ECOGRAPHY 27: 437–448, 2004

Presence-absence versus presence-only modelling methods for predicting bird habitat suitability

Lluís Brotons, Wilfried Thuiller, Miguel B. Araújo and Alexandre H. Hirzel

Brotons, L., Thuiller, W., Araújo, M. B. and Hirzel, A. H. 2004. Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. – Ecography 27: 437–448.

Habitat suitability models can be generated using methods requiring information on species presence or species presence and absence. Knowledge of the predictive performance of such methods becomes a critical issue to establish their optimal scope of application for mapping current species distributions under different constraints. Here, we use breeding bird atlas data in Catalonia as a working example and attempt to analyse the relative performance of two methods: the Ecological Niche factor Analysis (ENFA) using presence data only and Generalised Linear Models (GLM) using presence/absence data. Models were run on a set of forest species with similar habitat requirements, but with varying occurrence rates (prevalence) and niche positions (marginality). Our results support the idea that GLM predictions are more accurate than those obtained with ENFA. This was particularly true when species were using available habitats proportionally to their suitability, making absence data reliable and useful to enhance model calibration. Species marginality in niche space was also correlated to predictive accuracy, i.e. species with less restricted ecological requirements were modelled less accurately than species with more restricted requirements. This pattern was irrespective of the method employed. Models for wide-ranging and tolerant species were more sensitive to absence data, suggesting that presence/absence methods may be particularly important for predicting distributions of this type of species. We conclude that modellers should consider that species ecological characteristics are critical in determining the accuracy of models and that it is difficult to predict generalist species distributions accurately and this is independent of the method used. Being based on distinct approaches regarding adjustment to data and data quality, habitat distribution modelling methods cover different application areas, making it difficult to identify one that should be universally applicable. Our results suggest however, that if absence data is available, methods using this information should be preferably used in most situations.

L. Brotons (brotons@cefe.cnrs-mop.fr), Inst. Català d'Ornitologia, C./Girona 168, E-08037 Barcelona, Spain (present address: Centre d'Ecologie Fonctionnelle et Evolutive-CNRS, 1919 Route de Mende, F-34293 Montpellier, France). – W. Thuiller and M. B. Araújo, Centre d'Ecologie Fonctionnelle et Evolutive-CNRS, 1919 Route de Mende, F-34293 Montpellier, France. – A. H. Hirzel, Conservation Biology, Inst. of Zoology, Univ. of Bern, Baltzerstr. 6, CH-3012 Bern, Switzerland.

Mapping species distributions is a key issue in ecology and conservation since statement of hypotheses often relies on an accurate knowledge of where species occur. To map species distributions at large spatial scales, different approaches have been adopted the most common of which being the general atlas-distribution framework (Donald and Fuller 1998, Mitchel-Jones et al. 1999, Underhill and Gibbons 2002). The spatial positioning of data from large museum collections may also appear as an alternative in some cases (Peterson

Accepted 30 January 2004

ISSN 0906-7590

et al. 2002). Given the geographical extent of their coverage, the near-equal grid-cell sizes used, and the standardisation of their sampling methodology, atlases are among the most powerful tools available to analyse species' distributions and their governing factors (Donald and Fuller 1998). Nevertheless, most atlases have focused on reporting the occurrence of species and provide relatively poor quantitative information on species abundances or relative suitability of different locations. A recent large scale atlas work (Gibbons et al. 1993) has attempted to obtain quantitative estimates of variation in species abundances (Johnson and Sargeant 2002).

Habitat-suitability or niche-based modelling techniques use information on species locational records environmental factors to generate statistical functions that allow predictions of potentially suitable habitat distribution for species (for a review see Guisan and Zimmerman 2000). The projection of the generated functions to areas where environmental factors are known but species have not been sampled allows an optimal, cost effective, method to map species distributions in large regions and at low spatial resolutions (Hausser 1995, Guisan and Zimmerman 2000, Peterson et al. 2002). The recent development of techniques combined with an increasing availability of large-scale environmental information in digital format offers an opportunity to test and improve methodologies for quantitative mapping of species distributions. Appropriate data on species distributions have already been demonstrated to provide useful information for conservation planning. For instance, species extinctions seem more likely in areas with low suitability or in areas where species are less abundant. Including such information in reserve-selection procedures improves the ability to ensure long-term persistence of species (Araújo et al. 2002). Furthermore, habitat suitability models are increasingly being used to assess the impact of future land use or climate changes (Austin et al. 1996, Buckland et al. 1996, Peterson et al. 2002, Thuiller 2003a), or design ecological networks at large spatial scales (Bani et al. 2002).

There are different methods available to generate habitat suitability maps for species. A major difference between them is the quality of data needed. A first group of methods includes generalised linear models (GLM), generalised additive models (GAM), classification and regression tree analyses, and artificial neural networks (ANN). These methods require good quality presence/absence data in order to generate statistical functions or discriminative rules that allow habitat suitability to be ranked according to distributions of presence and absence of species (Manel et al. 1999, Guisan and Zimmerman 2000). A second group of methods include the Ecological Niche Factor Analysis (ENFA), Bioclim and Domain. These methods require presence data only and were developed to allow use data where knowledge of absences is inadequate or unavailable (Carpenter et al. 1993, Hirzel et al. 2002a, Farber and Kadmon 2003). Such methods rely on the definition of environmental envelopes around locations where species occur, which are then compared to the environmental conditions of background areas (Hirzel et al. 2002a). Using a virtual species with predefined habitat selection preferences, Hirzel et al. (2001) compared model performances of a method relying on species presence only (ENFA) with a method that requires both presence and absence data (GLM). Although both methods provided good predictions of the virtual species distribution, authors found that ENFA had a tendency to outperform in scenarios where species did not occupy all suitable habitat (i.e. many absences were thus modelled within suitable habitat). Conversely, when species were modelled to use all optimal habitats with a high probability and modelled to use sub-optimal habitats with lower probabilities, then GLM was more accurate.

Since data quality is likely to be a key issue affecting reliability of model predictions (Zaniewski et al. 2002, Stockwell and Peterson 2002a, b), knowledge of the predictive performance of methods and their domain of application becomes an important issue at early stages of project-development in surveys aimed at mapping species distributions. At present, we lack extensive tests of the relative performances of methods that compare species distribution models using presence-only or presence/absence data. While use of virtual species is useful as a preliminary exploration of methods' behaviour, it is important to use real data on species distributions to expand understanding of the relative performance of methods. Furthermore, since accurate data on absences is difficult to obtain, especially for mobile or inconspicuous species, it is particularly important to investigate the circumstances that make models using presence-only data to perform at least as well as models using presence/absence data. This should allow for a better understanding of the methods that are more adequate for particular applications. Another generally unexplored question is how dependent is the accuracy of a modelling approach to the ecological characteristics of the species and how these interact with species prevalence (i.e. proportion of occurrences in a data set) to determine model accuracy (but see Manel et al. 2001, Segurado and Araújo 2004). Given that species with more restricted ecological niches are more localised and less frequent, it is expected, even in cases in which data quality is poor, that they are better modelled and thus their distributions more easily predicted that more widespread species (Stockwell and Peterson 2002b, Segurado and Araújo 2004).

