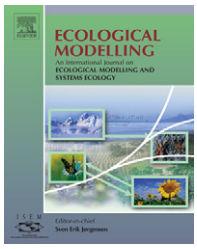
available at www.sciencedirect.com

ScienceDirect

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



Assessing the effects of pseudo-absences on predictive distribution model performance

*Rosa M. Chefaoui, Jorge M. Lobo**

Dpto. de Biología Evolutiva y Biodiversidad, Museo Nacional de Ciencias Naturales, c/ José Gutiérrez Abascal 2, 28006 Madrid, Spain

ARTICLE INFO

Article history:
Received 4 January 2007
Received in revised form 9 July 2007
Accepted 15 August 2007
Published on line 24 September 2007

Keywords:
Pseudo-absences
Distribution models
Model accuracy
Non-equilibrium
Graellsia isabelae
Iberian Peninsula

ABSTRACT

Modelling species distributions with presence data from atlases, museum collections and databases is challenging. In this paper, we compare seven procedures to generate pseudo-absence data, which in turn are used to generate GLM-logistic regressed models when reliable absence data are not available. We use pseudo-absences selected randomly or by means of presence-only methods (ENFA and MDE) to model the distribution of a threatened endemic Iberian moth species (*Graellsia isabelae*). The results show that the pseudo-absence selection method greatly influences the percentage of explained variability, the scores of the accuracy measures and, most importantly, the degree of constraint in the distribution estimated. As we extract pseudo-absences from environmental regions further from the optimum established by presence data, the models generated obtain better accuracy scores, and over-prediction increases. When variables other than environmental ones influence the distribution of the species (i.e., non-equilibrium state) and precise information on absences is non-existent, the random selection of pseudo-absences or their selection from environmental localities similar to those of species presence data generates the most constrained predictive distribution maps, because pseudo-absences can be located within environmentally suitable areas. This study shows that if we do not have reliable absence data, the method of pseudo-absence selection strongly conditions the obtained model, generating different model predictions in the gradient between potential and realized distributions.

1. Introduction

Reliable species distribution information on various scales is needed for both biogeographic and conservation purposes. Taking advantage of computing developments such as databases and GIS, many different initiatives aim to compile massive amounts of taxonomic and distribution information (Bisby, 2000). Atlases, museum data and databases can provide information relevant to the development of prediction maps (Dennis and Hardy, 1999; Reutter et al., 2003; Chefaoui et al., 2005; Hortal et al., 2005). Since these heterogeneous data sources do not indicate the locations where the species have not been found after a sufficiently intense collection effort, false absences can decrease the reliability of prediction models (see Anderson, 2003; Loiselle et al., 2003). Group discrimination techniques that use presence–absence data (Guisan and Zimmermann, 2000) seem to predict species distributions more accurately than profile techniques, which only use presence data (Ferrier and Watson, 1997; Manel et al., 1999; Hirzel et al., 2001; Guisan et al., 2002; Brotons et al., 2004; Gu and Swihart, 2004; Segurado and Araújo, 2004). However, group discrimination techniques are appropriate only in the case where absence data indicate the entire area unsuitable for the species are available. Since a quick and feasible method

* *Corresponding author*. Tel.: +34 91 4111328x1278; fax: +34 91 5645078.
E-mail address: mcnj117@mncn.csic.es (J.M. Lobo).
0304-3800/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2007.08.010

to overcome this problem is needed, the following approaches have been suggested: (i) randomly choosing absence points across all of the available territory (for example, Stockwell and Peters, 1999), (ii) selecting random absence points but weighting them in favour of areas known to contain true absences (Zaniewski et al., 2002), and (iii) including absence points identified from a circular buffer area around each presence point (Hirzel et al., 2001). Since all of these methods may produce false absences, even in areas that are environmentally favourable for the species, using a profile technique to calculate a habitat suitability map has been proposed as a way to select weighted absence points, which can subsequently be used with presence data in a logistic regression procedure (Engler et al., 2004). Absences obtained with this method, "pseudo-absences", can be considered an intermediate methodological approach between presence-only and presence-absence distribution models, which are especially useful when accurate absence data are not available.

In this study, two profile techniques were used to select pseudo-absences progressively near to the environmental domain of the presences, while also selecting them at random. Presence–absence models derived from these pseudo-absences and occurrence data from *Graellsia isabelae* (Graells, 1849) (Lepidoptera: Saturniidae), a protected moth endemic to Spain (see Fernández-Vidal, 1992), are compared with the purpose of showing that it is possible to achieve differently forecasted distributions depending on the method and the threshold used to select these pseudo-absences. The variation in these predictions will be subsequently related to the ambivalent capacity of distribution predictive models to represent realized and potential species distributions (sensu Svenning and Skov, 2004).

