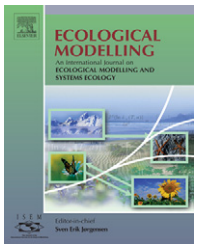
available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/ecolmodel



Evaluating the ability of habitat suitability models to predict species presences

Alexandre H. Hirzel[a,b,*], *Gwenaëlle Le Lay*[a], *Véronique Helfer*[a], *Christophe Randin*[a], *Antoine Guisan*[a]

[a] *Department of Ecology and Evolution, University of Lausanne, CH-1015 Lausanne, Switzerland*
[b] *Division of Conservation Biology, Institute of Zoology, University of Bern, CH-3012 Bern, Switzerland*

ARTICLE INFO

Article history:
Published on line 7 July 2006

Keywords:
Niche-based modelling
Model evaluation
Cross-validation
Generalised linear models (GLM)
Alpine plants
Swiss Alps

ABSTRACT

Models predicting species spatial distribution are increasingly applied to wildlife management issues, emphasising the need for reliable methods to evaluate the accuracy of their predictions. As many available datasets (e.g. museums, herbariums, atlas) do not provide reliable information about species absences, several presence-only based analyses have been developed. However, methods to evaluate the accuracy of their predictions are few and have never been validated. The aim of this paper is to compare existing and new presence-only evaluators to usual presence/absence measures.

We use a reliable, diverse, presence/absence dataset of 114 plant species to test how common presence/absence indices (Kappa, MaxKappa, AUC, adjusted D^2) compare to presence-only measures (AVI, CVI, Boyce index) for evaluating generalised linear models (GLM). Moreover we propose a new, threshold-independent evaluator, which we call "continuous Boyce index". All indices were implemented in the BIOMAPPER software.

We show that the presence-only evaluators are fairly correlated ($\rho > 0.7$) to the presence/absence ones. The Boyce indices are closer to AUC than to MaxKappa and are fairly insensitive to species prevalence. In addition, the Boyce indices provide predicted-to-expected ratio curves that offer further insights into the model quality: robustness, habitat suitability resolution and deviation from randomness. This information helps reclassifying predicted maps into meaningful habitat suitability classes. The continuous Boyce index is thus both a complement to usual evaluation of presence/absence models and a reliable measure of presence-only based predictions.

1. Introduction

Models predicting the spatial distribution of species (Boyce and McDonald, 1999; Guisan and Zimmermann, 2000; Manly et al., 2002; Pearce and Boyce, 2006) – sometimes called resource selection function or habitat suitability models – are currently gaining interest. As they often help both in understanding species niche requirements and predicting species potential distribution, their use has been especially promoted to tackle conservation issues, such as managing species distribution, assessing ecological impacts of various factors (e.g. pollution, climate change), risk of biological invasions or endangered species management (Scott et al., 2002; Guisan and Thuiller, 2005). These models statistically relate field observations to

* *Corresponding author at*: Laboratory of Conservation Biology, Department of Ecology and Evolution, University of Lausanne, CH-1015 Lausanne, Switzerland. Fax: +41 21 692 4105.
E-mail address: Alexandre.Hirzel@unil.ch (A.H. Hirzel).

0304-3800/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2006.05.017

a set of environmental variables, presumably reflecting some key factors of the niche, like climate, topography, geology or land-cover. They produce spatial predictions indicating the suitability of locations for a target species, community or biodiversity. Different types of modelling techniques are used to fit different types of biological information recorded at each sample site: (1) *presence-only*: occurrences of the target species are recorded; (2) *presence/absence*: each sample site is carefully monitored so as to assert with sufficient certainty whether the species is present or absent. With plants, for instance, it is commonly done by listing exhaustively all species present in each sample site. The reliability of absences depends on the species' characteristics (e.g. biology, behaviour, history) (Hirzel et al., 2001), their local abundance and ease of detection (Kéry, 2002), and the survey design (Mackenzie and Royle, 2005). More rarely, data record information about species' abundance or demography (e.g. growth rate, survival).

Although models based on presence-only and presence/absence data provide the same kind of predictions (e.g. habitat suitability scores), they generally cannot use the same technique. This is because presence-only methods cannot contrast their predictions with the characteristics of places where the species is absent. This partly explains why presence/absence methods have known a greater development. These differences, and the lack of absences, make comparison of the two model types difficult (Zaniewski et al., 2002).

Assessing the predictive power of a model is of paramount importance, both for theoretical and applied issues. However, while presence/absence models have received a lot of attention and many evaluators are available for them (Fielding and Bell, 1997), evaluation of presence-only models is lagging behind. There is therefore a crucial need for reliable presence-based evaluation measures, as well as an assessment of how they compare to the presence/absence measures.

The main problem of presence-only evaluation measures is the lack of absences to counterbalance the presences. It is thus difficult to discriminate a model predicting presence everywhere from a more contrasted model. Attempts to solve this problem have followed two main approaches: (1) a first approach is to generate pseudo-absences and then apply the standard presence/absence techniques (e.g. Zaniewski et al., 2002; Anderson et al., 2003). (2) A second approach is to assess how much the model predictions differ from random expectation (e.g. Boyce et al., 2002; Hirzel et al., 2002; Reutter et al., 2003). In this category, the index recently proposed by Boyce et al. (2002) offers new insights. We tested it thoroughly and derived a new evaluator from it, which does not depend on the choice of boundaries between habitat suitability classes. A third original approach, proposed by Ottaviani et al. (2004), is based on compositional analysis. However, it is restricted to cases where evaluation data are in the form of polygons or large mapping units (e.g. large grid cells in an atlas), and thus does not apply here.

In this paper, we present various presence-only evaluation measures. To validate them, we build 114 presence/absence models chosen for the reliability of their absences and evaluate them with presence-only and presence/absence evaluators. We test correspondence between them and discuss how the new "Boyce indices" can improve the interpretation and utilisation of habitat suitability models.

2. Materials and methods

We define a habitat suitability (HS) map as composed of cells (or pixels) whose quantitative values range from 0 to 1. These values indicate how close the local environment is to the species' optimal conditions, higher values standing for the most suitable areas. This map may result from any statistical analysis (Guisan and Zimmermann, 2000; Pearce and Boyce, 2006). The models' evaluation consists in quantifying how accurately the map is predicting the presence and absence of the species (Buckland and Elston, 1993; Manel et al., 2001), as given by a set of *evaluation points*. This set may consist either of verified presences and absences, or of verified presences only. Optimally, this data set should be completely independent from the data used to calibrate the model, e.g. collected on other areas (Randin et al., in press). However, due to time and money constraints, most studies have only one dataset and have to split it between a calibration and an evaluation sets. This is the method we use in this paper.