Here, we use breeding-bird atlas distribution data as a working example and attempt to analyse the relative performances of these two types of methods on a set of

forest species with similar habitat requirements but with varying prevalence and ecological niche characteristics. We first ask whether methods using presence data only perform equally well than those using presence/absence and whether hypothetical differences in performance hold when evaluating predictive habitat-suitability methods on independent data sets. We then investigate the role of species' ecological niche and prevalence on model accuracy and investigate whether these factors affect model accuracy in interaction with the method used.

Methods

Bird data

In our assessment of habitat-suitability methodologies, we used species occurrence data from a subset of the Catalan Breeding Bird Atlas (Estrada et al. 2004). The CBBA is a large-scale survey that covers the whole of the Catalan region (northeastern Iberian Peninsula, 31 000 km^2, Fig. 1). Within the study area 1550 1×1 km cells were selected (covering ca 6% of the total area extent) to conduct standardised surveys of species presence during the breeding seasons 1999–2001. Cells were selected by volunteers in a stratified fashion assuming that they should cover the main habitat types present in the 10×10 km UTM grid cell in which they were located (Hirzel and Guisan 2002). On each selected cell, two one-hour visits were conducted and the presence of species was investigated. The first visit was made in March–April and the second during May–June to better cover the breeding phenology span of different species. In this paper, we included species with at least 15 occurrences. We then selected a sub-set of species that spanned a range of possible prevalence values (Table 1). Overall we modelled 30 forest species as judged from their habitat selection patterns in the Mediterranean area (Table 1).

Species presence records were assumed to be reliable. The same could not be said for absence records; indeed, failing to detect a species does not guarantee the species is absent from that cell. Presence is a probabilistic function mainly affected by species abundance and detectability. By assuming that a species' detectability is constant across habitats, we considered that absences in this study were either reliable or associated to habitats in which abundance of species was low. However, the assumption that absence indicates areas where species are not present due to a negative species-environmental relationship is not necessarily a valid one. This assumption may not hold for a variety of reasons including

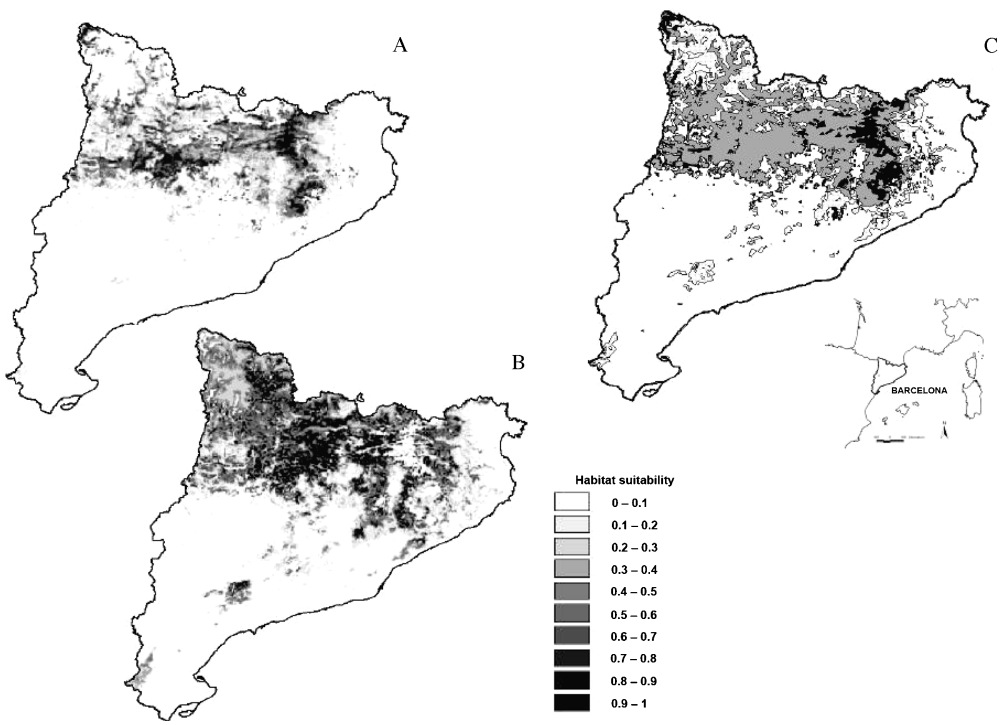

Fig. 1. Maps showing the predicted distributions of GLM (A) and ENFA (B) and the discrepancies between the two methods (C) for one of the forest species used, the nuthatch *Sitta europaea*. The discrepancy map was created by crossing predicted binary presence/absence maps after choosing for each modelling method a threshold maximising specificity and sensitivity. In (C) black cells show areas where ENFA predicted species absence and GLM presence, whereas light grey cells show areas where ENFA predicted presence and GLM predicted absence. Dark grey indicate coincidence in model predictions.

Table 1. Species and model accuracies ±SE as estimated from area under the ROC curve (AUC). § Marginality as estimated by Biomapper algorithm. Species are sorted according to their prevalence which was calculated as proportion of presences within the data set of 1550 cells sampled.

