Effects of species' traits and data characteristics on distribution models of threatened invertebrates

R. M. Chefaoui, J. M. Lobo & J. Hortal

Chefaoui, R. M., Lobo, J. M. & Hortal, J., 2011. Effects of species' traits and data characteristics on distribution models of threatened invertebrates. *Animal Biodiversity and Conservation*, 34.2: 229–247.

Abstract

Effects of species' traits and data characteristics on distribution models of threatened invertebrates.— The lack of information about the distribution of threatened species inhibits the development of strategies for their conservation. This is a particularly important problem when considering invertebrates. Here we evaluate the effects of species' traits and data characteristics on the accuracy of species distribution models (SDM) of 20 threatened Iberian invertebrates. We found that the accuracy of the predictions was mostly affected by the characteristics of the data. Species whose distributions were most accurately modelled were those with a greater sample size or smaller relative occurrence area (ROA). Species in habitats that were difficult to detect using GIS data, such as riparian species, tended to be more difficult to predict.

Key words: Ecological traits, Geographical distribution range, Iberian peninsula, Predictive accuracy, Sample size, Species distribution modelling.

Resumen

Efectos de las características ecológicas y de los datos sobre los modelos de distribución de invertebrados protegidos.— La escasez de información sobre la distribución de las especies amenazadas impide el desarrollo de estrategias para su conservación, un problema particularmente importante en el caso de los invertebrados. En este trabajo se evalúan los efectos que las características ecológicas y de los datos ejercen sobre la precisión de los modelos de distribución de 20 especies ibéricas de invertebrados amenazados. Se encontró que la precisión en los modelos predictivos se ve afectada mayoritariamente por las características de los datos. Las especies que obtienen modelos de distribución más precisos son aquellas con mayor tamaño de muestra o menor área de ocurrencia relativa (ROA). Además, las especies relacionadas con hábitats difíciles de detectar mediante SIG, como las especies riparias, tienden a ser más difíciles de predecir.

Palabras clave: Características ecológicas, Modelos de distribución de especies, Península ibérica, Precisión del modelo, Rango de distribución geográfica, Tamaño de muestra.

(Received: 16 XII 10; Conditional acceptance: 23 II 11; Final acceptance: 29 III 11)

Rosa M. Chefaoui, Jorge M. Lobo & Joaquín Hortal, Depto. de Biodiversidad y Biología Evolutiva, Museo Nacional de Ciencias Naturales–CSIC, c/ José Gutiérrez Abascal 2, 28006 Madrid, España (Spain).– J. Hortal, Depto. de Ecologia, Inst. de Ciências Biológicas, Univ. Federal de Goiás, 74001–970 Goiania, GO, Brazil.

Corresponding author: Rosa M. Chefaoui. E–mail: rosa.chef@gmail.com

ISSN: 1578–665X © 2011 Museu de Ciències Naturals de Barcelona

Introduction

Including rare and threatened species in the prioritisation of protected areas is particularly challenging because of the low spatial congruence (*i.e.*, coincidence in space) between species ranges (Grenyer et al., 2006) and the difficulties associated with mapping their distributions. Data scarcity is often ameliorated with the help of GIS–based models and analytical techniques. Species distribution modelling is nowadays a well–established set of research techniques (see Franklin, 2009 and references therein), and many studies use data from museum collections and literature to model the distributions of species (*e.g.*, Reutter et al., 2003; Brotons et al., 2004; Elith & Leathwick, 2007). Species distribution models (SDMs) are especially important when working with hyperdiverse invertebrates, where the difficulty of conducting extensive surveys makes biodiversity databases based on data from museums and atlases a necessary alternative to obtain presence records for mapping distributions (*e.g.,* Chefaoui et al., 2005; Lobo et al., 2006, 2010; Chefaoui & Lobo, 2007). Unfortunately, the quality of data on many species in biodiversity databases is usually compromised by sampling bias and/or deficient survey effort (Hortal et al., 2007), a problem that is particularly important for many invertebrate groups (Lobo et al., 2007; Hortal et al., 2008). Under these circumstances, systematic conservation planning for invertebrate taxa generally entails modelling species with diverse characteristics and ecological requirements using poor quality data, often with no time for detailed 'species–by–species' assessment by experts (see Cabeza et al., 2010). Using automated SDM protocols to predict the distribution of invertebrates from presence–only data is hampered by: (i) the use of heterogeneous biological data sources generally without any survey effort measure; (ii) the environmental and spatially biased character of this information; (iii) the lack of accurate absence data; (iv) the difficulty of identifying the best predictor variables for each species; and (v) the difficulty of finding a reliable accuracy measure of SDM performance that allows model success to be compared between different species (see discussion in Lobo et al., 2008, 2010; Jiménez–Valverde et al., 2008; Rocchini et al., 2011).

In an attempt to understand the limitations and possibilities of SDM techniques, many studies have addressed how the characteristics of the data and different ecological or geographical species' traits affect model accuracy. An increase in model accuracy has been related to greater sample sizes (Stockwell & Peterson, 2002; McPherson et al., 2004; Wisz et al., 2008; Mateo et al., 2010), and also to species with more specialized requirements (Brotons et al., 2004; Seoane et al., 2005), less mobility (Pöyry et al., 2008), more between–year population constancy (Carrascal et al., 2006), longer life spans in plants (Hanspach et al., 2010; Syphard & Franklin, 2010), specific types of response to fire disturbance in plants (Syphard & Franklin, 2010), and smaller geographic ranges (Stockwell & Peterson, 2002; Segurado & Araújo, 2004; Hernández et al., 2006). Nevertheless, relationships between model performance and species traits are strongly dependent on the modelling technique, and also on the characteristics of the data itself. These characteristics refer to sample size and the proportion of the occupied area over the total area of the territory under study (the relative occurrence area or ROA; Lobo, 2008; Lobo et al., 2008; Santika, 2010). Thus, a better understanding of how species' traits and data characteristics influence the results of different modelling methods could help refine the use of SDMs.

The main aim of this study was to determine how ecological traits and data characteristics influence the predictive performance of SDMs in the case of threatened insects and other invertebrate species. More precisely, we examined the relationship between three general measures of model accuracy (AUC, sensitivity and specificity), and (i) two characteristics of the data used, namely sample size and ROA, and (ii) several ecological traits, including niche specialization (marginality), the total extent of the distribution range (herein TER), dispersal ability, trophic group, habitat type and habitat detectability. To do this, we applied three SDM procedures (Generalized Linear Models, GLMs; Generalized Additive Models, GAMs; and Neural Network Models, NNETs) to model the distribution in the Iberian Peninsula of 20 threatened invertebrate species that have different ecological traits and data characteristics. We used presence data from museum collections and atlases, and pseudo–absences (Zaniewski et al., 2002; Chefaoui & Lobo, 2008). We later evaluated the influence of the data characteristics and species traits on model performance measures using non–parametric statistical tests.