2.1. Evaluation indices

Most measures currently used in the literature are based on presence/absence information. Their first step generally consists of choosing a habitat suitability threshold (often 0.5) supposed to separate unsuitable areas (HS below threshold) where the species should be absent, from suitable areas (HS above threshold) where it should be present. From this Boolean map, one builds the *confusion matrix*, which counts how many presence and absence evaluation points occur in the suitable and unsuitable areas (Fig. 1). Many evaluators are based on this matrix (see Fielding and Bell, 1997), the commonest being the Kappa index K (Cohen, 1960):

$$K = \frac{N\sum x_{ii} - \sum x_{i\cdot}x_{\cdot i}}{N^2 - \sum x_{i\cdot}x_{\cdot i}} \quad (1)$$

where x and N are counts of evaluation points as defined in Fig. 1. K varies from −1 to 1, high values indicating a good agreement between prediction and data, and 0 corresponds to random agreement.

These methods depend strongly on the suitability cut-off threshold, which is often chosen arbitrarily. Alternatively, one may use threshold-independent methods, like K_{max} (or MaxKappa, Guisan et al., 1998) and area under the curve (AUC, Zweig and Campbell, 1993; Fielding and Bell, 1997). The K_{max} index is the highest Kappa obtained when varying the thresh-

		Observed		Margin sums
		Presence	Absence	
Predicted	Presence	x_{11}	x_{12}	$x_{1\bullet}$
	Absence	x_{21}	x_{22}	$x_{2\bullet}$
Margin sums		$x_{\bullet 1}$	$x_{\bullet 2}$	N

Fig. 1 – Contingency table of the model predictions against the actual observations. The x_{ij} represent counts of evaluation points, with $N = \Sigma x_{ij}$.

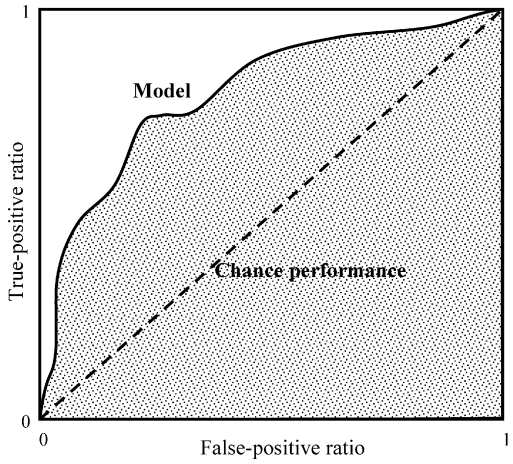


Fig. 2 – Curve of the true presence fraction ($=x_{11}/x_{.1}$) against the false presence fraction ($=x_{12}/x_{.2}$) computed for all possible cut-off points between 0 and 1. The AUC is the "area under the curve" and practically varies between 0.5 (not different from random expectation) and 1 (best model).

old from 0 to 1. The AUC is obtained by plotting, for each threshold in this range, the proportion of true positive $x_{11}/x_{.1}$ against the proportion of false positive $x_{12}/x_{.2}$ and by computing the area under the curve thus defined (Fig. 2). The AUC varies between 0 (worse-than-random model), 0.5 (random model) and 1 (best discriminating model).

When absence data are unreliable or unavailable, the model evaluation should be assessed for presences only. For this purpose, one possibility is to compare the model results to what would be expected from chance alone. Two simple evaluators are the absolute validation index (AVI) and contrast validation index (CVI) (Hirzel and Arlettaz, 2003; Hirzel et al., 2004). The AVI is the proportion of presence evaluation points falling above some fixed HS threshold (e.g. 0.5); it varies from 0 to 1. The CVI is the AVI minus the AVI of a model predicting presence everywhere (chance model), and varies from 0 to 0.5. As for the Kappa index, this approach suffers from having to choose an arbitrary threshold.

Boyce et al. (2002) proposed a way to relieve somewhat the threshold constraint. Their method consists in partitioning the habitat suitability range into b classes (or bins), instead of only two. For each class i, it calculates two frequencies: (1) P_i, the *predicted frequency* of evaluation points:

$$P_i = \frac{p_i}{\sum_{j=1}^{b} p_j} \quad (2)$$

where p_i is the number of evaluation points predicted by the model to fall in the habitat suitability class i and $\sum p_j$ is the total number of evaluation points; (2) E_i, the *expected frequency* of evaluation points, i.e. the frequency expected from a random distribution across the study area. This is given by the relative area covered by each class:

$$E_i = \frac{a_i}{\sum_{j=1}^{b} a_j} \quad (3)$$

where a_i is the number of grid cells belonging to habitat suitability class i, or area covered by the class i, and Σa_j is the overall number of cells in the whole study area.

Finally, for each class i, the *predicted-to-expected* (P/E) ratio F_i is given by

$$F_i = \frac{P_i}{E_i} \quad (4)$$

If the habitat model properly delineates the species suitable areas, a low suitability class should contain fewer evaluation presences than expected by chance, resulting in $F_i < 1$. Conversely, high suitability classes should have F_i increasingly higher than 1. The plot of P/E against the mean habitat suitability of each class thus provides a handy interpretation tool. In such a context, a good model is expected to show a monotonically increasing curve, i.e. F_i increase as suitability increases. Boyce et al. (2002) measure this monotonic increase by the Spearman rank correlation coefficient between F_i and i. This "Boyce Index" B_b varies from −1 to 1. Positive values indicate a model whose predictions are consistent with the presences distribution in the evaluation dataset, values close to zero mean that the model is not different from a chance model, negative values indicate an incorrect model, which predicts poor quality areas where presences are more frequent.

The main shortcoming of the Boyce index is its sensitivity to the number of suitability classes b and to their boundaries (Boyce et al., 2002; personal observations). To fix this problem, we derived a new evaluator based on a "moving window" of width W (say $W = 0.1$) instead of fixed classes. Computation starts with a first class covering the suitability range [0,W] whose P/E ratio is plotted against the average suitability value of the class, $W/2$. Then, the moving window is shifted from a small amount upwards and the P/E is plotted again. This operation is repeated until the moving window reaches the last possible range [1 − W, 1]. This provides a smooth P/E curve, on which a "continuous Boyce index" $B_{cont(W)}$ is computed.