Species names	GLM (AUC)		ENFA (AUC)		Prevalence	Marginality Index §
	Calibration	Evaluation	Calibration	Evaluation		
Emberiza citrinella	0.95±0.036	0.93±0.041	0.88±0.015	0.95±0.024	0.01	0.73
Parus palustris	1.00±0.002	0.94±0.028	0.82±0.045	0.91±0.013	0.01	0.93
Phoenicuros phoenicuros	0.91±0.027	0.57±0.182	0.67±0.012	0.66±0.111	0.02	0.65
Regulus regulus	0.99±0.003	0.91±0.044	0.94±0.043	0.91±0.091	0.03	0.99
Sylvia hortensis	0.89±0.021	0.76±0.069	0.72±0.004	0.68±0.056	0.03	0.59
Anthus trivialis	0.90±0.024	0.79±0.066	0.80±0.028	0.76±0.071	0.04	0.69
Prunella modularis	1.00±0.001	0.91±0.001	0.94±0.019	0.89±0.053	0.05	0.88
Sylvia borin	0.82±0.022	0.79±0.043	0.71±0.028	0.72±0.028	0.06	0.66
Sitta europaea	0.93±0.010	0.91±0.016	0.82±0.023	0.81±0.037	0.11	0.77
Muscicapa striata	0.76±0.022	0.68±0.040	0.58±0.014	0.60±0.019	0.12	0.62
Emberiza cia	0.88±0.011	0.86±0.019	0.76±0.025	0.75±0.041	0.21	0.55
Turdus philomelos	0.81±0.015	0.78±0.022	0.70±0.016	0.71±0.024	0.25	0.57
Phylloscopus collybita	0.84±0.014	0.75±0.027	0.69±0.017	0.64±0.026	0.25	0.62
Parus ater	0.92±0.009	0.89±0.016	0.67±0.017	0.64±0.027	0.27	0.66
Sylvia cantillans	0.83±0.013	0.81±0.020	0.69±0.017	0.68±0.025	0.29	0.47
Oriolus oriolus	0.77±0.015	0.68±0.026	0.66±0.013	0.62±0.021	0.29	0.54
Lullula arborea	0.87±0.011	0.78±0.021	0.73±0.017	0.70±0.027	0.30	0.47
Regulus ignicapillus	0.88±0.010	0.84±0.018	0.81±0.016	0.76±0.025	0.35	0.60
Parus cristatus	0.83±0.012	0.84±0.018	0.74±0.013	0.76±0.022	0.37	0.52
Streptopelia turtur	0.81±0.013	0.81±0.020	0.72±0.014	0.69±0.021	0.41	0.56
Aegithalos caudatus	0.86±0.011	0.83±0.019	0.78±0.015	0.81±0.023	0.38	0.57
Phylloscopus bonelli	0.84±0.012	0.79±0.021	0.69±0.016	0.69±0.025	0.43	0.47
Troglodytes troglodytes	0.87±0.010	0.80±0.020	0.77±0.014	0.72±0.023	0.48	0.54
Sylvia melanocephala	0.93±0.007	0.90±0.014	0.75±0.015	0.75±0.022	0.50	0.63
Parus caeruleus	0.88±0.010	0.82±0.020	0.77±0.014	0.76±0.023	0.51	0.52
Garrulus glandarius	0.85±0.011	0.81±0.020	0.76±0.015	0.72±0.025	0.55	0.51
Emberiza cirlus	0.84±0.012	0.81±0.020	0.75±0.015	0.73±0.024	0.55	0.47
Erithacus rubecula	0.91±0.009	0.90±0.014	0.83±0.013	0.83±0.020	0.57	0.53
Luscinia megharynchos	0.83±0.014	0.80±0.023	0.70±0.018	0.69±0.027	0.66	0.51
Turdus merula	0.88±0.015	0.85±0.022	0.60±0.016	0.64±0.024	0.85	0.41

habitat population dynamics, fragmentation, rate of dispersal or history, which may induce species absence from otherwise optimal habitat (Loehle and LeBlanc 1996, Araújo and Williams 2000). If the role of such events is significant and the species is not in equilibrium with its environment, absence data may affect model building whatever methodology is used, but this will be specially true when absence data is also included in the model calibration. On the other hand, if absences are indeed related to low suitable habitat for the species (i.e. the species is near the equilibrium with the environment), the information provided by them should improve the performance of methods relying on both presence and absence data (Hirzel et al. 2001).

Environmental data

Environmental variables (ENV) were generated from available GIS (Geographical Information Systems) layers. Habitat composition was analysed from land-use layers generated by the Cartographic Institute of Catalonia (ICC) and Agriculture Department (DARP, Table 2). After successive processes of simplification and classification, land-use maps were resampled to a 50 m pixel resolution and converted to several boolean maps (i.e. one per each land use category) which allowed the generation of final variables describing each 1×1 km cell (Table 2).

We also used climatic variables (temperature, precipitation and solar radiation) which were obtained from the Catalan Digital Atlas (CDA, Ninyerola et al. 2000), whereas data on topography was obtained from a Digital Elevation Model (DEM) generated by the ICC from topographic 1:50 000 maps. To obtain a value for each cell we calculated the mean value for all pixels (200 m side) in that cell (Table 2). We finally used three more variables that allowed the detection of geographic patterns in species distributions that were not captured by habitat or climatic ENV. These variables were the mean latitude and longitude co-ordinate for each cell and the mean distance to the sea (Table 2).

Statistical models

Methods based on presence/absence data

Different methods have been envisaged to build predictive models based on presence/absence data. Amongst them generalised linear models have been extensively tested elsewhere and have proved robust in a number of independent situations (Manel et al. 1999, Pearce and

Table 2. Environmental variables (ENV) used to generate habitat suitability models of the 30 forest bird species used in the comparison of methods. Unless otherwise mentioned, variables referred to 1 × 1 km squares correspond to means obtained from averaging individual values from pixels contained each 1 × 1 km square. Cartographic sources are indicated when necessary.

Descriptor type	Variable description [units]	Range
Forest		
	Coniferous forest [1]	0–400
	Esclerophylous [1]	0–400
	Deciduous forest [1]	0–400
	Pinus halepensis forest [2]	0–400
	Pinus sylvestris forest [2]	0–400
	Abies alba-*Pinus uncinata* forest pixels in 1 × 1 km squares [2]	0–400
	Pinus nigra forest [2]	0–400
	Other *Pinus* forest [2]	0–400
	Quercus suber forest [2]	0–400
	Quercus ilex forest [2]	0–400
	Quercus humilis forest [2]	0–393
	Other deciduous forest [2]	0–400
	Distance to nearest forest patch [log m] [2]	0–10
Agriculture		
	Dry herbaceous cropland (cereals) [1]	0–400
	Irrigated herbaceous cropland pixels (corn) [1]	0–400
	Dry arboreal cropland (olive tree, almond) [1]	0–400
	Irrigated arboreal cropland (fruit trees) [1]	0–400
	Vineyard [1]	0–400
Low vegetation cover		
	Scrub [1]	0–400
	Bare ground (rocks) [1]	0–400
Landscape		
	Number of land uses in 1 × 1 km squares (based on land use cover 1997, urban and industrial categories clumped) [1]	1–11
Human impact		
	Low density urbanization [1]	0–190
	Distance to cities > 10 000 inhabitants [log m] [2]	0–11
	Infrastructure (transport network and urban areas) [1]	0–400
	Distance to main roads of the primary road network [log m] [1]	0–10
	Distance to roads of the secondary road network [log m] [1]	0–10
Climate		
	Mean solar radiation [3] [10 $kJm^2 \times day^{-1}$]	19–961
	Mean accumulated summer precipitation (June–September) [$l \times m^{-2}$] [3]	50–500
	Mean accumulated of mean winter temperatures (December–March) [°C] [3]	−50–105
Topography		
	Mean altitude [m] [1]	0–2850
	Mean slope [degrees] [1]	0–39
Geography		
	Mean latitude [degrees] [1]	2.70–3.80
	Mean longitude [degrees] [1]	45.70–46.50
	Mean distance to the sea [km] [1]	33–100

[1] 'Institut Cartogràfic de Catalunya' (ICC).
[2] 'Departament de Medi Ambient de la Generalitat de Catalunya' (DAM).
[3] 'Centre de Recerca Ecologica i Aplicacions Forestals' (CREAF).