Methods

Study area

The study area was the Ibero–Balearic region (western Mediterranean), which comprises 587,663 km^2. Data on species occurrences were gathered from atlases and bibliographic sources, using 10 km–resolution (*i.e.,* 100 km^2) UTM cells due to the lack of geographical precision of most data sources. Environmental data were also referenced to the same resolution. The study area was therefore divided into 6,150 cells of 100 km^2, which constitute the units of analysis.

Biological data

We arbitrarily selected 20 species of threatened and/or protected (Bern Convention and Habitat Directive) invertebrates found in Spain (17 Arthropoda and 3 Mollusca; table 1). Based on different catalogues (see http://www.mma.es; Galante & Verdú, 2000; Verdú & Galante, 2006*)*, we selected species that fulfilled two requirements: (i) their presence had been recorded in a minimum of ten 10 x 10 km grid cells; and (ii) they had different biological and ecological traits and data characteristics. Occurrence data

Table 1. Data and species characteristics that may influence model accuracy: N. Sample size; ROA. Relative occurrence area; M. Marginality; TER. Total extent of the distribution range, in three categories from more restricted to wider distribution (C–I. Iberian [a] and Ibero–Maghrebian [b]; C–II. European; C–III. Euroasiatic); HT. Habitat types (T–I. Woods [a] and mountainous habitats [b]; T–II. Grasslands [a], varied habitats [b] and rocky slopes [c]; T–III. Riparian [a] and humid habitats [b]); HD. Habitat detectability (H. High; L. Low); TLC. Trophic level categories (Pl. Polyphagous; Cr. Carnivorous; Om. Omnivorous; Ol. Oligophagous; Ph. Phytophagous; NPh. Non–phytophagous); FC. Flight capacity.

Tabla 1. Características de los datos y de las especies que pueden influir en la precisión de los modelos: N. Tamaño de muestra; ROA. Área de presencia relativa; M. Marginalidad; TER. Extensión total del rango de distribución, en tres categorías desde la distribución más restringida a más amplia (C–I. Ibérica [a] e ibero–magrebí [b]; C–II. Europea; C–III. Euroasiática); HT. Tipos de hábitats (T–I. Bosques [a] y hábitats montañosos [b]; T–II. Praderas [a]; hábitats mixtos [b] y pendientes rocosas [c]; T–III. Hábitats húmedos [b] y riparios [a]); HD. Detectabilidad del hábitat (H. Alta; L. Baja); TLC. Categorías de nivel trófico (Pl. Polífago; Cr. Carnívoro; Om. Omnívoro; Ol. Oligófago; Ph. Fitófagas; NPh. No fitófagas); FC. Capacidad de vuelo.

	Data characteristics		Species characteristics					
Species	N	ROA	M	TER	HT	HD	TLC	FC
Cerambyx cerdo	152	0.796	0.768	C–III	T–I(a)	H	Pl(Ph)	Yes
Coenagrion mercuriale	87	0.629	0.455	C–I(b)	T–III(a)	L	Cr(NPh)	Yes
Cupido lorquinii	87	0.267	1.201	C–I(b)	T–II(a)	H	Om(NPh)	Yes
Elona quimperiana	41	0.141	2.869	C–I(b)	T–II(b)	L	Om(NPh)	No
Eriogaster catax	12	0.067	2.538	C–III	T–I(a)	H	Pl(Ph)	Yes
Euphydryas aurinia	749	0.851	1.154	C–III	T–I(a)	H	Ol(Ph)	Yes
Geomalacus maculosus	37	0.114	2.397	C–II	T–III(b)	L	Pl(Ph)	No
Graellsia isabelae	138	0.212	2.240	C–I(a)	T–I(a)	H	Ol(Ph)	Yes
Lucanus cervus	456	0.625	1.915	C–III	T–I(a)	H	Pl(Ph)	Yes
Macromia splendens	10	0.436	1.797	C–I(b)	T–III(a)	L	Cr(NPh)	Yes
Macrothele calpeiana	92	0.076	1.624	C–I(a)	T–II(b)	L	Cr(NPh)	No
Maculinea alcon	49	0.212	2.528	C–II	T–II(a)	H	Om(NPh)	Yes
Maculinea arion	166	0.310	3.397	C–III	T–II(a)	H	Om(NPh)	Yes
Maculinea nausithous	17	0.041	4.584	C–II	T–II(a)	H	Om(NPh)	Yes
Oxygastra curtisi	21	0.612	1.971	C–II	T–III(a)	L	Om(NPh)	Yes
Parnassius apollo	314	0.459	3.600	C–III	T–I(b)	L	Pl(Ph)	Yes
Parnassius mnemosyne	42	0.017	5.897	C–III	T–I(b)	H	Ol(Ph)	Yes
Rosalia alpina	47	0.132	3.656	C–III	T–I(a)	H	Pl(Ph)	Yes
Vertigo moulinsiana	20	0.064	1.261	C–II	T–III(b)	L	Cr(NPh)	No
Zerynthia rumina	1,107	0.927	0.376	C–I (b)	T–II(c)	L	Ol(Ph)	Yes

were obtained from the abovementioned catalogues, and from a diverse array of bibliographic sources (Soria et al., 1986; Castillejo, 1990; Rosas et al., 1992; Viejo Montesinos, 1992; Grosso–Silva, 1999; Grupo de Trabajo sobre Lucanidae Ibéricos, 2000; García–Barros & Herranz, 2001; Pérez–Bote et al., 2001; Raimundo et al., 2001; López–Sebastián et al., 2002; Martínez–Orti, 2004).

Because accurate absence data were not available, we used pseudo–absences to perform model training and validation. We identified environmental pseudo–absences located outside the climatic domain, defined by the available presences (see Lobo et al., 2010). To establish such a climatic domain we used a profile technique; a multidimensional envelope containing all presence data in a multivariate environmental space (Busby, 1991; Lobo et al., 2006) was calculated for each species using the maximum and minimum scores for each topographic, climatic and lithological variable mentioned in table 2.

Table 2. Predictor variables used to generate distribution models for the species. The appropriate variables for each species were previously selected by individual logistic regression analyses (see text).

Tabla 2. Variables predictivas usadas para generar los modelos de distribución de especies. Las variables apropiadas para cada especie fueron previamente seleccionadas individualmente mediante análisis de regresión logística (ver texto).

Predictor variables	Minimum–Maximum values
Topographic variables	
Maximum elevation (m)	1–3,399
Mean elevation (m)	1–2,721
Minimum elevation (m)	0–2,521
Elevation range (m)	0–2,291
Climatic variables	
Winter precipitation (Jan., Feb., March) (mm)	491–9,579
Spring precipitation (April, May, June) (mm)	463–6,236
Summer precipitation (July, August, Sept.) (mm)	66–4,724
Autumn precipitation (Oct., Nov., Dec.) (mm)	607–6,140
Temperature range (ºC)	11–32
Maximum Winter Temperature (ºC)	1–18
Mean Winter Temperature (ºC)	–4–13
Minimum Winter Temperature (ºC)	–8–10
Maximum Spring Temperature (ºC)	6–23
Mean Spring Temperature (ºC)	0–17
Minimum Spring Temperature (ºC)	–5–12
Maximum Summer Temperature (ºC)	19–35
Mean Summer Temperature (ºC)	10–26
Minimum Summer Temperature (ºC)	2–20
Maximum Autumn Temperature (ºC)	9–25
Mean Autumn Temperature (ºC)	2–21
Minimum Autumn Temperature (ºC)	–3–15
Aridity	0–1.64
Lithological variables	
Area of acid soil (km^2)	0–100
Area of calcareous soil (km^2)	0–100
Area of acid sediments (km^2)	0–100
Area of calcareous sediments (km^2)	0–100
Spatial variables	
Latitude (Y)	390000–4860000
Longitude (X)	–20000–1060058

We then created environmental pseudo–absences equalling ten times the number of presences (prevalence = 0.1). This way we included as many absences as possible in the training data, while avoiding biases caused by the inclusion of an extremely high number of absences (*e.g.*, prevalences below 0.01) (King & Zeng, 2000; Dixon et al., 2005; Jiménez–Valverde & Lobo, 2006; Jiménez–Valverde et al., 2009).