One of the main differences between the Boyce indices and the classical evaluators is that they require the HS prediction to be computed on the whole study area. For maps with low number of pixels, the whole information can be imported into a statistical application; however, in most cases their computation must be done within a GIS, or a program having direct access to the GIS data files. We have thus implemented all these evaluators into the free software BIOMAPPER (Hirzel et al., 2006), which works directly on GIS files and controls the free statistical application R (R Development Core Team, 2005) to compute the GLMs.

2.2. *Cross-validation*

All the evaluation methods presented above provide a single measure of the model predictive power. k-fold cross-validation is a resampling approach that allows assessment of the robustness of this measure (Van Houwelingen and Le Cessie, 1990; Fielding and Bell, 1997; Hastie et al., 2001). Moreover, it also enables one to evaluate a model even when the species dataset is small, as it ensures an optimal use of the data to calibrate and evaluate the model. Cross-validation consists of randomly dividing the dataset into k independent

partitions, using $k-1$ of them to calibrate the model, and computing the evaluator on the left-out partition. This procedure is repeated k times, each time leaving out another partition. This produces k estimations of the evaluator, allowing assessment of its central tendency and variance (in this study, we used median and 90%-confidence interval). The number of partitions typically varies between 3 and 10, depending on the number of species points. This method assumes that the k partitions are independent. See Hastie et al. (2001) for more details on cross-validation processes.

2.3. Test dataset

In order to test and compare the above evaluators, reliable data were required. We used data from 539 non-forest mountain vegetation plots located in a ~700 km^2 area in the external calcareous Alps of Canton de Vaud (6°60′–7°10′E and 46°10′–46°30′N, altitude ranging from 375 to 3210 m) in Switzerland. The sampling points were randomly stratified by classes of elevation, slope and aspect. The plants of each sampling point were exhaustively inventoried (Randin et al., in press). Among these plants, we selected those species that fulfilled three criteria: (i) species are easily detectable in the field during the sampling period, so that absences cannot be due to the species being undetected; (ii) species cannot be confounded with another sister species; (iii) species are at least present in 10 vegetation plots. The points (i) and (ii) guarantied reliable presences and absences. We ended up with 114 species that were present in at least 10 cells and at most 249. These data were collected during three field-sampling periods, in the summers 2002–2004.

2.4. Environmental variables

The environmental variables we used to fit the models are known to have a major direct ecophysiological impact on plant species (Pearson et al., 2002; Dirnböck et al., 2003; Körner, 2003). They were all calculated with a 25 m × 25 m spatial resolution, as derived from the digital elevation models (DEM) available in the study area (MNT25, Swisstopo). We calculated slope from the DEM to account for gravitational processes acting upon vegetation. The TOPO index indicates local convexity of the topography, which is partly correlated with water accumulation, snow persistence, nitrogen enrichment or wind protection. Basic climatic variables were obtained by spatial interpolation of climatic weather stations (monthly data for the period 1961–1990), and then transformed into three physiologically meaningful bioclimatic variables: degree-days (with 0 °C as threshold of plant growth), moisture index over the growing season (June–August) and potential global solar radiation of the growing season (see Table 1 and references therein). Details on these variables can be obtained in Randin et al. (in press).

2.5. Habitat suitability modelling

We used generalised linear models (GLM, McCullagh and Nelder, 1989) with a binomial probability distribution and a logit link to compute the habitat suitability maps based on presence/absence data. In a first step, we computed generalised linear models (GLMs) (as implemented in R, R Development Core Team, 2005) for the whole presence/absence dataset. To prevent the models' sensitivity to species prevalence, we weighted the absence points so as to have a presence/absence ratio of 1:1. The independent variables were those listed in Table 1. For each species, the relevant variables of the model were selected by a stepwise procedure based on the AIC criterion (Akaike, 1973; S-Plus, 1999); whenever a variable was retained in its squared form, we forced its linear form to be also included (T. Hastie, personal communication). The retained model was then fixed and, in a second step, we applied a k-fold cross-validation process (with $k=5$) using only the retained variables. To ensure a similar presence/absence ratio between all partitions, the random selection of the k partitions was done independently for the presence and absences data. Presences and absences were weighted as in the full model. This produced five GLM models and five HS maps for each species.

2.6. Evaluation index comparisons

Each HS map resulting from the cross-validation was evaluated by D^2_{adj}, K, K_{max}, AUC, AVI and CVI. In comparison, to investigate the sensitivity and the significance of the Boyce index, we tested it with several number of classes b: B_2, B_4, B_5 and B_{10}, as well as, in its continuous form, with the corresponding class sizes: $B_{cont(0.5)}$, $B_{cont(0.25)}$, $B_{cont(0.2)}$ and $B_{cont(0.1)}$. B_{10} corresponds to the number of classes used by Boyce et al. (2002). We tested B_2 as it was expected to be more comparable to two classes evaluators like K, AVI and CVI. Moreover, the GLM fit to the calibration data was evaluated by the adjusted explained deviance D^2_{adj}, which corresponds to the amount of deviance explained by the model corrected by the effective number of degrees of freedom used to build the model (Guisan

Table 1 – Environmental variables used to model habitat suitability of the 114 plants

Variables	Details	References
Moisture (mm day^{-1})	Monthly average of daily atmospheric water balance from July to September	Zimmermann and Kienast (1999)
Degree-days (°C day)	Number of days with mean temperature above 0 °C time their mean temperature	Prentice et al. (1992)
Global solar radiation (kJ m^{-2} day^{-1})	Monthly average of daily global solar radiation from July to September	Kumar et al. (1997)
Slope (degrees)	Slope inclination	ArcInfo (2004)
Topo	Topographic convexity	Zimmermann (unpublished)

and Zimmermann, 2000). We computed the median and 90%-confidence interval of each evaluator for the five HS maps. We finally plotted the medians of each evaluator against each other for all species and computed their Pearson's correlation coefficient ρ. We validated the Boyce indices by comparison to AUC and K_{max}, as they are common threshold-independent presence/absence evaluators.

3. Results

The chosen species cover a wide spectrum of ecological niche types and sample size. The quality of their habitat suitability models range from very bad to excellent. All the investigated evaluation measures convey similar information, with Pearson correlation coefficients greater than 0.5 in most cases (Table 2a). In particular, for the models where more than 50 presence points were available, most evaluators show more than 70% of correlation (Table 2b).