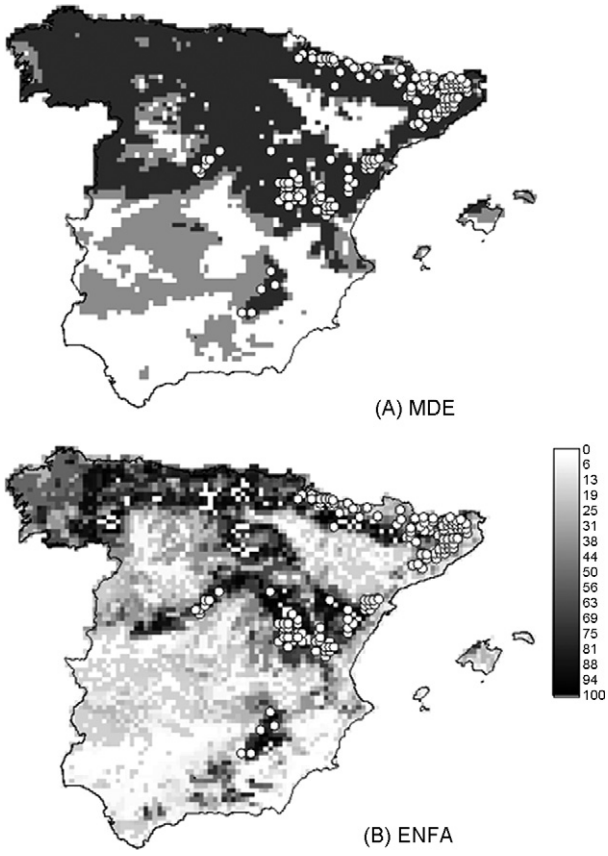


Fig. 1 – Habitat suitability maps representing the potential distribution obtained from presence-only models. (A) Dark grey indicates suitable area obtained from a multi-dimensional envelope model (MDE); light grey indicates potential area added by increasing maximum and minimum scores 10% for each environmental variable (Expanded-MDE). (B) Scale on the right shows different habitat suitability (HS) scores obtained from an ENFA model with the same environmental variables.

2. Methods

2.1. Study area and biological data

The area considered was mainland Spain and the Balearic islands. Since this species has an eastern Iberian distribution, we assume that our study area included most of the suitable habitat area for this species. The studied area comprised 498,150 km^2 divided into 5270 cells of 10 km × 10 km, to which biological and environmental data are referred.

G. isabelae, a sedentary and non-gregarious caterpillar, lives in pine forests and has five developmental stages. From June to August, the larvae feed before metamorphosizing into pupae. Since *G. isabelae* is a conspicuous and well-known species (adults are beautiful and exhibit marked sexual dimorphism), occurrence records can be considered reliable.

Species-presence data were mainly obtained from a distribution atlas (Galante and Verdú, 2000), as well as unpublished data from the Valencia region (Baixeras, 2001; J. Baixeras, personal communication, 2004) and other bibliographic references (Viejo, 1992; García-Barros and Herranz, 2001; López-Sebastián et al., 2002). Because species data came from diverse sources and some references were old (since 1849), all data were checked by comparing their locations with the distribution of pinewoods to eliminate possible outliers. As a result, six presence data points were discarded. A total of 136 presence data points with a spatial resolution of 100 km^2 (UTM cells) were considered (see Fig. 1).

2.2. Predictor variables

The explanatory variables used in the preparation of distribution models (Table 1) come from different sources and have been set up with the aid of IDRISI Kilimanjaro software (Clark Labs, 2003). Topographic variables, elevation and slope were extracted from a global DEM with a 1 km spatial resolution (Clark Labs, 2000). Aspect diversity was calculated by means of the Shannon Index, which estimated the aspect variation in all 1-km pixels composing each 100 km^2 cell. Temperature and precipitation data were provided by the Spanish Instituto Nacional de Meteorología. Aridity was calculated as: $I_a = 1/((P/T)+10) \times 10^2$, where P is the mean annual precipitation and T is the mean annual temperature (see Verdú and Galante, 2002). In addition, four lithological variables were digitized from a lithological map (Instituto Geográfico and Nacional., 1995). The resulting polygon vec-

Table 1 – Explanatory variables used to generate the distribution models

Predictor variables	Minimum–maximum values
Environmental variables	
Mean elevation (m)	0–2722
Aspect diversity	0–16
Slope (°)	0–46
Summer precipitation (July, August and September) (mm)	6.6–472
Annual precipitation (mm)	178–2201
Aridity	0–1.64
Minimum annual temperature (°C)	−3.6 to 14.3
Maximum annual temperature (°C)	9.1–24.9
Area with acidic stony soils (km^2)	0–100
Area with calcareous stony soils (km^2)	0–100
Area with acidic sediments (km^2)	0–100
Area with calcareous sediments (km^2)	0–100
Spatial variables (in UTM coordinates)	
Latitude (Y)	3990000–4860000
Longitude (X)	−20000 to 1060058

Spatial variables were used only with presence–absence GLM models.

tor layers were rasterized at 1 km^2 resolution, and the areas of calcareous deposits, siliceous sediments, stony acidic soils and calcareous soils were subsequently calculated for each cell. These variables were included to incorporate the basic–acidic nature of the soils and their hardness, variables that can be relevant to explain plant species distribution. The third-degree polynomial of the central latitude (Lat) and longitude (Lon) of each grid cell (Trend Surface Analysis; see Legendre and Legendre, 1998) was included after the environmental variables in order to determine if it helped explain anything more about the species distribution (see Lobo et al., 2006). All continuous independent variables were referenced to the same 10 km × 10 km UTM grid square as species data. Predictor environment variables were standardized to 0 mean and 1 standard deviation to eliminate the effect of varying measurement scales. Finally, latitude and longitude were standardized in the same way as the environmental variables.

2.3. *Presence-only models*

We used Multi-Dimensional Niche Envelope (MDE; Busby, 1991; Lobo et al., 2006) and Ecological Niche Factor Analysis (ENFA; Hirzel et al., 2002) to elaborate the presence-only models. These models were generated from the presence data ($n = 136$) and the information from 12 environmental predictor variables (Table 1). For the MDE model, maximum and minimum scores for all environmental variables from presence cells were used to select the suitable grid squares, with environment scores falling within that range. Thus, the generally appropriate environmental conditions for the species were established according to the environmental conditions of the observed presence points. In the Expanded-MDE, this range was expanded by 10% to guarantee that absences selected were environmentally distant from presence localities. MDE and Expanded-MDE were generated in EXCEL spreadsheets, while binary maps were elaborated with IDRISI Kilimanjaro.