As pseudo–absences were randomly selected from the area outside each envelope, they *a priori* excluded the possibility that some environmentally suitable localities where the species does not occur (either because it has not been able to colonize there or because it recently became extinct) would be counted as absences. Geographical predictions thus obtained would tend to approximate the potential distributions of the studied species rather than their realized distributions, as would occur if we were using random pseudo–absences (see Chefaoui & Lobo, 2008; Jiménez–Valverde et al., 2008; Lobo et al., 2010; see also Beaumont et al., 2009). In addition, by choosing pseudo–absences far from the environmental domain occupied by the presence data the discriminant ability of the environmental predictors would be maximized, because no pseudo–absences would be located in environmental domains similar to those occupied by the species presences. Using this kind of pseudo–absences inevitably inflates the AUC values obtained to measure model accuracy (Chefaoui & Lobo, 2008) because the localities with unsuitable environmental conditions are almost always well predicted. In this study we assumed that such inflation of AUC values was similar for all species, independently of their degree of equilibrium with the environment or how narrow their environmental tolerances were. In our case, low AUC values would highlight the inability of some predictor variables to discriminate suitable from unsuitable conditions. It should be noted here that we were interested in assessing the effects of species' traits and data characteristics on the accuracy of models aimed at representing the potential distribution of species. Therefore, we used different techniques and/or predictors to identify any patterns that consistently emerged despite the slightly different assumptions and flexibility in functions of each modelling strategy, rather than to assess the performance of different SDM techniques.

Predictor variables

Due to the heterogeneity in the ecological roles, life histories and adaptations of the invertebrates studied, we selected the best set of predictor variables (table 2) for each species from a range of topographic, climatic, lithological and spatial variables by means of a selection procedure (see below). We extracted topographic variables (maximum, mean and minimum elevation) from a global digital elevation model with 1–km spatial resolution (Clark Labs, 2000); elevation range was calculated as the difference between maximum and minimum elevation in each cell. GIS–layers accounting for minimum, mean and maximum temperature and precipitation for each season at 1–km resolution based on observations from weather stations were provided by the Spanish State Agency of Meteorology (http://www.aemet.es/). We calculated aridity as

$$Ia = 1/ (P/T + 10) \times 10^2$$

where P is the mean annual precipitation and T the mean annual temperature (see Verdú & Galante, 2002). We digitized four lithology variables from a lithology map (Instituto Geográfico Nacional, 1995), and subsequently calculated the area of calcareous deposits, siliceous sediments, stony acidic soils and calcareous soils on each 100 km^2 UTM cell. Finally, we extracted two spatial variables per cell: latitude (Lat) and longitude (Lon) of the centroid of each cell, and generated a trend surface with the third order polynomial of longitude and latitude (*i.e.,* Trend Surface Analysis). The inclusion of these spatial variables after environmental predictors can help to represent the effect of unaccounted–for predictors and/or other factors known to generate spatial patterns in species distributions (see Legendre & Legendre, 1998).

All predictor variables were extracted and handled using IDRISI Kilimanjaro GIS software (Clark Labs, 2003) to the 10 x 10 km UTM grid cells. All these variables (including latitude and longitude) were standardized to zero mean and one standard deviation to eliminate the effect of varying measurement scales.

Species distribution models

Species presence data, pseudo–absences and the selected predictor variables for each species were used to generate predictive functions for species distributions, by means of three different and widely used SDM techniques: Generalized Linear Models (GLMs), Generalized Additive Models (GAMs) and Neural Network Models (NNETs). GLMs (McCullagh & Nelder, 1989) were elaborated assuming a logistic relationship between the dependent and the explanatory variables (*i.e.,* link function), and a binomial error distribution of the dependent variable. To select the best explanatory variables for each species, presence–absence data were regressed against each one of the explanatory variables, using Statistica software (Statsoft, 2001). We evaluated the linear, quadratic and cubic functions for each variable, in order to account for possible curvilinear relationships (Austin, 1980). In addition, we chose the most appropriate spatial variables for each species after a backward–stepwise elimination of non–significant terms from the third–degree polynomial of latitude and longitude. The selected explanatory variables were used in the GAM models using penalized regression splines (Wood & Augustin, 2002) and in the NNET models fitting a single–hidden–layer neural network, with skip–layer connections (Ripley, 1996). All Species Distribution Models were fitted in R (R Development Core Team, 2008).

Measures of model performance

Given that the sample size for some species was small, we opted not to split it into representative training and evaluation datasets. We thus implemented a 'leave–one–out' jack–knife procedure (Olden et al., 2002) to validate models for all species. For this procedure, each observation is excluded and the model is parameterized using the remaining $n - 1$ observations to obtain a predicted probability score for the excluded observation; this procedure yields relatively unbiased estimates of model performance (Olden et al., 2002).

Table 3. Accuracy measures and resulting area size for each studied species and modelling technique used. All areas are measured as the number of grid cells (of 100 km^2 each): GAM. Generalized additive models; GLM. Generalized linear models; NNET. Neural network models; SD. Standard deviation.

Tabla 3. Medidas de precisión y tamaño de área resultante para cada especie estudiada y técnica predictiva utilizada. Todas las áreas se miden por el número de celdas (de 100 km^2 cada una): GAM. Modelos aditivos generalizados; GLM. Modelos generalizados lineales; NNET. Modelos de redes neuronales; SD. Desviación estándar.