Except for those based on very wide classes (B_2 and $B_{cont(0.5)}$), Boyce indices are highly correlated together, and are moreover highly consistent with presence/absence evaluators. They tend to be more correlated to AUC than to K_{max}; this tendency is stronger for the whole dataset (Table 2a) than for the 48 most prevalent species (Table 2b). In spite of having the same class pattern, B_2 and $B_{cont(0.5)}$ are but weakly correlated to K, AVI and CVI; they both give the poorest correlations and will not be considered further. The Boyce indices most consistent with the AUC and K_{max} were those with the largest number of classes (or smallest window sizes), B_{10}, $B_{cont(0.1)}$, B_5 and $B_{cont(0.2)}$. We tested more than 10 classes but the variance of the P/E curves becomes too high (results not shown). To simplify the discussion, we will from now on only consider one evaluator of each type: D^2_{adj}, AUC, K_{max}, B_{10}, $B_{cont(0.1)}$ and CVI (correlation graphs shown in Fig. 3). Although all Boyce indices are highly correlated with K, we chose not to consider this evaluator as it is not threshold-independent.

Fig. 4 shows the sensitivity of these evaluators to species prevalence. K_{max} is sensitive to species prevalence, tending to give higher values to common-species models, while D^2_{adj} and CVI tend to give higher scores to low prevalence models. The other evaluators are insensitive to prevalence (Fig. 4).

Table 2a – Pearson's correlation coefficients between pairs of evaluation indices (median of the k-fold values) on all species (n = 114)

	Presence/absence evaluation indices				Presence-only evaluation indices									
	N_1 [a]	D^2_{adj}	AUC	K	K_{max}	B_2	B_4	B_5	B_{10}	$B_{cont(0.1)}$	$B_{cont(0.2)}$	$B_{cont(0.25)}$	$B_{cont(0.5)}$	AVI	CVI
N_1 [a]	1.00														
D^2_{adj}	−0.40	1.00													
AUC	−0.24	**0.86**	1.00												
K	0.19	0.54	0.63	1.00											
K_{max}	0.47	0.38	0.55	0.79	1.00										
B_2	0.17	0.34	0.47	0.65	0.56	1.00									
B_4	0.11	0.63	0.63	0.74	0.64	0.53	1.00								
B_5	0.02	0.73	0.73	0.74	0.62	0.49	**0.81**	1.00							
B_{10}	−0.19	**0.84**	0.79	0.76	0.56	0.42	0.80	**0.89**	1.00						
$B_{cont(0.1)}$	−0.16	**0.83**	**0.80**	0.80	0.60	0.46	**0.83**	**0.90**	**0.96**	1.00					
$B_{cont(0.2)}$	0.02	0.68	0.69	**0.83**	0.67	0.60	**0.87**	**0.86**	**0.85**	**0.91**	1.00				
$B_{cont(0.25)}$	0.05	0.62	0.64	**0.82**	0.66	0.63	**0.82**	0.79	0.79	**0.85**	**0.98**	1.00			
$B_{cont(0.5)}$	−0.10	0.40	0.50	0.62	0.46	0.65	0.54	0.47	0.52	0.57	0.67	0.71	1.00		
AVI	−0.17	0.80	0.77	0.77	0.55	0.57	0.73	0.76	**0.83**	**0.86**	**0.80**	0.77	0.62	1.00	
CVI	−0.32	0.80	0.74	0.67	0.43	0.52	0.66	0.69	0.79	**0.80**	0.74	0.71	0.64	**0.96**	1.00

Correlations ≥ 0.8 are emphasised.

[a] Number of presence points.

4. Discussion

On the range covered by the 114 studied plant species, and according to the environmental characteristics of our study area, all evaluators convey correlated information. This is an important result meaning that the presence-only evaluators can be trusted.

4.1. *Evaluator comparisons*

As expected, the quality of the 114 GLM models fitted varies greatly. Overall, the tested evaluators agree about their ranking, in particular among prevalence-insensitive indices (AUC, B_{10}, $B_{cont(0.1)}$). The results show that Boyce indices tend to give poor results when computed on a small number of classes. In particular, B_2 and $B_{cont(0.5)}$ have a low correlation with almost all evaluators, including those based on a fixed threshold as

Table 2b – As in Table 2a but only on species with more than 50 presence points (n = 48)

	Presence/absence evaluation indices					Presence-only evaluation indices								
	N_1[a]	D^2_{adj}	AUC	K	K_{max}	B_2	B_4	B_5	B_{10}	$B_{cont(0.1)}$	$B_{cont(0.2)}$	$B_{cont(0.25)}$	$B_{cont(0.5)}$	AVI	CVI
N_1[a]	1.00														
D^2_{adj}	−0.10	1.00													
AUC	−0.07	**0.93**	1.00												
K	0.00	**0.90**	**0.81**	1.00											
K_{max}	0.32	**0.85**	**0.87**	**0.82**	1.00										
B_2	−0.06	0.45	0.57	0.45	0.46	1.00									
B_4	0.19	0.62	0.56	0.71	0.70	0.35	1.00								
B_5	0.07	0.74	0.70	**0.80**	0.76	0.43	0.77	1.00							
B_{10}	−0.05	**0.89**	**0.81**	**0.91**	0.79	0.42	0.78	**0.88**	1.00						
$B_{cont(0.1)}$	−0.07	**0.90**	**0.84**	**0.93**	**0.82**	0.47	0.79	**0.88**	**0.98**	1.00					
$B_{cont(0.2)}$	−0.04	0.78	0.74	**0.86**	0.76	0.49	**0.82**	**0.82**	**0.89**	**0.93**	1.00				
$B_{cont(0.25)}$	−0.06	0.75	0.72	**0.85**	0.73	0.50	0.75	0.74	**0.85**	**0.89**	**0.98**	1.00			
$B_{cont(0.5)}$	−0.38	0.52	0.52	0.52	0.41	0.32	0.41	0.39	0.53	0.58	0.63	0.69	1.00		
AVI	−0.02	**0.91**	**0.83**	**0.97**	**0.82**	0.41	0.64	0.75	**0.89**	**0.91**	**0.82**	**0.82**	0.55	1.00	
CVI	−0.29	**0.85**	0.76	**0.86**	0.65	0.37	0.54	0.65	**0.82**	**0.84**	0.78	0.80	0.69	**0.91**	1.00

Correlations ≥ 0.8 are emphasised.

[a] Number of presence points.