ENFA was performed using BIOMAPPER 3.1 software (Hirzel et al., 2004). The ENFA modelling technique (similar to Principle Component Analysis in that it generates orthogonal axes) computes a group of uncorrelated factors with ecological meaning (marginality and specialization), summarizing the main environmental gradients in the region considered. Habitat suitability (HS) is modelled using the selected factors to estimate the ecogeographic degree of similarity between each grid square and the environmental preferences of the species; that is, this method estimates the probability that a given cell belongs to the environmental domain of the presence observations. The resulting habitat suitability map has scores (HS values) that vary from 0 (minimum habitat suitability) to 100 (maximum). The predictor variables were normalized through a Box-Cox transformation (Sokal and Rohlf, 1981), and a "distance geometric-mean" algorithm, which provides a good generalization of the niche (Hirzel and Arlettaz, 2003), was chosen to perform the analyses.

2.4. *Pseudo-absences*

Identifying unsuitable habitats by profile techniques enables us to produce reliable pseudo-absences for presence–absence modelling. Previous results (A. Jiménez-Valverde, J.M. Lobo, J. Hortal, unpublished data) clearly demonstrate that it is necessary to use as much good absence data as possible, especially when dealing with small numbers of presences, to correctly classify the absence zone (see also Thuiller et al., 2004). However, to avoid biases caused by the inclusion of an extremely high number of absences (King and Zeng, 2000; Dixon et al., 2005), we selected 10 times more absences than presences (1360) from each model. To compare the effect of obtaining pseudo-absences with each method, we also randomly selected absences from all regions excepting occurrence cells. Seven groups of pseudo-absences were obtained: one at random, one from MDE, one from Expanded-MDE and four from ENFA. From the ENFA model, the sets were extracted according to four different habitat suitability (HS) thresholds: HS ≤ 10 (ENFA-10), HS ≤ 20 (ENFA-20), HS ≤ 30 (ENFA-30) and HS ≤ 40 (ENFA-40). The upper limit of the selected HS threshold was established as the mean HS score of presences (67) minus its standard deviation (26).

2.5. *Presence–absence models*

The 136 presence data points and each set of 1360 pseudo-absences were subsequently analyzed with the stepwise logistic regression method using Generalized Linear Models (GLM). GLM are an extension of classic linear regression models that allow for analysis of non-linear effects among variables and non-normal distributions of the independent variables (McCullagh and Nelder, 1989). The relationship between the dependent and the explanatory variables (the link function) is logit, and a binomial distribution of the dependent variable was assumed for this analysis.

The species presence–absence data for each of the 10 km × 10 km UTM cells were first compared to linear, quadratic and cubic functions of each environmental vari-

able in order to account for possible curvilinear relationships (Austin, 1980). Next, a model using all environmental variables was built, adding the variables sequentially, in order of their estimated importance (i.e., in a forward–backward stepwise procedure). Lastly, the third-degree polynomial of the central latitude and longitude of each cell was included in the model (TSA) to account for spatial variation due to historic, biotic or environmental factors otherwise not directly considered by this analysis (Legendre and Legendre, 1998). Backwards-stepwise regression, with 9 terms of the equation used as predictor variables, removed insignificant spatial terms. Significant terms ($p < 0.05$) were retained and included in the final environmental model. Including spatial variables after environmental ones partially prevented the model from accurately representing ecological niches but allowed us to increase the explanatory capacity of the model by incorporating unconsidered non-environmental factors. The STATISTICA 6.0 package (StatSoft Inc., 2001) was used for all statistical computations.

2.6. *Validation and cut-off threshold*

The Receiver Operating Characteristic (ROC; Zweig and Campbell, 1993; Schröder, 2004) was used to measure performance of the models. A ROC curve is a plot of sensitivity (ratio of correctly classified positives to the total number of positive cases) versus 1 – specificity (false positive rate) at all possible thresholds of presence–absence classification. The area under the ROC function (AUC), independent of the presence–absence threshold (Fielding, 2002), is widely used as a measure of model prediction accuracy. An AUC value of 0.5, from a possible range of 0–1, indicates that prediction of species presence–absence does not deviate from that of a random assignment, while an AUC score of 1 indicates perfect presence–absence prediction. Prediction maps were also compared by calculating sensitivity and specificity (percentages of correctly predicted presences and absences).

To compare observed and predicted maps, a cut-off point is needed to transform continuous probabilities obtained in GLM models to binary probabilities (i.e., presence–absence). The sensitivity–specificity difference minimizer (Liu et al., 2005; Jiménez-Valverde and Lobo, 2006, 2007) was used to select this threshold due to its generally good performance. These three accuracy measures (AUC, sensitivity and specificity) were computed with the aid of a jackknifing procedure (see Olden et al., 2002; Engler et al., 2004). With a dataset of n observations, the model was recalculated n times, leaving out one observation in turn. Each one of the regression models based on the $n-1$ observations was then applied to the excluded observation, and these models derived predictions for all observations, which were used again to calculate new sensitivity, specificity and AUC jackknife-derived scores.

Since all the resulting models use pseudo-absences, both specificity and AUC scores estimate the degree of accuracy of the absence information used in the model training process. Thus, a high specificity score only implies that most of the data considered as absence data are correctly predicted and does not imply a high performance in the prediction of the unknown true absences.

3. Results

Habitat suitability maps from the profile modelling techniques show remarkable differences (Fig. 1). The suitable area predicted by Expanded-MDE is 31% greater (356,700 km^2) than the area predicted by MDE (244,100 km^2). The predicted area generated by applying the four ENFA threshold-related models decreased with increases in the HS threshold; increasing the HS threshold from 10 to 40 produces a 53% reduction in the suitable area (Fig. 2 and Table 2). The mean ENFA habitat suitability score values for the 136 presence data points was 67.3 ± 25.6 (S.D.) with HS scores oscillating between 5 and 100; 52 presence points had very high suitability scores ($HS > 75$), 45 had high suitability scores ($75 \geq HS > 50$), 33 had low suitability scores ($50 \geq HS > 25$), and 6 had very low suitability scores ($HS \leq 25$).