Species	AUC			Specificity			Sensitivity			Area (in grid cells)		
	GAM	GLM	NNET	GAM	GLM	NNET	GAM	GLM	NNET	GAM	GLM	NNET
Cerambyx cerdo	0.9402	0.9557	0.8353	0.8620	0.8746	0.7565	0.8618	0.8750	0.7565	4,328	4,458	2,507
Coenagrion mercuriale	0.9196	0.9414	0.8020	0.8275	0.8735	0.7298	0.8275	0.8735	0.7356	3,857	4,165	1943
Cupido lorquinii	0.9730	0.8936	0.9788	0.9563	0.9827	0.9310	0.9540	0.8045	0.9310	1,021	859	822
Elona quimperiana	0.9866	0.9692	0.9885	0.9512	0.9439	0.9463	0.9512	0.9512	0.9512	594	551	398
Eriogaster catax	0.9430	0.9062	0.9840	0.9083	0.9583	0.9166	0.9166	0.8333	0.9166	4,845	4,866	5,234
Euphydryas aurinia	0.9896	0.9909	0.9835	0.9524	0.9571	0.9508	0.9519	0.9572	0.9506	4,127	4,159	4,019
Geomalacus maculosus	0.9554	0.9654	0.9385	0.8918	0.9189	0.8918	0.8918	0.9189	0.8918	784	747	676
Graellsia isabelae	0.9934	0.9927	0.9709	0.9594	0.9507	0.9275	0.9565	0.9492	0.9275	1,021	962	998
Lucanus cervus	0.9926	0.9924	0.9818	0.9700	0.9649	0.9547	0.9692	0.9649	0.9539	2,983	3,243	3,056
Macromia splendens	0.8830	0.8030	0.8570	0.7600	0.7000	0.7900	0.8000	0.7000	0.8000	1,361	848	266
Macrothele calpeiana	0.9933	0.9321	0.9626	0.9565	0.9782	0.9130	0.9565	0.8804	0.9130	454	369	318
Maculinea alcon	0.9729	0.9668	0.9536	0.9346	0.9265	0.9183	0.9387	0.9183	0.9183	1,182	938	1,024
Maculinea arion	0.9927	0.9912	0.9785	0.9596	0.9698	0.9337	0.9578	0.9698	0.9337	1,205	1,141	1,074
Maculinea nausithous	0.9861	0.9081	0.9892	0.9411	0.9882	0.9411	0.9411	0.8235	0.9411	214	172	285
Oxygastra curtisi	0.8749	0.8544	0.8920	0.8904	0.9190	0.8095	0.8095	0.7619	0.8095	677	493	557
Parnassius apollo	0.9932	0.9917	0.9869	0.9722	0.9746	0.9726	0.9713	0.9745	0.9713	1,850	1,622	1,884
Parnassius mnemosyne	0.9953	0.9462	0.9977	0.9761	0.9880	0.9976	0.9761	0.9047	1.0000	230	147	125

Table 3. (Cont.)

Species	AUC			Specificity			Sensitivity			Area (in grid cells)		
	GAM	GLM	NNET	GAM	GLM	NNET	GAM	GLM	NNET	GAM	GLM	NNET
Rosalia alpina												
	0.9881	0.9426	0.9838	0.9446	0.9148	0.9361	0.9361	0.9148	0.9361	545	485	341
Vertigo moulinsiana												
	0.9745	0.9288	0.8575	0.9350	0.8550	0.8050	0.9500	0.8500	0.8000	405	444	1,326
Zerynthia rumina												
	0.9860	0.9888	0.9660	0.9428	0.9551	0.9265	0.9430	0.9548	0.9268	5,422	5,596	5,455
Mean ± SD												
	0.9666	0.9430	0.9444	0.9245	0.9296	0.8974	0.9230	0.8890	0.8982	1,855.2	1,813.2	1,615.4
	± 0.03	± 0.04	± 0.05	± 0.05	± 0.06	± 0.07	± 0.05	± 0.07	± 0.07	±1,716.3	±1,823.7	±1,638.4

After repeating this procedure *n* times (one per observation), we used these new jack–knife probabilities to calculate three measures of model performance: (i) the area under the ROC curve (AUC) (Zweig & Campbell, 1993; Schröder, 2004), (ii) sensitivity (proportion of correctly predicted presences) and (iii) specificity (proportion of correctly predicted absences). Sensitivity and specificity were calculated fixing the threshold probability according to the prevalence of the data (0.1; see Jiménez–Valverde & Lobo, 2006). To transform the continuous probabilities obtained in SDMs to binary results (*i.e.,* presence–absence) we used the sensitivity–specificity sum maximizer criteria (Jiménez–Valverde & Lobo, 2006, 2007). All measures ranged from 0 (poor quality model) to 1 (excellent prediction).

Data characteristics

We evaluated the influence of two characteristics on model performance: sample size (N) and the Relative Occurrence Area (ROA). ROA is the ratio between the area of the distribution range of the species within the studied region, and the total area of such region (Lobo, 2008; Jiménez–Valverde et al., 2008). Here, the area of the study region is the whole area of the Ibero–Balearic region (see above), and the distribution range of the species within such region was estimated as the minimum convex polygon (*i.e.* the smallest polygon in which no internal angle exceeds 180 degrees) that contains all presence sites (also called convex–hull; Burgman & Fox, 2003). Thus, ROA measures whether the allocation of presence points in the study area shows a relatively wide distribution (as ROA values tend to 1) or a more restricted pattern.

Species traits

We examined the correlation between model accuracy and six ecological and biogeographical characteristics of the species: niche marginality, the total extent of the distribution range (TER), habitat type, habitat detectability, trophic group, and dispersal ability. Raw data on these species' traits were collected from published information on their life histories and biogeography, and then classified into categories. The degree of specialization of each species was estimated from its marginality scores obtained with ENFA (Hirzel et al., 2007). ENFA measures the average position of the species' niche according to the observed localities of presence in relation to the average environmental conditions in the study area; high marginality values indicate a tendency to inhabit extreme conditions regarding the overall conditions in the considered region. TER is a qualitative variable with three categories that represent the total extent and the general distribution of the species: Iberian and Ibero–Maghrebian species (C–I), European species (C–II), and Euroasiatic species (C–III). The type of habitat generally inhabited by the species was also classified into three categories: T–I (woodlands and mountainous habitats), T–II (open habitats such as grasslands, rocky slopes, etc) and T–III (humid and riparian conditions). Habitat detectability refers to the ease of detecting suitable habitat patches for each species using GIS–based data. Each species was classified according to its belonging to habitats of either low– or high–detectability. Low–detectability habitats were considered as those that are usually smaller than the resolution used in GIS data on land cover, including microhabitats such as specific host plants, under stones or river banks. Conversely, high–detectability habitats were taken to be those that were easily identifiable using GIS data, such as extensive woodlands, grasslands or mountainous areas. Species were also classified into two trophic groups according to their trophic range, phytophagous (P) or non–phytophagous (NP) species. Finally, the dispersal ability of the species was measured as a binary variable accounting for whether they are able to fly or not (table 1).

Table 4. Relationships between the three measures of model accuracy (AUC, sensitivity and specificity) and data or species' characteristics. Spearman rank correlation coefficients (R) used to assess the effect of continuous variables (upper rows); partial correlations (R_P) used to assess the individual relevance of N and ROA (lower rows). The effects of qualitative species characteristics were assessed using Kruskal–Wallis (H) and Mann–Whitney U–test (Z), either on the direct values of the accuracy measures (upper rows), or on the regression residuals on N and ROA (lower rows): M. Marginality; HT. Habitat type; TL. Trophic level; FC. Flight capacity; HD. Habitat detectability; * Statistically significant relationships ($p < 0.05$). Variables significant after applying a Bonferroni correction ($p < 0.0060$) are shown in bold. TER is the total extent of the distribution range (see text and table 1).