Table 3 – Median value taken by the evaluators (minimum and maximum in brackets)

Evaluator	Value
D^2_{adj}	0.34 (0.07, 0.75)
AUC	0.88 (0.63, 0.99)
K	0.27 (−0.04, 0.70)
K_{max}	0.49 (0.10, 0.82)
B_{10}	−0.07 (−0.67, 0.65)
$B_{cont(0.1)}$	−0.07 (−0.66, 0.61)
CVI	0.24 (−0.12, 0.98)

AVI, CVI and K. As the Boyce indices are based on the Spearman rank correlation, they are more sensitive to larger number of classes. However, classes cannot be added indefinitely as the variance among cross-validation partitions increases as their width decreases. Ten classes (class width = 0.1) seems to be the optimum, advocating for B_{10} and $B_{cont(0.1)}$. We prefer the later as it does not depend on any particular class cutting thresholds.

The agreement between presence/absence and presence-only measures tends to be lower when the species prevalence is below 10% (<50 presences) (Tables 2). This is because a low number of presences prevent presence-only evaluators from assessing the overall quality of the model, whilst presence/absence evaluators can still rely on the fit between predicted and observed absences. Therefore, when presences are scarce, presence/absence evaluators often give an intermediate score to the model on the base of absence predictions, whilst presence-only evaluators assess the model as poor (cf. evaluator ranges in Table 3 and Fig. 5). This redemption of the model by the absences may be acceptable if the cost of overlooking suitable habitat is not too high. It is important to take such evaluation scale shifts into account for wildlife management applications.

Why some models better predict absences than presences may come from various causes. First, unreliable species data may bring too much noise for a proper niche modelling. In our case, as the species were carefully selected for the reliability of their presence/absence dataset, we can mostly rule out this effect. Second, the model accuracy depends on the environmental variables relevance for the species. Although we chose good general predictors for plants (Dirnböck et al., 2003), some more specialised variables are obviously missing for the badly modelled species. This was expected, and actually sought, as we wanted the model quality to cover as wide a palette as possible. Thus, when the environmental variables are irrelevant to the species niche, the model cannot efficiently predict presences.

Most measures evaluate how well a model can predict absence and presence. By contrast, the Boyce indices assess the model ability to consistently predict several levels of suitability. A complementary evaluation would be to apply the same Boyce approach to the absences. In that case, one would expect negative P/E curves, i.e. negative Spearman ranks. The combination of these two facets of the Boyce indices seems promising, as it would bring the power of continuous evaluation to presence/absence models.

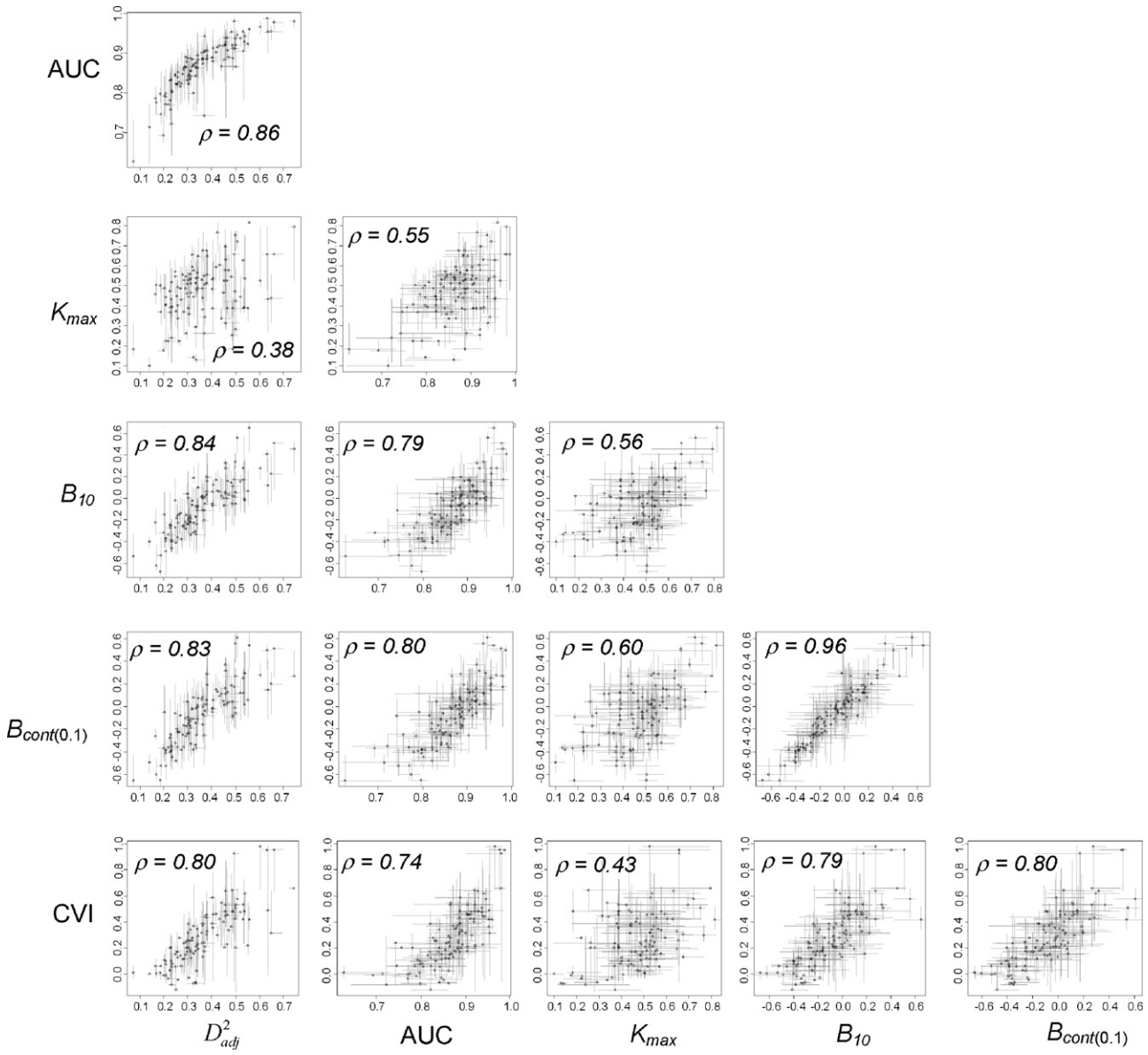


Fig. 3 – Relationship between the main evaluators. Each point represents the median value of the evaluators computed on the cross-validation partitions. The grey lines represent the interquartile range. The Pearson's correlation coefficient is indicated for each pair of evaluators.