Ferrier 2000, Osborne et al. 2001, Thuiller et al. 2003). Therefore, to analyse binary data such as the presence/absence of species within each sampled cell, we applied generalised linear regression techniques with binomial error distribution (logistic regression, GLM, McCullagh and Nelder 1989). We included as potential predictors in model building all linear and quadratic terms, which excluded environmental predictors showing correlations > 0.9. To select the most parsimonious model, we used an automatic stepwise model-selection procedure starting from a null model containing the intercept only. The "step.glm" function in S-Plus builds models by adding new terms and investigating how much they improve the fit, and by dropping terms that do not degrade the fit by a significant amount (Anon. 1999). Quadratic terms were included only if they improved their linear counterpart. The statistic used to select the final model was the Akaike Informaton Criteria (AIC, Chambers and Hastie 1997). It is important to stress GLM was used with a predictive rather than inductive goal. In such circumstances accuracy of model predictions is more important than significance of particular ecological terms (Legendre and Lengendre 1998).

Methods requiring presence data only

Different methods have been proposed to predict species distributions based on presence data only. These methods search for an "environmental envelope"

characteristic of the points in which the species is present in order to extrapolate to the remaining area under study (Guisan and Zimmerman 2000). To analyse these kinds of data we used the Ecological Niche Factor Analysis (ENFA) released in the BIOMAPPER package (Hirzel et al. 2002b). ENFA quantifies the niche occupied by a species by comparing its distribution in ecological space ("the species distribution") with the distribution of all cells (the "global distribution") (Hirzel et al. 2002a). ENFA focuses on the marginality of the species (how the species mean differs from the global mean) and environmental tolerance (how the species variance compares to the global variance). Species marginality gives indication of the species niche position whereas species tolerance it is negatively associated to species specialisation and refers to its niche width, or breadth. ENFA uses a factor analysis with orthogonal rotations to 1) transform the predictor variables to a set of uncorrelated factors, and 2) to construct axes in a way that accounts for all marginality of the species in the first axis, and that minimizes tolerance in the following axes. There are different algorithms available in BIOMAPPER to build habitat suitability maps from ENFA analysis (Hirzel et al. 2002b). Following Hirzel and Arlettaz (2004) we used the geometric mean algorithm, which takes into account the density of observation points in environmental space by computing the geometric mean to all observation points. We used a Box-Cox transformation of the environmental variables to enhance normality except in the cases when transformation produced near binary outcomes (Hirzel et al. 2002a).

Marginal species are likely to be less tolerant in most conditions, and species marginality and tolerance were indeed highly correlated in our set of forest species ($r = -0.76$, $p < 0.0001$). Furthermore, ecologically marginal species may tend to be less tolerant to changes in ecological conditions leading to restricted distributions. Species marginality and species prevalence were also significantly correlated in our data set ($r = -0.68$, $p < 0.0001$). In order to allow the independent assessment of the different components of species niche and prevalence, we conducted a Principal Component Analysis using species marginality, tolerance and prevalence as original variables. After a varimax transformation of the principal components maximising their correspondence to the original variables, we succeed to obtain two independent components: 1) a marginality component positively associated to species marginality ($r = 0.90$) and more weakly, negatively to tolerance ($r = -0.60$), and 2) a prevalence component identifying a gradient of species prevalence ($r = 0.90$) parallel to that of species tolerance ($r = 0.70$), separating less tolerant and scarcer species from more tolerant and abundant ones. These two components were finally used as predictors of model accuracy in further analyses.

Evaluation of habitat suitability models

We used cross-validation to evaluate predictive model accuracy and divided the data in two different sets, by randomly assigning 70% of occurrence values for each species to a calibration data-set and 30% of the remaining occurrences to an independent evaluation data set. The calibration data set was used to develop the habitat model that was evaluated on the evaluation data set (Fielding and Bell 1997).

There are practical difficulties in evaluating predictions from presence-only data models with traditional evaluation methods (Pearce and Ferrier 2000) given that absence data is usually missing and therefore can not be used to evaluate model predictions. A possible method is to compare the suitability of areas where the species is present with that of the background environment (Hirzel et al. 2001). Other authors have used correlations with known, or reference distributions, to evaluate models performance (Hirzel et al. 2001, Zaniewski et al. 2002, Boyce et al. 2002). However, in our case complete or reference species distributions were unavailable. Predic ting species absences is an important issue even when information has not been explicitly incorporated into model development (Stockwell and Peterson 2002a). Therefore, we assessed the accuracy of both ENFA and GLM models on the calibration and evaluation data sets using both presences and absences. By means of misclassification, results from probabilistic models are often judged as successful if predicted probabilities > 0.5 correspond with observed occurrences and values < 0.5 with absences and prediction errors (false positives and false negatives) are low. However, this dichotomy is arbitrary and lacks any ecological justification. A more powerful approach is to assess model success across a range of dichotomies from different cut-off points using the receiver operating characteristics (ROC) plots. The ROC plot is based on a series of misclassification matrices computed for a range of cut-offs from 0 to 1. It then plots on the y-axis the true positive fraction, against the false positive fraction from the same misclassification matrix (Fielding and Bell 1997, Pearce and Ferrier 2000). The area under the ROC curve (AUC) is a convenient measure of overall fit and commonly varies between 0.5 (for chance performance) and 1 (perfect fit). We obtained AUC and its standard error with a custom function in S-Plus software (Anon. 1999).

Comparison of accuracy between modelling methodologies

We first test for overall differences between modelling method (GLM vs ENFA) and data-set (calibration vs evaluation) by means of repeated measures ANOVA using modelling method and data-set as within-subject factors in the design according to species. We then used

repeated measures ANCOVA designs to assess how accuracy of habitat models varied between method and data-set using these factors as within-factors subjects but also adding to the design the two principal components summarising niche characteristics and species prevalence (i.e. the marginality component and the prevalence component).