*Tabla 4. Relaciones entre las tres medidas de precisión del modelo (AUC, sensitividad y especificidad) y las características de los datos o de las especies. Los coeficientes de correlación de Spearman (*R*) se usan para evaluar el efecto de las variables continuas (filas superiores); las correlaciones parciales (*Rp*) se usan para evaluar la relevancia individual de N y ROA (filas inferiores). Los efectos de las características cualitativas de las especies sobre los valores directos de las medidas de precisión (filas superiores), o sobre los residuos de su regresión sobre N y ROA (filas inferiores), se evaluaron mediante Kruskal–Wallis (*H*) y el test U de Mann–Whitney (*Z*): M. Marginalidad; HT. Tipo de hábitat; TL. Nivel Trófico; FC. Capacidad de vuelo; HD. Detectabilidad del hábitat; * Relaciones estadísticamente significativas (p < 0,05). Las variables significativas tras aplicar la corrección de Bonferroni (p < 0,0060) se muestran en negrita. TER es la extensión total del rango de distribución (ver texto y tabla 1).*

		Data characteristics		Species characteristics					
		N	ROA	M	TER	HT	TL	FC	HD
GAM									
	AUC	$R = 0.47$	$R = -0.25$	$R = 0.41$	$H = 3.39$	$H = 7.85$	$Z = 1.36$	$Z = 0.28$	$Z = 1.25$
		$p = 0.03$*	$p = 0.28$	$p = 0.08$	$p = 0.2$	$p = 0.02$*	$p = 0.2$	$p = 0.8$	$p = 0.21$
		$R_P = 0.75$	**$R_P = -0.74$**	$R = 0.16$	$H = 5.98$	$H = 8.52$	$Z = 1.21$	$Z = 0.66$	$Z = 2.16$
		$p < 0.001$	**$p < 0.001$**	$p = 0.48$	$p = 0.05$	$p = 0.01$*	$p = 0.23$	$p = 0.5$	$p = 0.03$*
	Sensitivity	$R = 0.52$	$R = -0.16$	$R = 0.30$	$H = 3.20$	$H = 7.13$	$Z = 0.94$	$Z = 0.14$	$Z = 1.21$
		$p = 0.02$*	$p = 0.48$	$p = 0.19$	$p = 0.20$	$p = 0.03$*	$p = 0.34$	$p = 0.88$	$p = 0.22$
		$R_P = 0.77$	**$R_P = -0.78$**	$R = 0.11$	$H = 4.05$	$H = 6.17$	$Z = 0.45$	$Z = 0.66$	$Z = 2.01$
		$p < 0.001$	**$p < 0.001$**	$p = 0.63$	$p = 0.13$	$p = 0.04$*	$p = 0.65$	$p = 0.5$	$p = 0.04$*
	Specificity	$R = 0.51$	$R = -0.15$	$R = 0.38$	$H = 4.11$	$H = 8.21$	$Z = 1.21$	$Z = 0.18$	$Z = 1.40$
		$p = 0.019$*	$p = 0.51$	$p = 0.09$	$p = 0.13$	$p = 0.02$*	$p = 0.22$	$p = 0.85$	$p = 0.16$
		$R_P = 0.66$	**$R_P = -0.66$**	$R = 0.25$	$H = 3.16$	$H = 3.76$	$Z = 0.38$	$Z = 1.23$	$Z = 1.56$
		$p = 0.002$	**$p = 0.002$**	$p = 0.29$	$p = 0.2$	$p = 0.15$	$p = 0.7$	$p = 0.22$	$p = 0.12$
GLM									
	AUC	**$R = 0.75$**	$R = 0.33$	$R = 0.08$	$H = 2.75$	$H = 5.04$	$Z = 2.19$	$Z = 0.28$	$Z = 0.87$
		$p < 0.001$	$p = 0.15$	$p = 0.72$	$p = 0.25$	$p = 0.08$	$p = 0.02$*	$p = 0.77$	$p = 0.38$
		$R_P = 0.54$	$R_P = -0.31$	$R = 0.19$	$H = 2.63$	$H = 2.71$	$Z = 0.98$	$Z = 0.00$	$Z = 1.18$
		$p = 0.016$*	$p = 0.2$	$p = 0.42$	$p = 0.27$	$p = 0.25$	$p = 0.32$	$p = 1.00$	$p = 0.24$
	Sensitivity	**$R = 0.74$**	$R = 0.29$	$R = 0.15$	$H = 3.72$	$H = 4.52$	$Z = 2.04$	$Z = 0.09$	$Z = 0.42$
		$p < 0.001$	$p = 0.21$	$p = 0.52$	$p = 0.15$	$p = 0.10$	$p = 0.04$*	$p = 0.92$	$p = 0.68$
		$R_P = 0.57$	$R_P = -0.37$	$R = 0.27$	$H = 2.93$	$H = 2.57$	$Z = 1.06$	$Z = -0.28$	$Z = 0.57$
		$p = 0.01$*	$p = 0.12$	$p = 0.24$	$p = 0.23$	$p = 0.27$	$p = 0.29$	$p = 0.77$	$p = 0.57$
	Specificity	$R = 0.26$	$R = -0.26$	$R = 0.39$	$H = 1.29$	$H = 9.15$	$Z = 0.38$	$Z = 0.75$	$Z = 1.71$
		$p = 0.26$	$p = 0.26$	$p = 0.08$	$p = 0.52$	$p = 0.01$*	$p = 0.71$	$p = 0.45$	$p = 0.09$
		$R_P = 0.51$	$R_P = -0.49$	$R = 0.26$	$H = 1.93$	$H = 2.34$	$Z = -0.22$	$Z = 1.42$	$Z = 1.4$
		$p = 0.023$*	$p = 0.03$*	$p = 0.27$	$p = 0.38$	$p = 0.31$	$p = 0.82$	$p = 0.15$	$p = 0.16$

Table 4. (Cont.)

		Data characteristics		Species characteristics					
		N	ROA	M	TER	HT	TL	FC	HD
NNET									
	AUC	R = 0.035	R = –0.39	**R = 0.69**	H = 3.45	H = 8.75	Z = 1.29	Z = 0.56	Z = 1.78
		p = 0.88	p = 0.08	**p < 0.001**	p = 0.17	p = 0.01*	p = 0.19	p = 0.57	p = 0.07
		R_P = 0.70	**R_P = –0.72**	R = 0.42	H = 4.04	H = 4.53	Z = 0.45	Z = 1.32	Z = 1.63
		p < 0.001	**p < 0.001**	p = 0.07	p = 0.13	p = 0.1	p = 0.65	p = 0.18	p = 0.10
	Sensitivity	R = 0.32	R = –0.15	**R = 0.59**	H = 4.12	H = 8.84	Z = 1.43	Z = 0.80	Z = 1.71
		p = 0.17	p = 0.53	**p < 0.001**	p = 0.13	p = 0.01*	p = 0.15	p = 0.42	p = 0.08
		R_P = 0.73	**R_P = –0.74**	R = 0.49	H = 4.66	H = 5.33	Z = 0.68	Z = 1.51	Z = 1.86
		p < 0.001	**p < 0.001**	p = 0.02*	p = 0.09	p = 0.07	p = 0.49	p = 0.13	p = 0.06
	Specificity	R = 0.34	R = –0.14	**R = 0.57**	H = 4.38	H = 8.94	Z = 1.51	Z = 0.85	Z = 1.78
		p = 0.14	p = 0.56	**p < 0.001**	p = 0.11	p = 0.01*	p = 0.13	p = 0.39	p = 0.07
		R_P = 0.73	**R_P = –0.74**	R = 0.48	H = 5.03	H = 6.09	Z = 0.91	Z = 1.42	Z = 1.94
		p < 0.001	**p < 0.001**	p = 0.03*	p = 0.08	p = 0.04*	p = 0.36	p = 0.16	p = 0.05