Five to seven predictor variables were selected in the seven final GLM logistic models ($p \leq 0.05$), highlighting the relevance of some explanatory variables in all models: mean elevation, summer precipitation and aridity (not shown). Spatial variables added after environmental ones only slightly improved the explanatory capacity of the models (around 1% of total variability), except the model in which pseudo-absences were selected at random (around 5% of added variability). Final GLM models in which pseudo-absences were selected by a profile technique accounted for a high percentage of total explained deviance (from 87.6% to 97.6%, see Table 3), while GLM models with pseudo-absences selected at random had a lower explanatory capacity (around 68% of total deviance). In general, models using absence data further away environmentally from the presence data possessed a higher explanatory capacity (Table 3).

All of the models that used pseudo-absences selected by profile techniques had impressive sensitivity, specificity and AUC scores (mean ± 95% confidence interval: 0.9792 ± 0.0123, 0.9843 ± 0.0126 and 0.9952 ± 0.0066, respectively), which were significantly higher than those obtained by selecting pseudo-absences at random (Table 3). Jackknife estimates of the three accuracy measures showed that the model results were highly stable: they differed by less than 2% of the estimates obtained using all the observations (Table 3). As with the explained percent deviance, models that used absence data that where environmentally further away from the presence data also had higher accuracy scores (Table 3).

After selecting the threshold value, continuous GLM probability maps were converted to binomial distributions (Fig. 2). The restrictive character of GLM versus profile techniques is evident; a suitable area generated by GLM models was smaller than that derived from profile-techniques (from 31% to 48% smaller; see Table 2). The GLM model performed using pseudo-absences derived from Expanded-MDE was the one that generated wider forecasted areas. Interestingly, in the case of ENFA-derived GLM models, the estimated species distribution area decreased with gradual increases in the habitat suitability threshold used to discern pseudo-absences. Moreover, when pseudo-absence data were randomly selected, the predicted species distribution area was almost 40% smaller than the most restricted distribution area estimated using pseudo-absences derived from profile techniques (Table 2 and

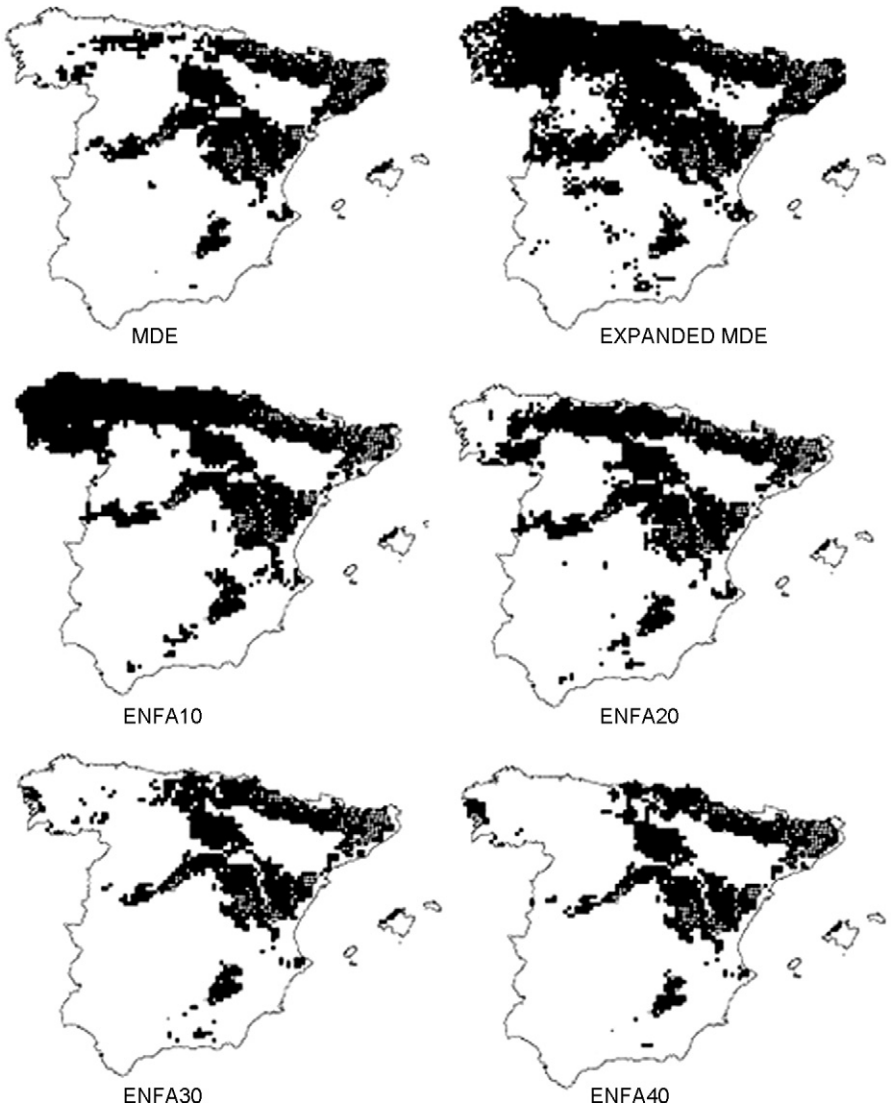


Fig. 2 – Obtained distribution maps from logistic GLM models using pseudo-absences derived from profile techniques, which vary in the threshold applied to select probable absence points (see Section 2).