Results

Overall accuracies of models

Overall model accuracy estimated with the ROC method performed better than random in every case analysed (Table 1). AUC values were higher for GLM models than for ENFA models (Table 3, Figs 2 and 3) and were also higher when evaluated for the calibration data compared to the evaluation data set (Table 3, Fig. 2). We also found that change in predictive accuracies between the calibration and the evaluation data sets was larger for GLM than for ENFA models, indicating that the loss in predictive performance when applied to an independent data set not used for model construction is higher for GLM (Table 3, Fig. 2). In addition to overall differences in predictive accuracy, we detected considerable variation in species spatial distributions projected with GLM and ENFA (Fig. 1). There was a general tendency for ENFA to overestimate the spatial extent of the distributions, especially on the edges of those estimated by GLM; in some cases areas estimated to have high species' probabilities of occurrence with GLM were overlooked by ENFA (e.g. Fig. 1).

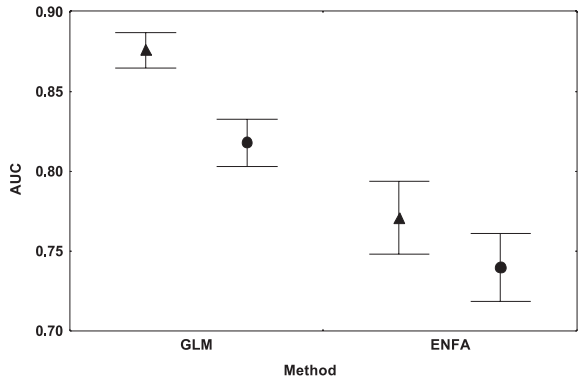

Fig. 2. Mean values of model accuracies for GLM and ENFA methods applied on both the calibration (70% of sample cells, triangles) and the evaluation data (remaining 30% of sample cells, circles, n = 30 species). Whiskers represent the standard error of the estimates.

Effects of species niche characteristics and prevalence on model performance

Species distributions in ecological space had a major role in determining species model accuracies, with AUC being generally higher for more marginal species (marginality component, Table 3). The observation that marginal species were modelled more accurately was coincident for the two methods tested and for the data sets used as indicated by the lack of significant interactions, which suggests that this effect was robust to methodological considerations (Table 3, Fig. 4).

The prevalence component did not have an overall consistent effect on model performance (Table 3, Fig. 5). However, there was a significant difference in the effect of this factor on model performance between methods in

Table 3. Repeated-measures ANCOVA conducted on the predictive model accuracies of GLM and ENFA models on 30 forest species in Catalonia. The within subject effects considered are method (two levels, GLM vs ENFA, see section methods) and data-set (two levels, calibration vs evaluation). Species marginality component and the prevalence component (see methods) were used as continuous predictors in the ANCOVA analyses. Significant results are emphasized in **bold**.

Source of variation	Model accuracy (AUC)		
	DF	F	p
Between subject effects			
Marginality component	**1**	**33.61**	**< 0.0001**
Prevalence component	1	0.02	0.89
Error	27		
Within subject effects			
Method	**1**	**141.05**	**< 0.0001**
Method × Marginality component	1	2.15	0.15
Method × Prevalence component	1	3.54	0.07
Error	27		
Data-set	**1**	**23.64**	**< 0.0001**
Data-set × Marginality component	1	0.03	0.87
Data-set × Prevalence component	**1**	**4.71**	**< 0.05**
Error	27		
Method × Data-set	**1**	**25.69**	**< 0.0001**
Method × Data-set × Marginality component	1	0.49	0.49
Method × Prevalence component	**1**	**11.29**	**< 0.001**
Error	27		

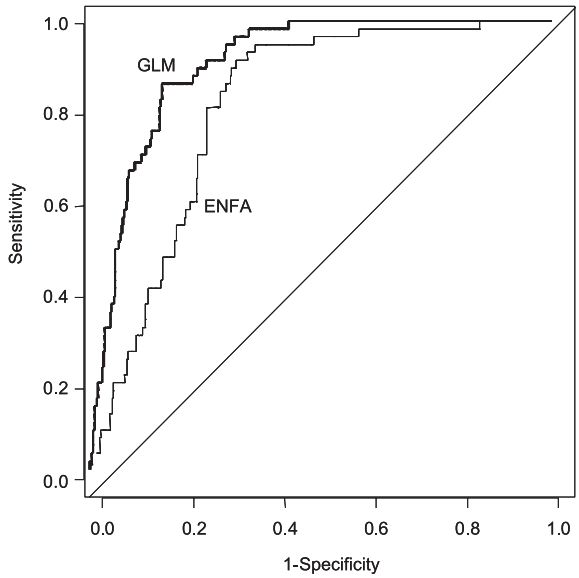

Fig. 3. ROC plot for GLM and ENFA methods for the evaluation data set on one of the species analysed, the nuthatch, *Sitta europaea* (see Fig. 1 for predicted distributions, and Table 1 for actual ROC values). Sensitivity represents the true positive fraction and 1-specificity the false positives fraction for a range of cut-offs used to classified modelled probabilities into presence absence data.

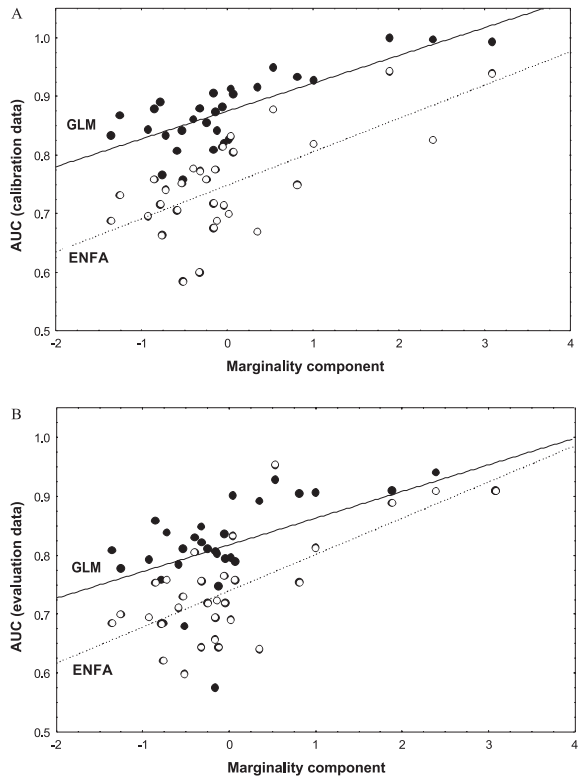

Fig. 4. Relationship between model accuracies for GLM and ENFA methods and species marginality component (index describing how far from the mean conditions of the study area the species optimum is found, n = 30 species). A. Calibration data (70% of sample cells). B. Evaluation data (remaining 30% of sample cells).

the evaluation data set but not in the calibration data set, with higher values of the prevalence component associated with higher AUC values in GLM models but not in ENFA, which remained unaffected by this factor (Table 3, Fig. 5). This effect resulted in a stronger overall positive relationship between the prevalence component and predictive accuracy in the evaluation data set than in the calibration data set (Table 3, Fig. 5).

