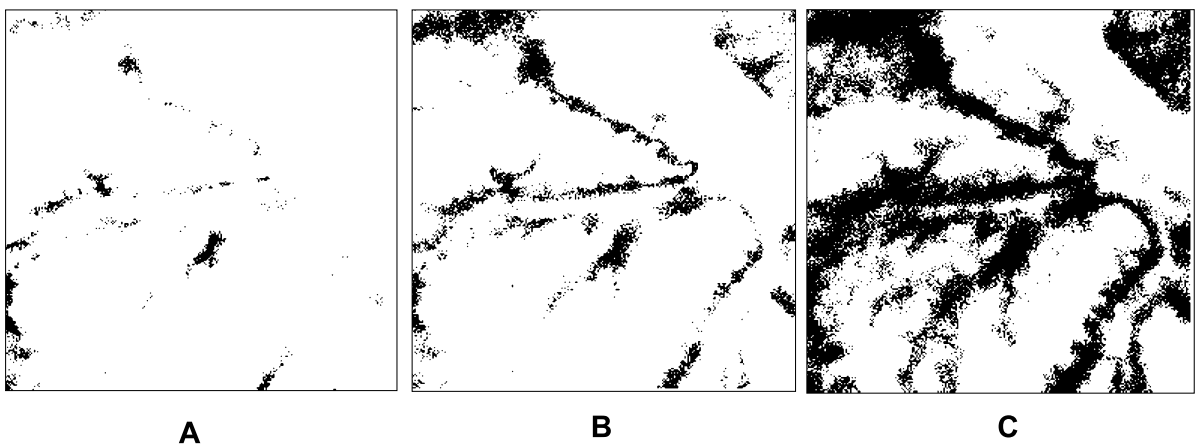

Fig. 5. Distribution maps of the virtual species for three colonisation scenarios. Black points are the cells where the species is present and the white ones are those where it is absent. Map A represents the 'spreading' scenario: the species entered the area from the southwest and is currently propagating in all directions, settling down in the most suitable areas. Map B shows the 'equilibrium' scenario in which the species occupies uniformly all the suitable areas. Map C presents the 'overabundance' scenario in which very high densities force the species to occupy less adequate areas.

of each habitat suitability analysis. They can be viewed as three historical phases of colonisation— the fundamental niche does not change but the realised one does: (1) a 'spreading phase' showing a density gradient from the south-west corner of the map to the north-east corner, (2) an 'equilibrium phase' where the species is abundant enough to occupy all the available suitable areas, and (3) an 'overabundance phase' where the species is so numerous that it has to spread in less suitable areas. (Fig. 5).

The 'equilibrium' distribution map was computed as follows. To each cell of the 'truth' habitat suitability map was added a random value taken in the range $[-0.2, 0.2]$ (uniform distribution); this was made in order to introduce some stochasticity in the model. If the resulting habitat suitability coefficient was larger than 0.7, the cell was marked as occupied.

The 'overabundance' distribution map was computed in a similar way but with a 0.5 habitat suitability threshold to simulate the overflowing density.

The 'spreading' distribution needed an additional operation: each cell of the 'truth' habitat suitability map was beforehand multiplied by a value decreasing in $1/d^2$, d being the distance to a point arbitrarily placed south-westward of the south-west corner of the map. This gradient function was tuned to produce values ranging from 0 to 1, 0.5 lying approximately in the middle of the map. This new gradient map was then submitted to the same operations as the 'equilibrium' scenario (habitat suitability threshold = 0.7).

This generating method assured to obtain distribution maps with a presence density correlated with area suitability.

2.4. Sample maps

These distribution maps were then used to simulate 'field' sampling data usually resulting from the trapping/detecting/observation activities of field biologists. As the GLM and the ENFA do not need the same kind of data, it was necessary to generate two data sets: one presence/absence set for the GLM and one presence set only for the ENFA. In order to compare results, sampling sizes were identical for all scenarios and analyses. Two sampling sizes were addressed, 300 points and 1200 points.