Evaluation of the influence on model performance

We individually examined whether any of the data characteristics or species' traits correlated with the measures of model performance by using non–parametric statistical tests. The influence of continuous variables (N, ROA and marginality) was assessed using Spearman rank correlations (*Rs*) with each one of the accuracy measures (AUC, sensitivity and specificity). Here, partial correlation analysis was also used to estimate the single contribution of N and ROA on the variation of accuracy measures. The degree of association between model accuracy measures and the qualitative variables (TER, habitat type, habitat detectability, trophic group and dispersal ability) was established using non–parametric statistical tests such as Kruskal–Wallis or Mann–Whitney U.

In addition, to eliminate the influence of data characteristics, we regressed accuracy values against N and ROA. Residuals from these regression analyses were later submitted to a new correlation –either Kruskal–Wallis or Mann–Whitney U tests– to evaluate their relationships with the studied species' traits, applying both a standard significance level (p < 0.05) and a Bonferroni correction for multiple comparisons (p = 0.05/9 = 0.006).

Results

The high accuracy achieved on average (mean AUC ± SD = 0.951 ± 0.013; mean specificity = 0.917 ± 0.017; mean sensitivity = 0.903 ± 0.017; table 3) was to some extent expected, as absences lay outside the envelope defined by the presences and validation data were not spatially independent (Veloz, 2009). Neither AUC nor specificity or sensitivity values differed significantly between the three SDM techniques (Kruskal–Wallis test; n = 60; AUC: H = 3.98, p = 0.14; specificity: H = 3.26, p = 0.20; sensitivity: H = 4.20, p = 0.10). Neither did the area calculated for the potential distribution of the studied species differ significantly between the three modeling techniques (Kruskal–Wallis test; n = 60; H = 0.31, p = 0.86).

Among the considered variables N, ROA and, to a lesser extent, marginality significantly (p < 0.006) affected the accuracy of distribution models (table 4). Several traits (habitat type, trophic group and habitat detectability) were also associated with model accuracy measures (p < 0.05), although their influence was much lower than data characteristics and was not significant when a Bonferroni correction was applied. In contrast, TER and flight capacity did not seem to influence any measure of model accuracy.

As expected from previous studies, species with greater N obtained higher model accuracies; AUC values and sensitivity scores were higher when models were developed from samples for which there were more than 200 records (fig. 1; see appendix 1). Partial correlation analyses of both data characteristics (N and ROA) on accuracy measures showed that while sample size was always positively and significantly correlated with model accuracy, ROA was usually negatively correlated (seven out of nine; see table 4).

The species' traits showed less influence on model performance. Marginality values showed a statisti-

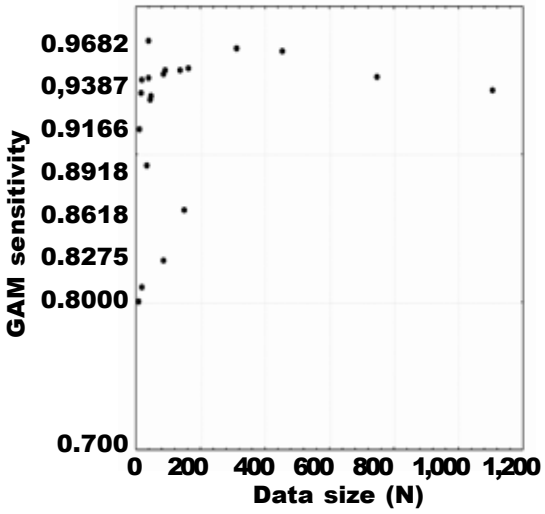


Fig. 1. Correlation between sensitivity of GAM models and data size (N), an example of how the number of occurrences influences accuracy scores. Similar results were obtained with specificity and AUC metrics (see Appendix 1).

Fig. 1. Correlación entre la sensitividad de los modelos GAM y el tamaño de muestra (N), un ejemplo de cómo el número de ocurrencias influye sobre los valores de precisión. Resultados similares se obtuvieron con la especificidad y los valores de AUC (ver Apéndice 1).

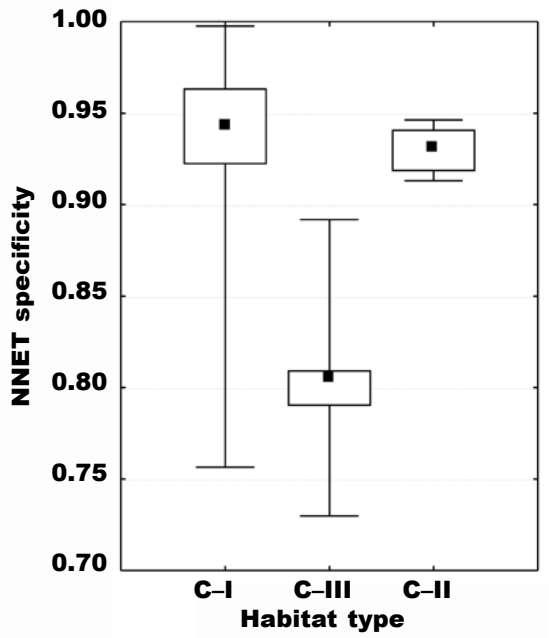


Fig. 2. Specificity of NNET results by habitat type (C–I. Woods and Mountainous habitats C–II. Grasslands and varied habitats, C–III. Riparian and humid habitats). Less accurate models are obtained for species associated to riparian and humid habitats. Similar results were obtained with sensitivity and AUC metrics (see Appendix 2). The middle point shows the median response for each habitat type and specificity score combination. The bottom and top of the box show the 25 and 75 percentiles respectively. The whiskers show minimum and maximum values.

Fig. 2. Resultados de especificidad de los modelos NNET en función del tipo de hábitat (C–I. Bosques y hábitats montañosos, C–II. Praderas y hábitats mixtos, C–III. Hábitats húmedos y riparios). Las especies asociadas a hábitats húmedos y riparios obtuvieron modelos menos precisos. Se obtuvieron resultados similares con las medidas de sensitividad y AUC (ver Apéndice 2). El punto central representa el valor de la combinación de la mediana para cada tipo de hábitat con los valores de especificidad, los límites inferiores y superiores de la caja muestran los percentiles 25 y 75 respectivamente. Los bigotes señalan el valor máximo y mínimo.

cally significant correlation with accuracy until the effect of N and ROA was removed. The only species' trait that remained relevant for accuracy measures after accounting for data characteristics was habitat type; predictions for species associated with humid and riparian conditions were poorer (see figs. 2, 3 and appendix 2). However, this association was not statistically significant when a Bonferroni correction was applied. Other associations, such as the trophic range of species and GLM accuracy or habitat detectability and GAM performance, also ceased to be significant under the more restrictive Bonferroni significance levels.