Fig. 3). This reduction in the predicted area always followed a well-defined geographical pattern; cells in the northwestern corner of the study area, where the species has never been collected, disappeared gradually from the potential distribution area.

4. Discussion

In this study, selecting pseudo-absences appears to be a good strategy to make GLM modelling possible when true

Table 2 – Predicted suitable and unsuitable areas of distribution for *Graellsia isabelae* in Spain according to the two profile techniques (ENFA and MDE) and suitable area predicted by the logistic GLM models from selected pseudo-absences at different thresholds from the profile techniques or randomly (see Section 2)

	MDE	Expanded-MDE	ENFA-10	ENFA-20	ENFA-30	ENFA-40	Random
Unsuitable area estimated by the profile technique	282,900	170,300	144,800	243,700	311,600	347,600	-
Suitable area estimated by the profile technique	244,100	356,700	382,200	283,300	215,400	179,400	-
Suitable area estimated by the GLM model	132,300	245,400	199,200	159,600	121,200	113,200	68,400

Values are expressed in km^2.

Table 3 – Comparison of the GLM models obtained from pseudo-absences generated at random and from two profile techniques (ENFA and MDE) at different threshold scores (see Section 2)

Method of pseudo-absence selection	Deviance	Explained deviance (%)	Sensitivity	Specificity	AUC
MDE					
Environmental	69.61	94.48			
Environmental + TSA	64.43	94.89	0.9779 (0.9619)	0.9816 (0.9711)	0.9990 (0.9824)
Expanded-MDE					
Environmental	42.29	96.65			
Environmental + TSA	42.29	96.65	0.9708 (0.9632)	0.9956 (0.9956)	0.9831 (0.9794)
ENFA-10					
Environmental	35.62	97.17			
Environmental + TSA	30.33	97.59	0.9926 (0.9779)	0.9941 (0.9765)	0.9998 (0.9910)
ENFA-20					
Environmental	55.64	95.59			
Environmental + TSA	45.53	96.39	0.9926 (0.9779)	0.9926 (0.9779)	0.9988 (0.9909)
ENFA-30					
Environmental	115.26	90.87			
Environmental + TSA	109.74	91.31	0.9779 (0.9632)	0.9764 (0.9633)	0.9961 (0.9896)
ENFA-40					
Environmental	171.14	86.45			
Environmental + TSA	156.08	87.64	0.9632 (0.9559)	0.9654 (0.9573)	0.9942 (0.9851)
Random					
Environmental	467.22	63.01			
Environmental + TSA	406.86	67.79	0.8970 (0.8823)	0.8970 (0.8860)	0.9599 (0.9505)

GLM models were built by including environment and environment + spatial (TSA) variables. The percentage of correctly predicted presences (sensitivity) and absences (specificity), as well as the area under the ROC function (AUC) are measurements derived from the confusion matrix to estimate model prediction accuracy. The average scores of these accuracy measures are showed in brackets after accomplishing a Jackknifing procedure in which all the regression models based on the $n-1$ observations were calculated and the model applied to that excluded one.

absence data are not available. Taking into account that an AUC value > 0.90 is qualified as outstanding (Hosmer and Lemeshow, 2000), our validation results for all the pseudo-absence selection approaches are excellent. Engler et al. (2004) show that GLM models using ENFA-weighted pseudo-absences provide significantly better results than those that use randomly chosen pseudo-absences or profile techniques

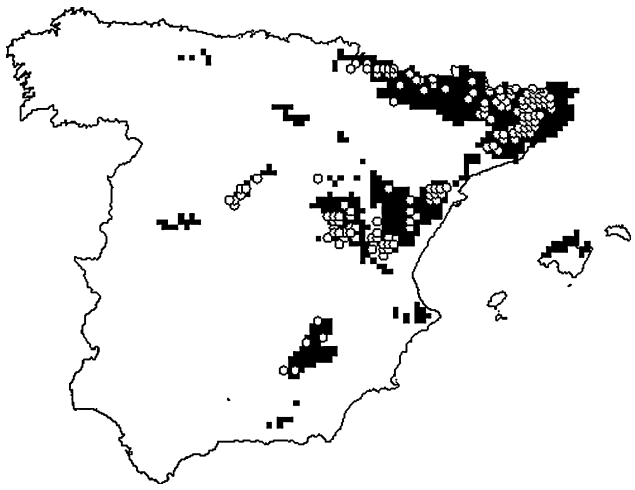


Fig. 3 – Obtained distribution map from logistic GLM model using randomly selected pseudo-absences. Dots represent the observed distribution of *Graellsia isabelae*.

such as ENFA alone, due mainly to the tendency of profile techniques to over-predict species distributions. In agreement with Engler et al. (2004), we find that this strategy provides a way to enhance the quality of GLM-based potential distribution maps. In our case, the GLM model derived from pseudo-absences extracted from cells with an ENFA habitat suitability score equal to or lower than 20 (ENFA-10 and ENFA-20) seems to be the most accurate, although Expanded-MDE pseudo-absence selection also provides rather good validation results. However, the profile method used and the environmental limits defined when selecting pseudo-absences greatly influences the percentage of explained variability, the scores of the accuracy measures and, most importantly, the degree of constraint on the distribution estimated.