Discussion

Our results showed that GLM using both presence and absence data predicted the distribution of forest species with higher accuracy than ENFA, which used presence data only. This supports the view that the forest species analysed used available habitats proportionally to their suitability, making absence data reliable and useful to enhance model calibration This is in line with the results obtained by Hirzel et al. (2001) using a modelling approach based on a virtual species with predetermined habitat preferences. The authors found that GLM performed significantly better than ENFA when estimating habitat suitability in an overabundance scenario in which species occupied all optimal habitats and occupied secondary habitats at lower probabilities. In this scenario, absence data is likely to be reliable and help to "fix the floor" of what is unsuitable habitat for each focal species. First, by giving a low weighting to occurrences in low-density habitats, absence data helps to identify low suitability areas that may have otherwise been classified as good habitats if only presence data were used. Occurrences in good but scarce habitats may also bias models based on presence only data because relative importance of such habitats may be overweighted by a larger number of observations in other habitat types. For instance in the case of the nuthatch, deciduous forest areas, which cover a limited surface within the study region, were ranked as low suitability by the presence only method. Indeed this species had a small number of occurrences in such areas that were overridden by the greater number of occurrences in other more abundant habitats (Fig. 1). Here the availability of absences may become critical to correctly assess the relative suitability of these areas in comparison with other areas equally suitable but where presences are more common due to the relative availability of different habitats in the area. Some authors have suggested that when lacking absence data, distribution models may be improved by generating random pseudo-absences from

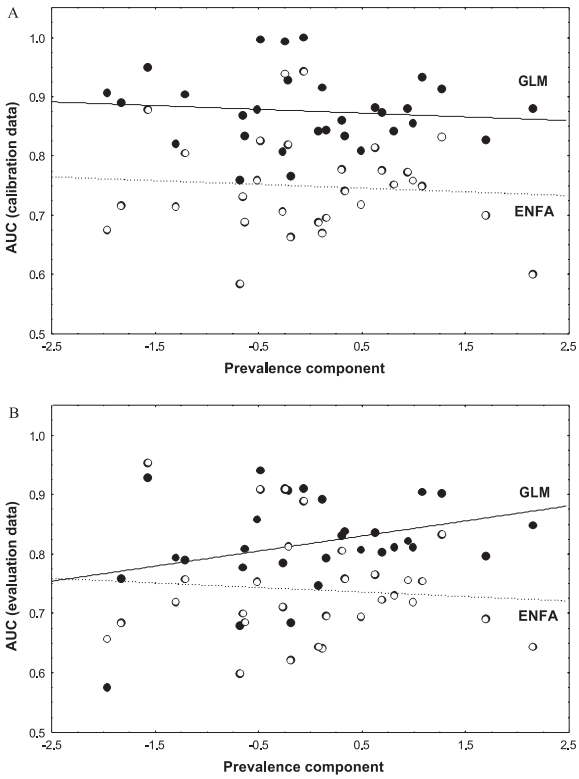

Fig. 5. Relationship between model accuracies for GLM and ENFA methods and the prevalence component (gradient separation abundant tolerant species from scarce less tolerant species, n = 30 species). A. Calibration data (70% of sample cells). B. Evaluation data (remaining 30% of sample cells).

background areas in which species data is missing (Stockwell and Peterson 2002a). However, this method may result in biased absence data if species are widespread or presence data is scarce (Boyce et al. 2002).

The approach employed to test model accuracy assumes that identification and prediction of locations from which species are absent is important. This is likely to be the case in predictive cartographic habitat modelling (e.g. vertebrate atlas studies), but may not be the case if identification of ecological mechanisms are more specific (i.e. selecting optimal areas for species reintroduction). In cases with poor data availability or assumptions of species equilibrium are strongly violated (i.e. use of museum data to produce distributions at large spatial scales, or distribution of invading or spreading species in which absence data is ecologically meaningless), evaluation of absence data becomes meaningless. Presence only methods may then make best use of available presence data. They also have the advantage of reducing the contributions of non-equilibrium factors in model predictions (Hirzel et al. 2002a). Recently, Zaniewski et al. (2002) have shown that although presence-absence based methods were more discriminant than ENFA, at a species level, they appeared to be less suitable to identify areas with high conservation concern in a multi-specific perspective. For example, if the objective were to protect rare or endangered species overestimating areas of potentially elevated biodiversity might be preferable than underestimating their existence and presence only methods may be useful (Zaniewski et al. 2002). In this case, however one should proceed with caution because optimistic predictions proved false may artificially increase the cost of conservation strategies (Araújo and Williams 2000).

Sudden changes in habitat quality may occur under natural conditions (Gates and Donald 2000) resulting in individuals not using optimal habitats or being present in low quality areas. Caution in the use of habitat suitability methods should be adopted if strong suspicions of non-equilibrium situations are expected. However, in large scale distribution modelling, most species, especially in rather mobile groups such as birds, are likely to be close to equilibrium with environmental conditions due to population dynamics and habitat selection mechanisms (Chamberlain and Fuller 1999). This is likely to be the case if factors causing non-equilibrium are related mainly to dispersal. In these cases, absences are likely to reflect low habitat suitability and therefore improve model performance. We argue that using of absence data in building presence/absence models is generally more appropriate than using presence only data. This should be particularly true when using data from intensive collection studies, such as breeding bird atlases where an important number of absences are indeed expected to be true absences and reflect low habitat suitability. Ecological interpretation of different habitat modelling methodologies is of great importance and may guide the final choice of available alternatives. Zaniewski et al. (2002) argue that pure presence-only methods such as ENFA are more likely to predict potential distributions that more closely resemble the fundamental niche of the species, whereas presence-absence modelling is more likely to reflect the present natural distribution derived from realized niche. However, both methods aim at predicting distributions by sampling real distributions, and therefore, they provide different estimations of the realised niche of the species (Loehle and LeBlanc 1996). Since presence only methods do not take into account the areas from which the species might be absent, they are less conservative in estimating the species' realised niche. On the other hand, they may better capture realised niche responses in species which are far from equilibrium with the environment and therefore are not yet using all habitats corresponding to their realized niche (Hirzel et al. 2001). It is important to emphasize that, being based on distinct approaches regarding adjustment to data and variation in data quality, habitat distribution modelling methods will likely cover different application areas and

it will be impossible to identify one among them as universally applicable (Elith and Burgman 2002a, b, Segurado and Araújo 2004). Therefore, the goals and assumptions of habitat modelling should be clear before they are applied to particular situations. Methods based on presence only data such as ENFA appear to fully cover habitat modelling focused on data in which absence data is not available, or when the main objective of the modelling is to identify overall suitable areas for a given species (i.e. the current distribution of the species is certainly unreliable). Otherwise, methodologies employing presence/absence methods should be prioritised.