Discussion

Several species' traits and data characteristics have shown to influence SDM performance (*e.g.* Brotons et al., 2004; Segurado & Araújo, 2004; Seoane et al., 2005; Hernández et al., 2006; Marmion et al., 2008). One of these characteristics is the prevalence in the dataset, which is generally thought to affect the accuracy of models (*e.g.* McPherson et al., 2004; Seoane et al., 2005; Marmion et al., 2008). These effects, however, might only appear in extreme prevalence values (see Jiménez–Valverde et al., 2009). Here we deliberately equalled the prevalence of all species' datasets to avoid its effect on model performance. We also removed the effect of data size and ROA using residual analyses, and our results showed that, when these mere methodological artefacts were controlled the supposed differences in model performance attributed to the ecological or biogeographical traits of

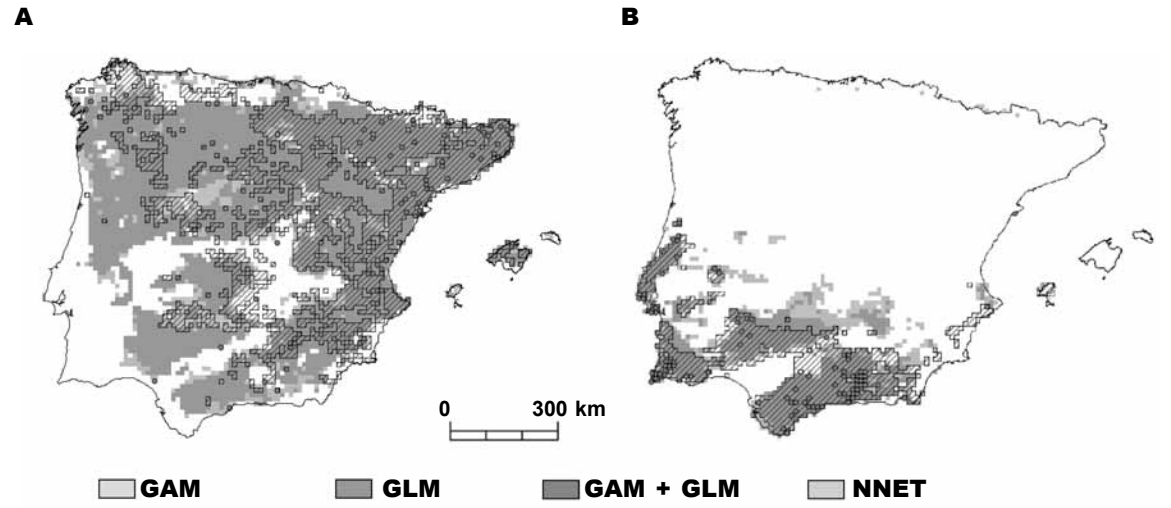


Fig. 3. Differences between the predictive maps produced for a riparian species, *Coenagrion mercuriale* (A) and a species not linked to riparian habitats, *Cupido lorquinii* (B). Although data for both species have the same sample size (N = 87), GAM and NNET models performed better for *C. lorquinii* than for *C. mercuriale*. Note that the difference in ROA values (*C. lorquinii* = 0.267; *C. mercuriale* = 0.629) could have also influenced this disparity.

Fig. 3. Diferencias entre los mapas predictivos obtenidos para una especie riparia, Coenagrion mercuriale *(A) y una especie no ligada a hábitats riparios,* Cupido lorquinii *(B). Aunque los datos de ambas especies tienen el mismo tamaño de muestra (N = 87), los modelos GAM y NNET obtuvieron mejores resultados para* C. lorquinii *que para* C. mercuriale*. Obsérvese que la diferencia en los valores de ROA (*C. lorquinii = *0,267;* C. mercuriale = *0,629) también podrían haber influido en esta disparidad.*

species tended to disappear. This result coincides with the findings of Santika (2010), who examined the influence of prevalence on simulated data. Such dependence on data characteristics, and the fact that these characteristics also affect the measures of SDM performance, makes us wonder whether it is possible to find an accuracy measure able to compare the performance of SDMs among different species (see Lobo et al., 2008).

Larger sample sizes have previously been shown to increase model accuracy (Stockwell & Peterson, 2002; McPherson et al., 2004; Hernández et al., 2006; Wisz et al., 2008; Mateo et al., 2010). In this work, sample size had a positive and significant effect on model performance. In spite of the significant and positive correlation between sample size and ROA (R_S = 0.65, p = 0.0017; fig. 4), ROA did not show any direct relationship with model performance when analyzed individually.

Ecological traits of several species also seemed to influence model performance, though to a lesser degree. The species with most restricted ecological requirements (*i.e.*, the most marginal species) were modelled more accurately than less specialized species, but only in the case of NNET. In contrast with other studies (Brotons et al., 2004; Segurado & Araújo, 2004; Luoto et al., 2005), we did not find any strong

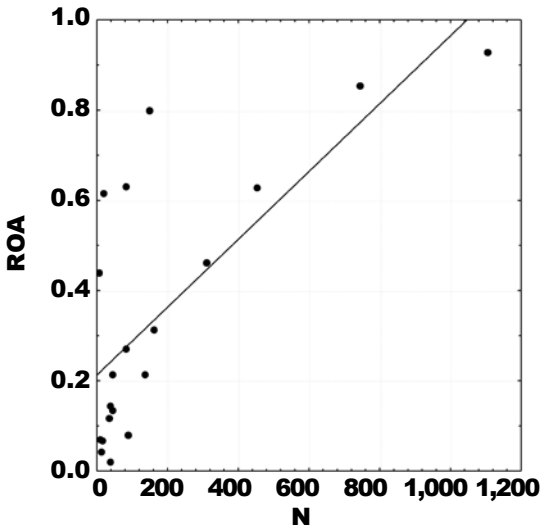


Fig. 4. Relationship between the values of sample size (N) and the relative occurrence area (ROA).

Fig. 4. Relación entre los valores del tamaño de muestra (N) y el área de presencia relativa (ROA).

relationship between the performance of GAM and GLM models and niche specialization (*i.e.,* marginality). This is in agreement with Pöyry et al. (2008) and Newbold et al. (2009), who were not able to detect any effect of the niche width of butterflies regarding model accuracy. Besides, all SDM techniques, but specially GAM and NNET, seemed to perform better with species not associated with riparian and humid conditions, a result also found by McPherson & Jetz (2007). Such poor performance may be associated with a poorer localization of wetlands in land cover maps. This hampers the inclusion of predictor variables related to the quality of aquatic habitats, thereby impeding the use of the true determinants of the distribution of riparian species. Finally, we did not detect any influence of the variables measuring flight capacity and the total extent of the distribution range of the species on model accuracy. Hence, it can be assumed that the SDM techniques used are not sensitive to either how widespread the species is outside the study area, or to its dispersal capacity.