ENFA data sets were generated by randomly picking points in the distribution maps in order to obtain the targeted sample size. The probability to pick one cell in a given area was correlated with its density, which was variously correlated with its suitability depending on the scenario. The 'spreading' scenario had only 418 occupied cells and it was therefore impossible to get the 1200 points sample size in this case.

The same presence points were used for the GLM but additional absence data were generated as follows. In order to take into account the spatial auto-correlation of the predictors, a seven-cell-radius circular buffer was drawn around each presence point; the absence points were then randomly drawn from the area out of these buffers (a procedure similar to that used by Akçakaya and Atwood, 1997).

The number of GLM presence/absence points was thus the double of the number of ENFA presence points, but the 'field' data (presence) were identical.

2.5. Result habitat suitability maps

The simulated data sets were then submitted as dependent variables to the GLM and the ENFA. The independent variables were a set of 12 predictors out of the 17 available; six of them were arbitrarily taken among those used to generate the 'truth' habitat suitability map (elevation, southern aspect frequency, distance to towns, distance to forests, slope $> 30°$ frequency, and distance to primary roads), and six other were new (distance to rivers, distance to lakes, distance to pastures, distance to agricultural meadows, rock frequency and bush frequency).

ENFA was entirely performed with the BIOMAPPER software (Hirzel et al., 2001). The predictors were first normalised by the Box–Cox algorithm (Sokal and Rohlf, 1981). Ecological niche factors were then computed on these normalised predictors and ENFA provided one marginality factor and 11 tolerance factors totally uncorrelated, each factor being a linear combination of the predictors. Among these factors we kept only those explaining a significant amount of total variance by comparison with a broken-stick distribution (always greater than 75%). Factor distributions were computed on six classes and these empirical distributions used to compute the habitat suitability maps (for detailed information of this method, see Hirzel et al., in press). A 7×7-gaussian filter was finally applied to them in order to smooth the step shape produced by this analysis.

The GLM were calibrated in the S-PLUS software (Mathsoft Inc.) using a binomial distribution and a stepwise variable selection procedure. As the niche coefficient was not a linear function of the predictors, we introduced in the input variables not only the 12 predictors already mentioned, but also their square power; bell shaped and truncated linear niche functions could thus be modelled with satisfying accuracy. Because of the high sensitivity of the stepwise process — which eliminates a part of the input predictors to retain only the most relevant — to the input order of the predictors, we tried several orders, retaining the model that explained the highest proportion of the variance. The habitat suitability was then expressed as a linear combination of the predictors and their square terms. The model was then implemented in the GIS and maps were produced that had finally to be transformed by the inverse logistic function to be scaled between 0 and 1 (for further explanations, see for example Guisan et al. (1999)).

2.6. Evaluation

The accuracy of the 'result' habitat suitability maps had finally to be assessed. With a real species we would have used independent evaluation data and calculated various statistics to assess the accuracy of the classification (reviewed in Fielding and Bell, 1997). But here, with a virtual species, the 'true' habitat suitability that the models were supposed to reproduce was perfectly known. More adapted statistics based on the Pearson correlation coefficient between the two maps could thus be used. In order to get round the pseudo-replication engendered by spatial auto-correlation between cells, we proceeded as follows: 250 cells were picked randomly and a determination coefficient R^2 (proportion of variance explained by the model) was computed between the values of these cells in the 'result' map and the 'truth' map (Mesplé et al., 1996). This process was replicated ten times and the mean and standard deviation of R^2 were computed. The mean R^2 was used to assess the accuracy of the models. The results obtained by both techniques in each scenario were compared by a bilateral Student t-test. Their sensitivities to distribution scenario and sample size were also assessed with a Student t-test for each method.

3. Results

Equilibrium and overabundance scenarios were addressed with two sample sizes (300 and 1200 points) for both analyses (ENFA and GLM) and the spreading scenario only with 300 sample points, which makes a total of ten habitat suitability maps.