Species niche characteristics, prevalence and model accuracy

We found that ecological niche position (marginality) plays a key role in determining predictive accuracy in models developed with both GLM and ENFA. In particular, less marginal bird species from which selected habitats differed little from the available environmental conditions in the study area were modelled less accurately than more marginal selective species. This result agrees with the results of Hepinstall et al. (2002) and Stockwell and Peterson (2002b) who also observed that the performance of bird habitat models was negatively correlated with the proportion of habitats used by a species with more generalist species being poorly modelled. Segurado and Araújo (2004) described similar pattern of increasing accuracy of model predictions for marginal amphibian and reptile species in Portugal. Stockwell and Peterson (2002b) offered as a biological explanation for this observation that widespread species often show local or regional differences in ecological characteristics. Modelling all these sub-populations together would effectively overestimate the species' ecological breath and hence reduce model accuracy. Therefore, the more widespread a species is, the more likely it is to use different habitats thus increasing the likelihood that more factors determine its distributions (Osborne and Suárez-Seoane 2002). Another potentially simpler explanation is that species described to have wider distributions or use a wide range of habitats in one area might not be limited by any of the measured predictive factors at the scale at which models are fitted. In both cases, an accurate prediction of species distributions becomes difficult and will benefit from availability of absence data to determine relative suitability among available habitats. By contrast, both GLM and ENFA methods seem to perform equally well on more marginal species, which offers a promising background to the development of models of marginal potentially threatened species from sources of poor quality data (Peterson et al. 2002).

A major but variable role of prevalence on the predictive accuracy of habitat models has been stressed by several studies (Araújo and Williams 2000, Pearce and Ferrier 2000, Manel et al. 2001, Stockwell and Peterson 2002b). For example, Araújo and Williams (2000) found that prevalence affected negatively the specificity component of model predictive ability (i.e. increasing false positives), while it would affect positively the sensitivity component of model predictive ability (i.e. reducing false negatives). On the other hand, Manel et al. (2001) found that predictive model accuracy assessed with the ROC method was independent of prevalence (an observation that was not supported by Segurado and Araújo 2004). However, a critical assessment of the effects of prevalence on model predictive accuracy is problematic because prevalence is likely to vary both with species ecological characteristics and relative sampling effort. More marginal, or less tolerant, species will tend to be less frequent and therefore, relatively fewer occurrences will be available than for species with a wider ecological distribution. On the other hand, relative lower sampling effort or bias in data collection may also decrease species prevalence. Prevalence is thus likely to affect model accuracy more strongly via indirect effects of species ecology. Thuiller (2003b) found that within a given species, accuracy is independent of prevalence supporting the view that among species effects of prevalence on model accuracy are likely to be associated to variability in species niche characteristics. In our study, we could not completely isolate prevalence from this factor. However, we found that independently of the marginality component, the effect of the prevalence component may still play a secondary role on predictive model accuracy (Hirzel et al. 2001, Karl et al. 2002). This role suggests that the effect of prevalence on predictive accuracy is moderately stronger in models using presence/absence data, because a relative increase in the amount of information derived from the additional presences may enhance its ability to discriminate the quality of the different sites. When using a presence only method an increase in the number of occurrences analysed did not render similar benefits to model accuracies. Indeed, using a virtual species, Hirzel et al. (2001) already showed that independently of data quality ENFA appeared to be robust to data quantity. This is supported by our results that the prevalence component did not affect accuracies of ENFA models independently from species ecology. In Hirzel et al. (2001), GLM was also found to be relatively robust to data quantity. However, our results suggested that higher prevalence for a given species ecology may enhance model accuracy on independent test data, raising the issue of the importance of testing habitat predictive models on evaluation data tests not used for model development (Fielding and Bell 1997, Beutel et al. 1999, Hirzel et al. 2001). Future studies should explicitly assess the influ-

ence of the relationship between sample size and ecology on the relative performance of habitat suitability models based on presence and absence methods.

Acknowledgements – We thank all volunteers that contributed to the collection of data for the Catalan Breeding Bird Atlas (CBBA), especially the two coordinators of the project, Joan Estrada and Vittorio Pedrocchi for their effort on data collection and management. The CBBA was financed by the "Departament de Medi Ambient de la Generalitat de Catalunya", "Fundació Territori i Paisatge" and "Sociedad Española de Ornitología". We thank G. Siriwardena and M. Mönkkönen for improving earlier drafts of the manuscripts. This research is a contribution to the Montpellier-Barcelona LEA "Mediterranean Ecosystems in a Changing World" and has been supported by a Marie Curie Fellowship of the European Community programme Improving Human Potential under the contract number HPMF-CT-2002-01987.

References

Anon. 1999. S-Plus 2000 Guide to statistics. Vol. 1. – Mathsoft, Seattle.

Araújo, M. B. and Williams, P. H. 2000. Selecting areas for species persistence using occurrence data. – Biol. Conserv. 96: 331–345.

Araújo, M. B., Williams, P. H. and Fuller, R. J. 2002. Dynamics of extinction and the selection of nature reserves. – Proc. Roy. Soc., Biol. Sci. 269: 1971–1980.

Austin, G. E. et al. 1996. Predicting the spatial distribution of buzzard *Buteo buteo* nesting areas using a Geographical Information System and remote sensing. – J. Appl. Ecol. 33: 1541–1550.

Bani, L. et al. 2002. The use of focal species in designing a habitat network for a lowland area of Lombardy, Italy. – Conserv. Biol. 16: 826–831.

Beutel, T. S., Beeton, R. J. S. and Baxter, G. S. 1999. Building better wildlife-habitat models. – Ecography 22: 219–223.

Boyce, M. et al. 2002. Evaluating resource selection functions. – Ecol. Modell. 157: 281–300.