In the case of *G. isabelae*, where presence data occur in the eastern territory, GLM spatial predictions exclude the Ebro Valley and other areas of low elevation. However, in the western area, where the species has not been observed, the lack of reliable absence data causes high variability in the predictions; the models tend to expand the suitable area for this species to the northwestern Iberian corner. Presence-only methods always generate wider potential distributional areas than GLM models derived from pseudo-absences. Moreover, those strategies in which pseudo-absences were selected from a smaller area environmentally distant from the optimum established by the presence data (Expanded-MDE and ENFA-10) generate final GLM models that explain a higher percentage of total variability, have higher accuracy scores

and wider distributions. Conversely, the profile techniques that generate wider unsuitable areas, such as MDE, ENFA-30 and ENFA-40, produce functions with lower percentages of explained deviance and poorer accuracy scores, but more restricted predictive distribution maps, similar to the observed distribution. The random selection of pseudo-absences generates the most constrained predictive distribution map because all absence data are included, even those data located within environmentally favourable areas.

Only an appropriate selection of presence and absence locations can guarantee the reliability of distribution model predictions. First, however, we must determine whether we would like to produce a distributional range closer to the potential or closer to the realized distribution. Species distributions should be considered abstractions of a dynamic reality. We can be interested in providing a distributional hypothesis able to reflect all the environmental suitable places in which a species can occur according to a group of environmental variables (the potential distribution). Profile techniques such as MDE and ENFA estimate the distribution of the species considering the environmental information of the localities in which the species has been observed, generating wide suitable distributions; this is because these techniques cannot incorporate the absence information on the climatically suitable localities in which the species does not occur. Many theoretical arguments and empirical studies show that it is possible to find reliable absence data in sites with environmentally favourable conditions (Ricklefs and Schluter, 1993; Hanski, 1998; Pulliam, 1988, 2000). Obviously, including such absence information in predictive modelling techniques should inevitably diminish the obtained range size until a distributional hypothesis nearer to the realized distribution is reached. That happens because some "a priori" favourable environmental localities are considered as absences. Hence, only the use of reliable presence and absence data and discrimination techniques such as GLMs allows the production of a reliable approach to model the "real" distribution of a species; a distribution in which contingent distribution restriction forces as historical factors, biotic interactions or dispersal limitations play an effective role.

The current distribution of *G. isabelae* is reasonably well known, due to its conspicuous nature. Thus, we are inclined to believe that the lack of presence data in suitable habitat areas in western Iberia indicates actual absence, and not a sampling artefact. Profile techniques indicate that the potential distributional range of *G. isabelae* is wider than realized in the western region. Thus, reasons other than environmental characteristics may be the cause of this non-equilibrium state in which species do not occupy all suitable habitats (see Austin, 2002; Guisan and Thuiller, 2005). Under equilibrium conditions, good absence data should always come from locations with unfavourable environments (see, for example, Hirzel and Arlettaz, 2003). Contrarily, in a non-equilibrium scenario, the cells considered environmentally unfavourable and chosen as pseudo-absences can influence the obtained predictive functions and the difference between potential and realized distributions (see Svenning and Skov, 2004). The principle difficulty lies in obtaining predictive distributional models that closely approximate the realized distribution of species under non-equilibrium conditions; obtaining these models causes reductions in goodness-of-fit, similar to those caused by using MDE, ENFA-30 or ENFA-40. This response is due to the fact that both presence and absence data may be possible under similar environmental conditions (see also Collingham et al., 2000). Hence, neither the coefficient of determination, sensitivity, specificity, nor AUC scores are appropriate measures of the performance of models if the objective is to obtain a model under non-equilibrium conditions. Selecting pseudo-absences environmentally distant from the presences unavoidably facilitates the production of models that over-predict presences, as well as the discrimination between presences and absences. The discrimination ability of distribution models must be evaluated according to the pursued purposes. Profile methods must be used if we want to generate a hypothesis on the potential distribution. Using discrimination methods and selecting pseudo-absences by Expanded-MDE and ENFA-10 methods also allows us to obtain models nearer to the potential distribution of the species because pseudo-absences are selected from environments dissimilar to those of species presence data. On the contrary MDE, ENFA-30 and ENFA-40 are better models to represent the approximate range of the realized distribution. Paradoxically, the random selection of pseudo-absences can be a satisfactory alternative procedure to model the realized distribution of the species, provided good absence data are not available, because we include in the modelling process many absences near the environmental domain of the presences. Since there is no single way to build, evaluate and interpret distribution models, it is necessary to carefully consider the available distribution and biological information of each species in an individualized way (Zimmermann and Kienast, 1999; Rushton et al., 2004; Soberón and Peterson, 2005). In conclusion, as the degree of prediction over-estimation varies with the method applied, the success that we can achieve using correlative static models and environmental predictors is determined by two factors: (1) the distribution equilibrium state of the species in the analyzed region and (2) the method used to select pseudo-absences.

How does one construct predictive models without using variables that describe the biotic or historical factors likely to influence the non-equilibrium, present-day distribution of the species? The major challenges of distribution modelling are accounting for the distributions of most species that are likely to be in non-equilibrium states. If it is not possible to assume that environmental factors are the unique determinants of species distributions (Davis et al., 1998; Iverson et al., 2004; Skov and Svenning, 2004; Thomas et al., 2004; Soberón and Peterson, 2005), perhaps including spatial variables along with environmental ones would help account for variability due to non-environmental factors (Legendre and Legendre, 1998; Lobo et al., 2004). In our case, the addition of spatial variables after environmental ones increases the explanatory capacity of the GLM models when pseudo-absences are randomly selected or when the habitat suitability range is augmented to select pseudo-absences. Although the additional deviance explained by environmental + TSA models can be small, it is important to remember that all environmental variables are spatially structured, and that changes in the environmental variables included in the models can cause a better recovery of the spatial variability in the dependent variable. For