In order to compare the predictive power of these 'result' maps, the proportion of explained variance (R^2) was computed on a sample of 250 pairs of points taken in the 'result' map and in the 'truth' map. This coefficient was computed ten times for each map and the mean (R^2) and standard deviation (S.D.) were computed. These average R^2 achieved by each analysis were then compared in pairs by mean of a Student t-test.

Due to the stochasticity added in the process of building the 'truth' habitat suitability map, it was impossible to obtain $R^2 = 1$. The best model would have been the map computed just before the addition of stochasticity and this one gave $\langle R^2 \rangle = 0.67$ (S.D. = 0.07). This is to be remembered when looking at the absolute signification of the results presented in Table 2.

In the 'spreading' scenario, the ENFA proved to be significantly more efficient than the GLM. In the equilibrium scenario, there was no significant difference between the two methods. In the case of 'overabundance', the GLM gave significantly better results with the sample size of 300, but when the sample was larger (1200 points) the difference between the two analyses disappeared.

Another interesting result was the sensitivity analysis of each method to the quality and quantity of the input data. This was achieved by intra-method pair comparisons using the same procedure as already described.

Table 3 shows that ENFA is generally robust, the predictive power of the maps being significantly different in only 30% of the cases. The 'overabundance' scenario was the most sensitive.

In contrast, the GLM is quite sensitive to data quality (scenario effect) but not to data quantity. Actually, predictive powers are always highly significantly different, except between maps produced with the same scenario but different sample sizes.

Table 2
Mean $\langle R^2 \rangle$ and standard deviation (S.D.) of the proportion of explained variance obtained by comparing ten times each 'result' map with the 'truth' map[a]

Scénario	GLM		ENFA		GLM = ENFA?
	$\langle R^2 \rangle$	S.D.	$\langle R^2 \rangle$	S.D.	P
Spreading, 300 points	0.38	0.03	**0.57**	0.05	3.9×10^{-9}****
Equilibrium, 300 points	0.53	0.04	0.55	0.03	1.3×10^{-1} NS
Overabundance, 300 points	**0.63**	0.02	0.57	0.04	2.2×10^{-4}***
Equilibrium, 1200 points	0.54	0.03	0.56	0.05	5.0×10^{-1} NS
Overabundance, 1200 points	0.63	0.04	0.60	0.03	1.2×10^{-1} NS

[a] The greater the value of $\langle R^2 \rangle$, the higher the predictive power of the 'result' map. ENFA proved to be better in the 'spreading' scenario, whereas GLM was better in the 'overabundance' scenario with the small sample size. The probability (P) of the GLM and ENFA to have different predictive power was computed with a bilateral Student t-test. When the difference is significant, the best analysis is emphasised. Note that the best R^2 that could be achieved was 0.67. NS, Non-significant; *** $10^{-4} < P < 10^{-3}$, **** $P < 10^{-4}$.

Table 3
Sensitivity analyses of ENFA and GLM[a]

		Spreading (300 points)	Overabundance (300 points)	Equilibrium (300 points)	Overabundance (1200 points)
ENFA	Equilibrium (300 points)	NS (0.24)			
	Overabundance (300 points)	NS (0.67)	NS (0.38)		
	Equilibrium (1200 points)	NS (0.43)	NS (0.84)	NS (0.63)	
	Overabundance (1200 points)	NS (0.12)	**	*	*
GLM	Equilibrium (300 points)	****			
	Overabundance (300 points)	****	****		
	Equilibrium (1200 points)	****	NS (0.30)	****	
	Overabundance (1200 points)	****	****	NS (0.95)	****

[a] The upper part of the table compares predictive power of maps produced by the ENFA in each scenario. The lower part compares the maps produced by the GLM. ENFA maps have generally a similar predictive power (t-test not significant), while the GLM maps are generally different (t-test very highly significant). When the predictive powers of two 'result' maps do not differ significantly (NS), the t-test probability is given between parentheses. Significant results: * $0.01 < P < 0.05$, ** $0.001 < P < 0.01$, and **** $P < 0.0001$.