Buckland, S. T., Elston, D. A. and Beaney, S. J. 1996. Predicting distributional change, with application to bird distributions in northeast Scotland. – Glob. Ecol. Biogeogr. Lett. 5: 66–84.

Carpenter, G., Gillison, A. N. and Winter, J. 1993. Domain: a flexible modelling procedure for mapping potential distributions of plants and animals. – Biodiv. Conserv. 2: 667–680.

Chamber, J. M. and Hastie, T. J. 1997. Statistical models in S. – Chapman and Hall.

Chamberlain, D. E. and Fuller, R. J. 1999. Density-dependent habitat distribution in birds: issues of scale, habitat definition and habitat availability. – J. Avian Biol. 30: 427–436.

Donald, P. F. and Fuller, R. 1998. Ornithological atlas data: A review of uses and limitations. – Bird Study 45: 129–145.

Elith, J. and Burgman, M. A. 2002a. Predictions and their validation: rare plants in the Central Highlands, Victoria, Australia. – In: Scott, M. S. et al. (eds), Predicting species occurrences: issues of accuracy and scale. Island Press, pp. 303–319.

Elith, J. and Burgman, M. A. 2002b. Habitat models for PVA. – In: Brigham, C. A. and Schwartz, M. W. (eds), Population viability in plants. Springer, pp. 203–238.

Estrada, J. et al. 2004. Catalan breeding bird atlas (1999–2002). – Inst. Catalá d'Ornitologia, Barcelona, in press.

Farber, O. and Kadmon, R. 2003. Assessment of alternative approaches for bioclimatic modelling with special emphasis on the Mahalanobis distance. – Ecol. Modell. 160: 115–130.

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

Gates, S. and Donald, P. F. 2000. Local extinction of British farmland birds and the prediction of further loss. – J. Appl. Ecol. 37: 806–820.

Gibbons, D. W., Reid, J. B. and Chapman, R. A. 1993. The new atlas of breeding birds in Britain and Ireland: 1988–1991. – T and AD Poyser.

Guisan, A. and Zimmerman, N. E. 2000. Predictive habitat distribution models in ecology. – Ecol. Modell. 135: 147–186.

Hausser, J. 1995. Mammifères de Suisse. – Birkhäuser.

Hepinstall, J. A., Krohn, W. B. and Sader, S. A. 2002. Effects of niche width on the performance and agreement of avian habitat models. – In: Scott, M. S. et al. (eds), Predicting species occurrences: issues of accuracy and scale. Island Press, pp. 593–606.

Hirzel, A. H. and Guisan, A. 2002. Which is the optimal sampling strategy for habitat suitability modelling. – Ecol. Modell. 157: 331–341.

Hirzel, A. H. and Arlettaz, R. in press. Modelling habitat suitability for complex species distributions by the environmental-distance geometric mean. – Environ. Manage.

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

Hirzel, A. H., Hausser, J. and Perrin, N. 2002a. Biomapper 2.0. Div. of Conservation Biology, Bern, < http://www.unil.ch/biomapper > .

Hirzel, A. H. et al. 2002b. Ecological-niche factor analysis: How to compute habitat- suitability maps without absence data? – Ecology 83: 2027–2036.

Johnson, D. H. and Sargeant, G. A. 2002. Towards better atlases: improving presence-absence information. – In: Scott, M. S. et al. (eds), Predicting species occurrences: issues of accuracy and scale. Island Press, pp. 391–398.

Karl, J. W. et al. 2002. Species commonness and the accuracy of habitat-relationship models. – In: Scott, M. S. et al. (eds), Predicting species occurrences: issues of accuracy and scale. Island Press, pp. 573–580.

Legendre, P. and Legendre, L. 1998. Numerical ecology, 2nd ed. – Elsevier.

Loehle, C. and LeBlanc, D. 1996. Model-based assessments of climate change effects on forests: a critical review. – Ecol. Modell. 90: 1–31.

Manel, D. et al. 1999. Alternative methods for predicting species distribution: an illustration with Himalayan river birds. – J. Appl. Ecol. 36: 734–747.

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

McCullagh, P. and Nelder, J. A. 1989. Generalized Linear Models. – Chapman Hall.

Mitchell-Jones, A. J. et al. 1999. Atlas of European mammals. – Academic Press.

Ninyerola, M., Pons, X. and Roure, J. M. 2000. A methodological approach of climatological modelling of air temperature and precipitation through GIS techniques. – Int. J. Climatol. 20: 1823–1841.

Osborne, P. E. and Suárez-Seoane, S. 2002. Should data be partitioned before building large-scale distribution models? – Ecol. Modell. 157: 249–259.

Osborne, P. E., Alonso, J. C. and Bryant, R. G. 2001. Modelling landscape-scale habitat use using GIS and remote sensing: a case study with great bustards. – J. Appl. Ecol. 38: 458–471.

Pearce, J. and Ferrier, S. 2000. An evaluation of alternative algorithms for fitting species distribution models using logistic regression. – Ecol. Modell. 128: 127–147.

Peterson, A. T. et al. 2002. Future projections for Mexican faunas under global change scenarios. – Nature 416: 626–629.

Segurado, P. and Araújo, M. B. 2004. Evaluation of methods for modelling species probabilities of occurrence. – J. Biogeogr, in press.
Stockwell, D. R. and Peterson, A. T. 2002a. Controlling bias in biodiversity. – In: Scott, M. S. et al. (eds), Predicting species occurrences: issues of accuracy and scale. Island Press, pp. 537–546.
Stockwell, D. R. and Peterson, A. T. 2002b. Effects of sample size on accuracy of species distribution models. – Ecol. Modell. 148: 1–13.
Thuiller, W. 2003a. BIOMOD: Optimising predictions of species distributions and projecting potential future shifts under global change. – Global Change Biol. 9: 1353–1362.
Thuiller, W. 2003b. Impacts of global climate change on biodiversity in Europe: projections and uncertainties. – Thesis, Univ. of Montpellier, Montpellier.
Thuiller, W., Araújo, M. B. and Lavorel, S. 2003. Generalized Models versus Classification Tree Analysis: a comparative study for predicting spatial distributions of plant species at different scales. – J. Veg. Sci. 14: 669–680.
Underhill, L. and Gibbons, D. 2002. Mapping and monitoring bird populations: their conservation uses. – In: Norris, K. and Pain, D. J. (eds), Conserving bird biodiversity: general principles and applications. Univ. Press.
Zaniewski, A. E., Lehman, A. and Overton, J. 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. – Ecol. Modell. 157: 261–280.