example, the inclusion of significant environmental variables obtained from the GLM model built with pseudo-absences from ENFA-10 in the ENFA-20, ENFA-30 and ENFA-40 models does not noticeably reduce the explained deviance (from 95.0%, 88.8% and 82.8% to 93.4%, 84.1% and 80.5%), but the added percentage of variability explained by the spatial variables increases notably (7.3%, 10.0% and 13.3%, respectively). Another promising option is to consider some geographical variables as predictors indirectly related to the failure of a species to colonize the entire suitable territory (see Lobo et al., 2006). That is, if one includes a measure of connectivity between areas, or a "distance cost" (Hortal et al., 2005), one can quantify the dispersal effort necessary to inhabit areas farther from the area with well-known presences, directly integrating dispersal models and environmental data (Iverson et al., 2004). When the biological, historical and physiological information necessary to describe the realized distribution of species is lacking, our predictions should continue to be based on correlative statistical models in which the role of non-environmental processes must be considered. Good models need good data. Thus, the elaboration of reliable simulations on the realized distribution of species unavoidably requires good absence data, as well as the inclusion of non-environmental processes in the model procedure. Our study shows that if we do not have reliable absence data the method of pseudo-absence selection strongly conditions the obtained model, generating different model predictions in the gradient between potential and realized distributions.

Acknowledgements

Special thanks to Alberto Jiménez-Valverde and Joaquín Hortal for their valuable suggestions. This paper has been supported by a Fundación BBVA project (Diseño de una red de reservas para la protección de la Biodiversidad en América del sur austral utilizando modelos predictivos de distribución con taxones hiperdiversos) and a MEC Project (CGL2004-04309).

REFERENCES

Anderson, R.P., 2003. Real vs. artefactual absences in species distributions: tests for *Oryzomys albigularis* (Rodentia: Muridae) in Venezuela. J. Biogeogr. 30, 591–605.

Austin, M.P., 1980. Searching for a model for use in vegetation analysis. Vegetation 42, 11–21.

Austin, M.P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modeling. Ecol. Model. 157, 101–118.

Baixeras, J., 2001. Seguimiento de Poblaciones de *Graellsia isabelae* en Zonas de Actuación del Proyecto Life-Habitats. Universidad de Valencia, Valencia, Spain.

Bisby, F.A., 2000. The quiet revolution: biodiversity informatics and the internet. Science 289, 2309–2312.

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

Busby, J.R., 1991. BIOCLIM–a bioclimate analysis and prediction system. In: Margules, C.R., Austin, M.P. (Eds.), Nature Conservation: Cost Effective Biological Surveys and Data Analysis. CSIRO, Melbourne, Australia, pp. 64–68.

Chefaoui, R.M., Hortal, J., Lobo, J.M., 2005. Potential distribution modelling, niche characterization and conservation status assessment using GIS tools: a case study of Iberian *Copris* species. Biol. Conserv. 122, 327–338.

Clark Labs, 2000. Global Change Data Archive, vol. 3: 1 km Global Elevation Model. Clark University.

Clark Labs, 2003. Idrisi Kilimanjaro. GIS Software Package. Clark Labs, Worcester, MA.

Collingham, Y.C., Wadsworth, R.A., Willis, S.G., Huntley, B., Hulme, P.E., 2000. Predicting the spatial distribution of alien riparian species: issues of spatial scale and extent. J. Appl. Ecol. 37, 13–27.

Davis, A.J., Jenkinson, L.S., Lawton, J.H., Shorrocks, B., Wood, S., 1998. Making mistakes when predicting shifts in species range in response to global warming. Nature 391, 783–786.

Dennis, R.L.H., Hardy, P.B., 1999. Targeting squares for survey: predicting species richness and incidence of species for a butterfly atlas. Global Ecol. Biogeogr. 8, 443–454.

Dixon, P.M., Ellison, A.M., Gotelli, J., 2005. Improving the precision of estimates of the frequency or rare events. Ecology 86, 1114–1123.

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.

Fernández-Vidal, E.H., 1992. Comentarios acerca de la distribución geográfica francesa y notas taxonómicas sobre *Graellsia isabelae* (Graells, 1849). Shilap 77, 29–49.

Ferrier, S., Watson, G., 1997. An evaluation of the effectiveness of environmental surrogates and modelling techniques in predicting the distribution of biological diversity. NSW National Parks and Wildlife Service Department of Environment, Sport and Territories, Report to Environment Australia, Canberra, Australia.

Fielding, A.H., 2002. What are the appropriate characteristics of an accuracy measure? In: Scott, J.M., Heglund, P.J., Haufler, J.B., Morrison, M., Raphael, M.G., Wall, W.B., Samson, F. (Eds.), Predicting Species Occurrences. Issues of Accuracy and Scale. Island Press, Washington, USA, pp. 271–280.

Galante, E., Verdú, J.R., 2000. Los Artrópodos de la "Directiva Hábitat" en España. Ministerio de Medio Ambiente, Madrid, Spain, 247 pp.

García-Barros, E., Herranz, J., 2001. Nuevas localidades de *Proserpinus proserpina* (Pallas, 1772) y *Graellsia isabelae* (Graells, 1849) del centro peninsular. Shilap 114, 183–184.

Gu, W., Swihart, R.K., 2004. Absent or undetected? Effects of non-detection of species occurrence on wildlife-habitat models. Biol. Conserv. 116, 195–203.

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

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

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

Hanski, I., 1998. Metapopulation dynamics. Nature 396, 41–49.

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

Hirzel, A.H., Hausser, J., Chessel, D., Perrin, N., 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecology 7, 2027–2036.