4. Discussion

The three addressed scenarios were modelled with unequal success by the two analyses. The ENFA appeared to be very robust to data quality and quantity, none of the investigated cases presents a significantly better or poorer fit; the overall goodness of fit was good with an average explained variance proportion of 0.58 (S.D. = 0.02). On the other hand, GLM was moderately sensitive to data quality but not to data quantity (average explained variance, 0.52; S.D. = 0.11).

Relying on absence data is both the strength and the weakness of this analysis: when they really reflect low habitat suitability (like in the 'overabundance' scenario) the additional information improves the model, but when the absences are due to historical causes (like in the 'spreading' scenario) this information is fallacious and decreases the overall predictive power.

ENFA and GLM were not sensitive to sample size, as both analyses produce only slightly better results with 1200 points than with 300 points. An interesting sequel to this study would be to explore more thoroughly the effect of the sample size and particularly the minimal efficient size as it could give useful clues when conceiving a sampling design.

As this experiment was not designed to explore qualitatively the results of these methods, it was not clear which one produced the best ecological interpretation of the data. GLM stepwise procedure is highly sensitive to predictor input order when these are not fully uncorrelated; adding or removing a predictor often qualitatively modifies the resulting model. In contrast, ENFA is not at all sensitive to this 'input effects'. Thus, when ecological interpretation is the aim of the study, ENFA could be more useful even for situations in which a GLM should provide a higher correlation to observed data.

Spatial autocorrelation is always problematic with the use of geographical space (Legendre, 1993). In this study, the problem arises when comparing the 'result' habitat suitability maps with the 'truth' habitat suitability map. Independent data are needed in order to use adequately the t-test of significance (Sokal and Rohlf, 1981). To reduce the correlation between sample points, a small sample size (250 points, being separated by an average distance of 913 m) was used to compute the R^2 statistics. Nevertheless, spatial autocorrelation can never be totally removed and it is best to cope with it (Legendre, 1993). Several methods exist either to remove spatial autocorrelation or to take it into account when computing the significance test (for example, Clifford et al., 1989; Dutilleul, 1993). Unfortunately, none of them were suitable to our case (testing the equality of two R^2 values) and generalising them to include it was beyond the scope of this paper and would have added little information.

Actually, ENFA–GLM comparison (Table 2) and GLM sensitivity analysis (lower part of Table 3) are very clear: the t-tests are either highly significant ($P < 10^{-4}$) or not at all significant ($P > 0.1$). Taking autocorrelation into account would lower the actual sample size to an effective sample size (Dutilleul, 1993) but it would probably not lower it enough to qualitatively modify these results. Actually, ENFA sensitivity analysis only (Table 3, upper part) could be qualitatively modified should the degrees of freedom be lower; two hardly significant ($P = 0.03$) pair comparisons could become non-significant. This would prove ENFA even more robust.

The virtual species approach proved to be most serviceable. When comparing models on a real data basis, it is only possible to make assumptions about what is the true habitat suitability by using various expert and statistical evaluation methods. Many factors are out of reach and may introduce a bias that cannot be accurately assessed (Alldredge and Ratti, 1992; Paruelo and Tomasel, 1997; Guisan et al., 1999; Manel et al., 1999): they may be historical (disturbances, catastrophes, diseases or colonisation events) or spatial (dispersal barriers or corridors), or they can be ecological (interspecific competition). Real data are therefore only a snapshot of a dynamical situation and can only give a partial and instantaneous comprehension of the fundamental ecological niche. By generating a virtual species, the 'truth' is now completely reachable and resulting models can be accurately compared to it.

Moreover, previous works (Lek et al., 1996; Paruelo and Tomasel, 1997; Guisan et al., 1999; Manel et al., 1999; Özesmi and Özesmi, 1999) compare methods on a unique case; the method application domain may thus hardly be explored. In contrast, our 'true' ecological niche approach may be used to generate various kinds of data sets to test different situations. In this paper, we explored the effect of colonisation history on the accuracy of two models, but sensitivity to many other effects could be tested: sampling size, sampling bias, interspecific competition, etc. Field ecologists have proposed a wide panel of statistical analyses but it is often difficult to choose among them. By circumscribing theoretically their application domain, this virtual species approach helps to select the most appropriate model in each situation. The ability to manipulate the virtual species allow one to isolate and thus to better understand problems encountered when dealing with real species; this will end up in better-suited analyses.