4.2. *Interpreting the P/E curves*

While an evaluation index gives a summary of the model prediction ability, the continuous P/E curves provide a wealth of valuable insights into the model accuracy, offering three levels of information.

First, the variance among the cross-validation curves gives information about model *robustness* all along the HS range. The narrowness of the confidence interval reflects the model sensitivity to particular calibration points. As the variance often fluctuate along the curve, one can thus determine which parts of the model are the most accurate. For instance, a model could provide trustable prediction for low suitability regions (good absence prediction) but be more variable about high suitability (bad presence prediction). Such information provides a finer understanding of the weaknesses of the model. The manager can then take these weaknesses into account when applying the predictions to management decisions. Alternatively, these weaknesses may give clues about what parts of the predictions (e.g. absences) must be improved. In this study, we indicate robustness by using cross-validation and providing a confidence interval around the evaluators.

The second information level is related to the actual shape of the P/E curve. An ideal model would have a linear P/E curve and could thus predict habitat suitability with an infinitely fine resolution. It means that the suitability index is really proportional to the probability of use, as defined by Manly et al. (2002). Real curves however may exhibit non-linear (e.g. exponential) or staircase shapes. Wherever the local slope is flat or negative, the corresponding range of HS may be pooled into one

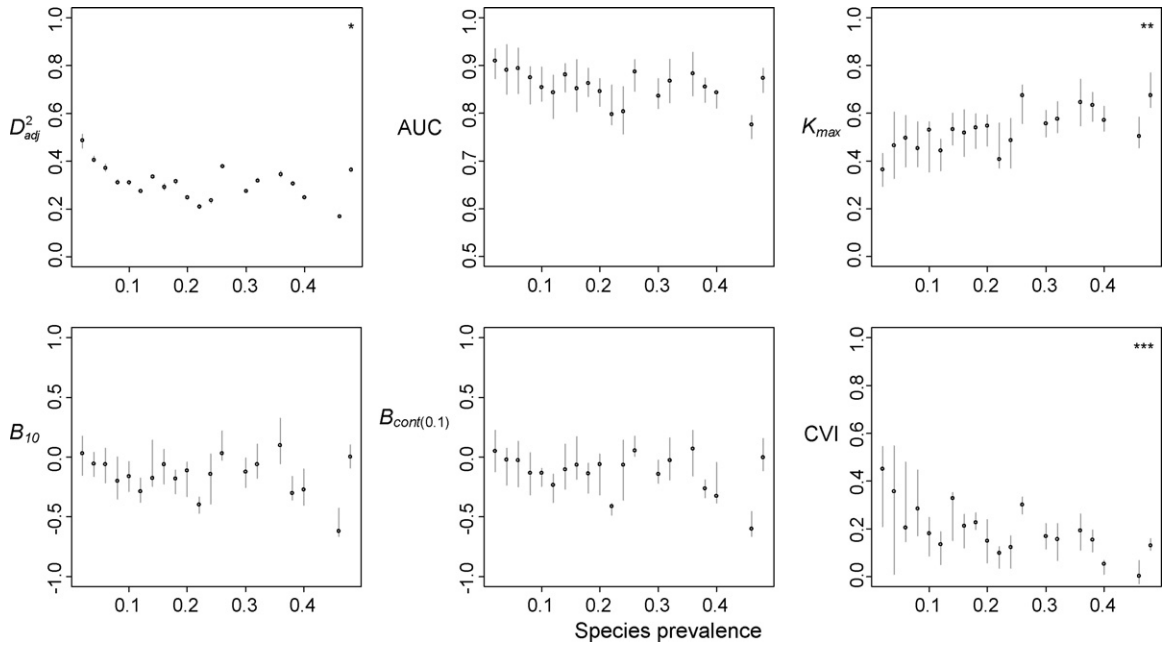


Fig. 4 – Evaluator sensitivity to species prevalence (proportion of presence points in the dataset). The points represent the index median value and the grey lines the interquartile range. *t*-test significance of the Pearson's correlation coefficient indicated by stars: $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