Hirzel, A.H., Hausser, J., Perrin, N., 2004. Biomapper 3.1. Division of Conservation Biology, University of Bern, Bern, Switzerland. http://www.unil.ch/biomapper.
Hortal, J., Nieto, M., Rodríguez, J., Lobo, J.M., 2005. Evaluating the roles of connectivity and environment on faunal turnover: patterns in recent and fossil Iberian mammals. In: Elewa, A.M.T. (Ed.), Migration in Organisms-Climate, Geography, Ecology. Springer–Verlag, Berlin, Germany, pp. 301–328.
Hosmer Jr., D.W., Lemeshow, S., 2000. Applied Logistic Regression. John Wiley and Sons (ed.), New York, USA, 373 pp.
Instituto Geográfico Nacional. 1995. Atlas nacional de España, vol. 1–2. Centro Nacional de Información, Madrid, Spain.
Iverson, L.R., Schwartz, M.W., Prasad, A.M., 2004. How fast and far might tree species migrate in the eastern United States due to climate change? Global Ecol. Biogeogr. 13, 209–219.
Jiménez-Valverde, A., Lobo, J.M., 2006. The ghost of unbalanced species distribution data in geographic model predictions. Div. Dist. 12, 521–524.
Jiménez-Valverde, A., Lobo, J.M., 2007. Threshold criteria for conversion of probability of species presence to either–or presence–absence. Acta Oecol. 31, 361–369.
King, G., Zeng, L., 2000. Logistic regression in rare events data. Political Anal. 9, 2.
Legendre, P., Legendre, L., 1998. Numerical Ecology. Elsevier, Amsterdam, Holland, 853 pp.
Liu, C., Berry, P.M., Dawson, T.P., Pearson, R.G., 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28, 385–393.
Lobo, J.M., Jay-Robert, P., Lumaret, J.P., 2004. Modelling the species richness for French Aphodiidae (Coleoptera, Scarabaeoidea). Ecography 27, 145–156.
Lobo, J.M., Verdu, J.R., Numa, C., 2006. Environmental and geographical factors affecting the Iberian distribution of flightless *Jekelius* species (Coleoptera: Geotrupidae). Div. Dist. 12, 179–188.
Loiselle, B.A., Howell, C.A., Graham, C.H., Goerck, J.M., Brooks, T., Smith, K.G., Williams, P.H., 2003. Avoiding pitfalls of using species distribution models in conservation planning. Conserv. Biol. 17, 1591–1600.
López-Sebastián, E., López, J.C., Juan, M.J., Selfa, J., 2002. Primeras citas de mariposa isabelina en la Comunidad Valenciana. Quercus 193, 10–13.
Manel, S., Dias, J.M., Ormerod, S.J., 1999. Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. Ecol. Model. 120, 337–347.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. Chapman and Hall, London, England, 511 pp.
Olden, J.D., Jackson, D.A., Peres-Neto, P.R., 2002. Predictive models of fish species distributions: a note on proper validation and chance predictions. Trans. Am. Fish. Soc. 131, 329–336.
Pulliam, H.R., 1988. Sources, sinks and population regulation. Am. Nat. 132, 652–661.
Pulliam, H.R., 2000. On the relationship between niche and distribution. Ecol. Lett. 3, 349–361.
Reutter, B., Helfer, V., Hirzel, A.H., Vogel, P., 2003. Modelling habitat-suitability using museum collections: an example with three sympatric *Apodemus* species from the Alps. J. Biogeogr. 30, 581–590.
Ricklefs, R.E., Schluter, D., 1993. Species Diversity in Ecological Communities. Historical and Geographical Perspectives. University Chicago Press, Chicago, USA, 416 pp.
Rushton, S.P., Ormerod, S.J., Kerby, G., 2004. New paradigms for modelling species distributions? J. Appl. Ecol. 41, 193–200.
Schröder, B., 2004. ROC Plotting and AUC Calculation Transferability Test. Potsdam University.
Segurado, P., Araújo, M.B., 2004. An evaluation of methods for modelling species distributions. J. Biogeogr. 31, 1555–1568.
Skov, F., Svenning, J.C., 2004. Potential impact of climatic change on the distribution of forest herbs in Europe. Ecography 27, 366–380.
Soberón, J., Peterson, T.A., 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. Biodivers. Inform. 2, 1–10.
Sokal, R.R., Rohlf, F.J., 1981. Biometry. Freeman, New York, USA, 859 pp.
StatSoft, Inc., 2001. STATISTICA (data analysis software system), version 6.
Stockwell, D., Peters, D., 1999. The GARP modelling system: problems and solutions to automated spatial prediction. Int. J. Geogr. Inf. Sci. 2, 143–158.
Svenning, J.C., Skov, F., 2004. Limited filling of the potential range in European tree species. Ecol. Lett. 7, 565–573.
Thomas, C.D., Cameron, A., Green, R.E., Bakkenes, M., Beaumont, L.J., Collingham, Y.C., Erasmus, B.F.N., Ferreira de Siqueira, M., Grainger, A., Hannah, L., Hughes, L., Huntley, B.S., van Jaarsveld, A., Midgley, G.F., Miles, L., Ortega-Huerta, M.A., Peterson, A.T., Phillips, O.L., Williams, S.E., 2004. Extinction risk from climate change. Nature 427, 145–148.
Thuiller, W., Brotons, L., Araújo, M.B., Lavorel, S., 2004. Effects of restricting environmental range of data to project current and future species distributions. Ecography 27, 165–172.
Verdú, J.R., Galante, E., 2002. Climatic stress, food availability and human activity as determinants of endemism patterns in the Mediterranean region: the case of dung beetles (Coleoptera, Scarabeoidea) in the Iberian Peninsula. Div. Dist. 8, 259–274.
Viejo, J.L., 1992. Biografía de un naturalista y biología del lepidóptero por él descrito. Graells y la Graellsia. Quercus 74, 22–30.
Zaniewski, A.E., Lehman, A., Overton, J., 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.