Although the error-free quality of simulated data has been often used in other domains to qualify method results (Ferré, 1995; Mesplé et al., 1996; Olley and Kochhar, 1996; Delay and Lamotte, 2000), it is less current in ecology (Paruelo and Tomasel, 1997; Kendall et al., 1999; Moilanen, 1999) and has never been applied to habitat-suitability model assessment.

There is a risk that virtual species do not simulate correctly the reality, introducing errors or biases in the results. In this study, we reduced this risk by several means. (1) We used real ecogeographical data; simulated predictors could be interesting in some cases, for example to explore the model sensitivity to their distribution, but as our study was focused on the quality of presence/absence data, there was no need to do so. The correlation between variables and their spatial auto-correlation, as well as their distribution, was therefore representative of what can be found in reality. (2) The niche function was made up of both linear and non-linear components, with some stochasticity. Therefore, the resulting niche shape did not favour one particular analysis. (3) Half of the predictors included as predictors in the analyses were not used to generate the virtual species. This simulated the fact that, in real cases, true predictors are generally unknown or not available and models are basically built on correlated variables

5. Conclusions

1. This paper gives insights on the domains of application of GLM and ENFA. It appears that the robustness of ENFA makes it particularly suitable and efficient when the quality of data is either poor (the absence data are unreliable) or unknown. The GLM offers slightly better results when the available presence/absence data are sufficiently good.
2. Virtual species simulation proved to be useful when assessing analysis predictive power in spatial ecology, allowing one to achieve a more accurate evaluation and to better control the experiment parameters.

Acknowledgements

This work was funded by the Swiss Federal Office for Environment, Forest and Landscape (OFEFP) (grant 0310.3600.305) and by the Laboratory of Conservation Biology, Institute of Ecology, University of Lausanne. The authors wish to thank Karine Fattebert and Antoine Guisan for their kind help with S-PLUS use and GLM methods. Pierre Dutilleul's and Jerôme Goudet's assistance during our struggle against spatial autocorrelation was also greatly appreciated. Many thanks to Sebastien Sachot, Antoine Guisan and two anonymous reviewers who helped to improve our manuscript. Finally, the authors are grateful to Nicolas Perrin, whose keen advice and sharp revisions guided us all through our work.

References

Akçakaya, H.R., Atwood, J.L., 1997. A habitat-based metapopulation model of the California Gnatcatcher. Conservat. Biol. 11, 422–434.

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

Alldredge, J.R., Ratti, J.T., 1992. Further comparison of some statistical techniques for analysis of resource selection. J. Wildl. Manag. 56, 1–9.

Augustin, N.H., Mugglestone, M.A., Buckland, S.T., 1996. An autologistic model for the spatial distribution of wildlife. J. Appl. Ecol. 33, 339–347.

Austin, M.P., Cunningham, R.B., Fleming, P.M., 1984. New approaches to direct gradient analysing using environmental scalars and statistical curve-fitting procedures. Vegetatio 55, 11–27.

Breitenmoser, U., Zimmermann, F., Olsson, P., Ryser, A., Angst, C., Jobin, A., Breitenmoser-Würsten, C., 1999. Beurteilung des Kantons St.Gallen als Habitat für den Luchs. KORA, Bern.

Clifford, P., Richardson, S., Hémon, D., 1989. Assessing the significance of the correlation between two spatial processes. Biometrics 45, 123–134.

Delay, F., Lamotte, J.-L., 2000. Numerical simulations of geological reservoirs: improving their conditioning through the use of entropy. Math. Comput. Simulat. 52, 311–331.

Dutilleul, P., 1993. Modifying the t-test for assessing the correlation between two spatial processes. Biometrics 49, 305–314.