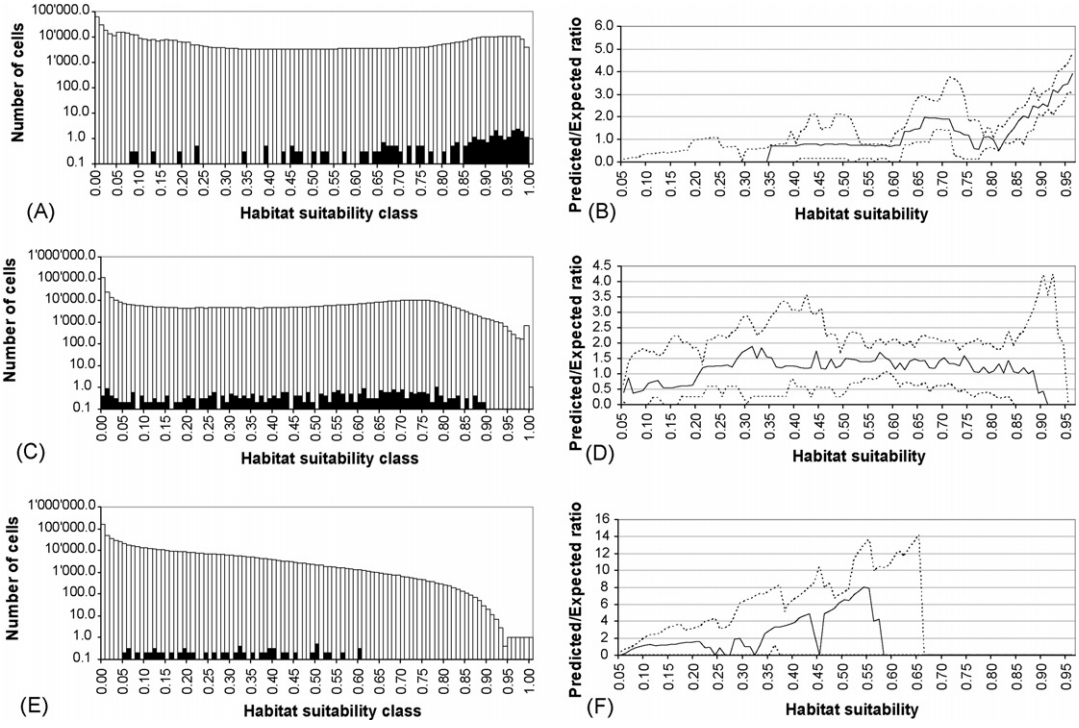


Fig. 5 – Typical examples of the three models: (A and B) good model; (C and D) random model (confidence interval brackets the 1-line all along); (E and F) bad model (presences fall in low suitability areas). In the left-hand column, the white bars show the number of cells belonging to each habitat suitability class, while black bars are the number of cells with asserted presence in these classes. These histograms average the results of the five cross-validation HS maps, where vertical scales are logarithmic. In the right-hand column, the predicted/expected curves computed by a moving window of width 0.1 (plain line = median, dashed lines = 90%-confidence interval). More details on the models include: (A and B) *Lathyrus pratensis*, $K_{max} = 0.68$, AUC = 0.92, $B_{10} = 0.82$, $B_{cont(0.1)} = 0.76$; (C and D) *Alchemilla xanthochlora*, $K_{max} = 0.68$, AUC = 0.87, $B_{10} = 0.00$, $B_{cont(0.1)} = 0.00$; (E and F) *Gypsophila repens*, $K_{max} = 0.37$, AUC = 0.78, $B_{10} = -0.26$, $B_{cont(0.1)} = -0.40$.

class without loss of information (Fig. 5). This means that any departure from the straight line actually decreases the *resolution* of the model predictions, i.e. its ability to distinguish many different classes of suitability. Model resolution is partly contained in the Boyce indices as they get penalised whenever the *P/E* curve goes down. However, being based on the Spearman correlation coefficient, they cannot discriminate between monotonic curves (e.g. linear, exponential and sigmoid curves would all get the same score).

The third information level refers to the maximum value reached by the *P/E* curve. This value reflects how much the model differs from chance expectation, or *deviation from randomness*. This score reflects the model ability to differentiate the species niche characteristics from those of the studied area. This measure must be taken with care as our experience has shown that it highly depends on the species niche breadth, the extent of the study area, the scale of the study (i.e. the environmental variables resolution), and the relevance of the chosen environmental variables. For instance, with a mountain species, a model built at the whole country level with climatic variables is bound to get a higher information index than one focusing on the mountain ranges; however, the country-wide model is obviously not better than the mountain-wide one. Practically, one must use this deviation from randomness only to compare models applied to the same species and the same study area.

4.3. *Reclassifying HS maps*

Most habitat suitability models (including GLMs) generate maps showing continuous gradients of suitability. This kind of output obviously conveys more information than a sheer presence/absence map and is more convenient for wildlife management support. However, the present study has shown that a continuous scale is often misleading. Even good predictive models suffer from uncertainty, making the use of a full continuous HS scale spurious. A reclassified map showing only a few classes may be more honest about its actual informative content. The problem of choosing objectively the HS class boundaries immediately arises. For binary reclassifications – presence/absence – methods already exists: the K_{max} approach readily provides the HS threshold that maximise the *K* index. The AUC approach allows weighting the risks of over- and under-prediction and computing the optimal threshold accordingly (Fielding and Bell, 1997). For more than two classes, the *P/E* curves provide a handy support for choosing (1) the number of classes and (2) their boundaries. The optimal number of classes may be defined by looking at the confidence interval around the continuous *P/E* curve (e.g. Fig. 5B), the goal being of finding how many HS classes (on the horizontal axis) may be defined while minimizing their overlap in *P/E* ratio (vertical axis). The continuous *P/E* curves also allow choosing the class boundaries objectively (Fig. 6). A first, natural boundary is defined by the *P/E* = 1 line: where *P/E* confidence interval is lower than 1, the model is predicting less presences than expected by chance, and the opposite when it is greater than 1; this may be used to distinguish unsuitable, marginal (random and uncertain) and suitable habitat (Fig. 6). Another natural threshold is the HS below which no presence ever occurs, suggesting uninhabitable conditions, which is equivalent to the threshold defining the minimal predicted area (MPA_{100}) of Engler et al. (2004). Additional thresholds may be placed at the steps of the curve. The *P/E* ratios also provide a reproducible HS partitioning scheme: one could define HS classes predicting twice more presences than expected, thrice more, and so on. Fig. 6 illustrates this process of HS partitioning. Note that with categorical variables, it would make little sense to have more HS classes than categories.

In conclusion, our work has shown that evaluating a habitat suitability model based only on presences is possible and a valuable exercise. Among the presence-only evaluators, the continuous Boyce index $B_{cont(0.1)}$ was most accurate for characterizing predictive capability among our sample of 114 plant distributions. Accordingly, we suggest that it is both a complement to usual evaluation of presence/absence models (e.g. GLMs and GAMs, Guisan et al., 2002) and a reliable measure of presence-only based predictions (e.g. ecological niche factor analysis: Hirzel et al., 2002; or resource selection functions: Manly et al., 2002). Such measures could also prove useful to evaluate presence/absence-based models when an accurate

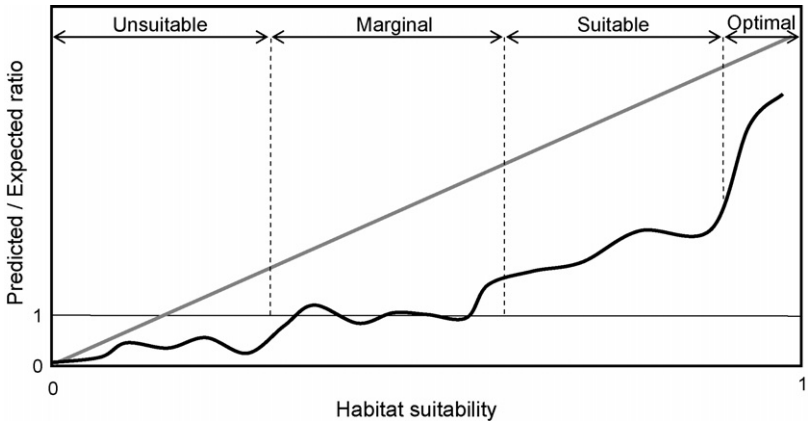


Fig. 6 – Predicted/expected curve shapes. An ideal model would give a straight *P/E* curve (plain grey line). Actual models often exhibit an irregular increase (black plain curve). The curve shape and its confidence interval (dashed curves) may be used to define the boundaries of habitat suitability classes (as suggested by the vertical dashed lines). The horizontal thin line at *P/E* = 1 would be the curve of a completely random model.

prediction of presences is crucial, as in the case of detecting new populations of threatened species (e.g. Engler et al., 2004). Moreover, we stress the need to reclassify habitat suitability maps so as to provide more honest and relevant predictions. The P/E curves described offer a handy support for this reclassification.

Acknowledgements

We wish to thank Mark S. Boyce, Gretchen G. Moisen and all the participants of the Riederalp Workshop, Switzerland, 2004, for stimulating discussions about the model evaluation, Patrick Patthey, Julie Jacquiéry and Pietro Persico for the first explorations of the Boyce index. We also wish to thank Jane Elith and two anonymous reviewers who helped improve this article. We are grateful to Fabien Fivaz for the help with R scripts as well as to all those who contributed to the field work: Pascal Vittoz, Stéfanie Maire, Dario Martinioni, Simone Peverelli, Chantal Peverelli, Séverine Wydler, Lorenzo De Stefani and Roxane Milleret.

REFERENCES

Akaike, H., 1973. Maximum likelihood identification of Gaussian autoregressive moving average models. Biometrika 60, 255–265.

Anderson, R.P., Lew, D., Peterson, A.T., 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecol. Model. 162, 211–232.

ArcInfo, 2004. ArcInfo Version 9.0. Environmental Systems Research Institute Inc., Redlands, CA.

Boyce, M.S., McDonald, L.L., 1999. Relating populations to habitats using resource selection functions. Trends Ecol. Evol. 14, 268–272.

Boyce, M.S., Vernier, P.R., Nielsen, S.E., Schmiegelow, F.K.A., 2002. Evaluating resource selection functions. Ecol. Model. 157, 281–300.

Buckland, S.T., Elston, D.A., 1993. Empirical models for the spatial distribution of wildlife. J. Appl. Ecol. 30, 478–495.

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

Dirnböck, T., Dullinger, S., Grabherr, G., 2003. A regional impact assessment of climate and land-use change on alpine vegetation. J. Biogeogr. 30, 401–417.

Engler, R., Guisan, A., Rechsteiner, L., 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. J. Appl. Ecol. 41, 263–274.

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

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

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

Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecol. Lett. 8, 993–1009.

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

Hastie, T., Tibshirani, R., Friedman, J.H., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, New York.

Hirzel, A.H., Arlettaz, R., 2003. Modelling habitat suitability for complex species distributions by the environmental-distance geometric mean. Environ. Manage. 32, 614–623.

Hirzel, A.H., Hausser, J., Chessel, D., Perrin, N., 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecology 83, 2027–2036.

Hirzel, A.H., Hausser, J., Perrin, N., 2006. Biomapper 3.2. Lab. for Conservation Biology, University of Lausanne, Lausanne, http://www.unil.ch/biomapper.

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

Hirzel, A.H., Posse, B., Oggier, P.-A., Glenz, Y.C., Arlettaz, R., 2004. Ecological requirements of a reintroduced species, with implications for release policy: the bearded vulture recolonizing the Alps. J. Appl. Ecol. 41, 1103–1116.

Kéry, M., 2002. Inferring the absence of a species: a case study of snakes. J. Wildlife Manage. 66, 330–338.

Körner, C., 2003. Alpine Plant Life. Springer, Berlin.

Kumar, L., Skidmore, A.K., Knowles, E., 1997. Modelling topographic variation in solar radiation in a GIS environment. Int. J. Geogr. Inf. Sci. 11, 475–497.

Mackenzie, D.I., Royle, J.A., 2005. Designing occupancy studies: general advice and allocating survey effort. J. Appl. Ecol. 42, 1105–1114.

Manel, S., Williams, H.C., Ormerod, S.J., 2001. Evaluating presence–absence models in ecology: the need to account for prevalence. J. Appl. Ecol. 38, 921–931.

Manly, B.F., McDonald, L.L., Thomas, D.L., McDonald, T.L., Erickson, W.P., 2002. Resource Selection by Animals: Statistical Design and Analysis for Field Studies, 2nd ed. Kluwer Academic Publishers, Dordrecht.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, 2nd ed. Chapman & Hall, London.

Ottaviani, D., Lasinio, G.J., Boitani, L., 2004. Two statistical methods to validate habitat suitability models using presence-only data. Ecol. Model. 179, 417–443.

Pearce, J.L., Boyce, M.S., 2006. Modelling distribution and abundance with presence-only data. J. Appl. Ecol. 43, 405–412.

Pearson, R.G., Dawson, T.P., Berry, P.M., Harrison, P.A., 2002. SPECIES: a spatial evaluation of climate impact on the envelope of species. Ecol. Model. 154, 289–300.

Prentice, C.I., Cramer, W., Harrison, S.P., Leemans, R., Monerud, R.A., Solomon, A.M., 1992. A global biome model based on plant physiology and dominance, soil properties and climate. J. Biogeogr. 19, 117–134.

R Development Core Team, 2005. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Randin, C., Dirnböck, T., Dullinger, S., Zimmermann, N.E., Zappa, M., Guisan, A., in press. Are species distribution models transferable in space? J. Biogeogr.

Reutter, B.A., Helfer, V., Hirzel, A.H., Vogel, P., 2003. Modelling habitat-suitability on the base of museum collections: an example with three sympatric *Apodemus* species from the Alps. J. Biogeogr. 30, 581–590.

Scott, J.M., Heglund, P.J., Morrison, M.L., Haufler, J.B., Raphael, M.G., Wall, W.A., Samson, F.B., 2002. Predicting Species Occurences: Issues of Scale and Accuracy. Island Press, Washington.

S-Plus, 1999. S-PLUS 2000 Guide to Statistics. Data Analysis Products Division, MathSoft, Seattle, WA.

Van Houwelingen, J.C., Le Cessie, S., 1990. Predictive value of statistical models. Stat. Med. 9, 1303–1325.
Zaniewski, A.E., Lehmann, A., Overton, J.M., 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. Ecol. Model. 157, 261–280.
Zimmermann, N.E., Kienast, F., 1999. Predictive mapping of alpine grasslands in Switzerland: species versus community approach. J. Veg. Sci. 10, 469–482.
Zweig, M.H., Campbell, G., 1993. Receiver-operating characteristic (Roc) plots—a fundamental evaluation tool in clinical medicine. Clin. Chem. 39, 561–577.