Eastman, J.R., 1997. Idrisi for Windows 2.0 User's Guide. Clark University, Worcester.

Ferré, L., 1995. Selection of components in principal component analysis: a comparison of methods. Computat. Stat. Data Anal. 19, 669–682.

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

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

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

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

Hirzel, A.H., Hausser, J., Perrin, N., 2001. Biomapper 1.0. Laboratory for Conservation Biology, Lausanne http://www.unil.ch/biomapper.

Hirzel, A., Hausser, J., Perrin, N. Ecological-Niche Factor Analysis: How to compute habitat-suitability maps without absence data? Ecology, in press.

Hutchinson, G.E., 1957. Concluding remarks. Cold Spring Harbour symposium on quantitative biology.

Jongman, R.H.G., ter Braak, C.J.F., Van Tongeren, O.F.R., 1987. Data Analysis in Community and Landscape Ecology. Cambridge University Press, Cambridge.

Kendall, B.E., Briggs, C.J., Murdoch, W.W., Turchin, P., Ellner, S.P., McCauley, E., Nisbet, R.M., Wood, S.N., 1999. Why do populations cycle? A synthesis of statistical and mechanistic modeling approaches. Ecology 80, 1789–1805.

Le Lay, G., Clergeau, P., Hubert-Moy, L., 2001. Computerized map of risk to manage wildlife species in urban areas. Environ. Manage. 27, 451–461.

Legendre, P., 1993. Spatial autocorrelation: trouble or new paradigm? Ecology 74, 1659–1673.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., Aulagnier, S., 1996. Application of neural networks to modelling nonlinear relationships in ecology. Ecol. Model. 90, 39–52.

Manel, S., Dias, J.M., Buckton, S.T., Ormerod, S.J., 1999. Alternative methods for predicting species distribution: an illustration with Himalayan river birds. J. Appl. Ecol. 36, 734–747.

McArdle, B.H., 1990. When are rare species not there? OIKOS 57, 276–277.

Mesplé, F., Troussellier, M., Casellas, C., Legendre, P., 1996. Evaluation of simple statistical criteria to qualify a simulation. Ecol. Model. 88, 9–18.

Mladenoff, D.J., Haight, R.C., Sickley, T.A., Wydeven, A.P., 1997. Causes and implications of species restoration in altered ecosystems. A spatial landscape projection of wolf population recovery. Bioscience 47, 21–31.

Moilanen, A., 1999. Patch occupancy models of metapopulation dynamics: efficient parameter estimation using implicit statistical inference. Ecology 80, 1031–1043.

Nicholls, A.O., 1989. How to make biological surveys go further with generalized linear model. Biol. Conservat. 50, 51–75.

Olley, P., Kochhar, A.K., 1996. Case simulation to assess learning systems. Eng. Appl. Artif. Intell. 9, 285–300.

Özesmi, S.L., Özesmi, U., 1999. An artificial neural network approach to spatial habitat modelling with interspecific interaction. Ecol. Model. 116, 15–31.

Palma, L., Beja, P., Rodgrigues, M., 1999. The use of sighting data to analyse Iberian lynx habitat and distribution. J. Appl. Ecol. 36, 812–824.

Paruelo, J.M., Tomasel, F., 1997. Prediction of functional characteristics of ecosystems: a comparison of artificial neural networks and regression models. Ecol. Model. 98, 173–186.

Sanchez-Zapata, J.A., Calvo, J.F., 1999. Raptor distribution in relation to landscape composition in semi-arid Mediterranean habitats. J. Appl. Ecol. 36, 254–262.

Sokal, R.R., Rohlf, F.J., 1981. Biometry: The Principles and Practice of Statistics in Biological Research. W.H. Freeman, New York.

Solow, A.R., 1993. Inferring extinction from sighting data. Ecology 74, 962–964.

Welsh, A.H., Cunningham, R.B., Donnelly, C.F., Lindenmayer, D.B., 1996. Modelling the abundance of rare species: statistical models for counts with extra zeros. Ecol. Model. 88, 297–308.