However, sample size and ROA altogether seem to interact with the influence of species' traits on model accuracy. Once the effects of these data characteristics are removed, only a few effects of species' traits remain. In particular, the residual analyses reveal a consistent, though weak, relationship between model performance and habitat detectability; species associated to easy–to–detect habitats are predicted more accurately by GAM models than those whose preferred habitats are smaller than the resolution of the available GIS layers. This also agrees with the results obtained by McPherson & Jetz (2007), where habitat detectability also had a secondary role on model accuracy. Besides, this result supports the idea commented above: the low detectability of riparian and humid habitats could be associated with the incapacity of the predictor variables used here (which represent the most commonly used ones) to capture the species' response to environmental conditions. On the other hand, the weak relationship between the better performance of GLMs for phytophagous species (in comparison with non–phytophagous species) disappears after removing the effect of N and ROA, revealing that this minor relationship could be a spurious statistical artefact. Further analyses are needed to evaluate whether other species traits not considered in this work are important for the performance of SDMs, beyond the mere limitations of data characteristics such as N or ROA.

The limitations of this study, such as data scarcity, low spatial resolution, and lack of reliable absence data and independent validation data sets, are common when working with rare invertebrate species. These constraints, and especially the lack of reliable absence data, are also under the common choice of using background absences, which are randomly selected from the considered extent. The use of background absences generates spatial representations of the distribution of the species that are placed in an unknown situation within the realized–potential gradient described by Jiménez–Valverde et al. (2008), depending on the Relative Occurrence Area (Lobo et al., 2010). Thus, the dependency of the accuracy measures on the ROA invalidates any further assessment of the relationships between these accuracy values and the predictor variables, which are also dependent on the ROA. To minimize this drawback, instead of using background absences, here we use pseudoabsences that are *a priori* located under environmentally unsuitable conditions. By accounting for the limitations of AUC as a measure of model accuracy, our approach identifies some factors that are related to the performance of representing potential distributions.

Our results confirm that although some species' traits may affect SDM performance, prediction accuracy is mostly affected by the characteristics of the data. The separate effects of N and ROA are difficult to determine due to the unavoidable correlation between them (species recorded in more cells have a higher probability of being widely distributed in the region). For this reason, an unknown proportion of the effect on model performance generally attributed to low sample sizes may be due to a less relative occurrence area of presence data in the studied region; *i.e.,* the inability to select reliable absences outside environmental domain used by the species when the number of observations is low (Austin & Meyers, 1996). Given the overall good results obtained by the three methods according to the standard measures of model evaluation, we consider more attention should be given to assessing the quality and/or adequacy of the data rather than selecting a particular SDM technique. Similar results were obtained by Syphard & Franklin (2010), who found that ecological and range characteristics of the species have a greater effect on model performance than the choice of SDM method. In this study, species were modelled more accurately when samples were larger no matter which technique was used. Moreover, ROA had an additive effect to that of sample size, showing that selecting coarse extents of analysis to model the distribution of geographically restricted species may result in trivial models. These models are able to discriminate such restricted distributions within a large geographical context and they therefore yield highly accurate measures of performance, but they are unable to provide reliable descriptions of the environmental response of the species (Lobo, 2008; Jiménez–Valverde et al., 2008; VanDerWal et al., 2009).

Our results suggest researchers should avoid any between–species comparison of SDM results while selecting the most adequate technique. We alternatively suggest carrying out species by species SDMs, ensuring that the amount of data available is sufficient and that the geographical focus (*i.e.,* extent) of the analysis is adequate to recover the environmental response of each particular species. In addition, special care should be taken while modelling species inhabiting inconspicuous habitats or strongly affected by interactions occurring at small spatial scales (see Hortal et al*.,* 2010). The problems associated with predicting the distributions of these species should be tackled either by using more precise predictors or by resizing the scale (*i.e.,* grain) of the analyses.

Acknowledgements

This research was supported by the Spanish Ministry of Science and Innovation through the project CGL2008–03878. J. H. received funding from a CSIC JAE–Doc grant.

References

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

Austin, M. P. & Meyers, J. A., 1996. Current approaches to modelling the environmental niche of eucalypts: implications for management of forest biodiversity. *Forest Ecology and Management*, 85: 95–106.

Beaumont, L. J., Gallagher, R. V., Downey, P. O., Thuiller, W., Leishman, M. R. & Hughes, L., 2009. Modelling the impact of *Hieracium* spp. on protected areas in Australia under future climates. *Ecography*, 32: 757–764.

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.

Burgman, M. A. & Fox, J. C., 2003. Bias in species range estimates from minimum convex polygons: implications for conservation and options for improved planning. *Animal Conservation*, 6: 19–28.

Busby, J. R., 1991. BIOCLIM–a bioclimate analysis and prediction system. In: *Nature Conservation: Cost effective biological surveys and data analysis:* 64–68 (C. R. Margules & M. P. Austin, Eds.). CSIRO, Melbourne.

Cabeza, M., Arponen, A., Jäättelä, L., Kujala, H., van Teeffelen, A. & Hanski, I., 2010. Conservation planning with insects at three different spatial scales. *Ecography*, 33: 54–63.

Carrascal, L. M., Seoane, J., Palomino, D., Alonso, C. L., Lobo, J. M., 2006. Species–specific features affect the ability of census–derived models to map winter avian distribution. *Ecological Research*, 21: 681–691.

Castillejo, J., 1990. Babosas de la Península Ibérica. I. Los Ariónidos. Catálogo crítico y mapas de distribución (Gastrópoda, Pulmonata, Arionidae). *Iberus*, 9: 331–345.

Chefaoui, R. M. & Lobo, J. M., 2007. Assessing the conservation status of an Iberian moth using pseudo–absences. *The Journal of Wildlife Management*, 8: 2507–2516.

– 2008. Assessing the effects of pseudo–absences on predictive distribution model performance. *Ecological Modelling*, 210: 478–486.

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. *Biological Conservation*, 122: 327–338.

Clark Labs, 2000. *Global Change Data Archive Vol. 3. 1 km Global Elevation Model*. Clark University.

– 2003. *Idrisi Kilimanjaro*. Clark Labs, Worcester, MA.

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

Elith, J. & Leathwick, J. R., 2007. Predicting species' distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Diversity and Distributions*, 13: 165–175.

Franklin, J., 2009. *Mapping species distributions: Spatial inference and prediction.* Cambridge University Press, New York.

Galante, E. & Verdú, J. R., 2000. *Los Artrópodos de la 'Directiva Hábitat' en España*. Ministerio de Medio Ambiente, Madrid.

García–Barros, E. & Herranz, J., 2001. Nuevas localidades de *Proserpinus proserpina* (Pallas, 1772) y *Graellsia isabelae* (Graells, 1849) del centro peninsular. *SHILAP Revista de Lepidopterología*, 29: 183–184.

Grenyer, R., Orme, C. D. L., Jackson, S. F., Thomas, G. H., Davies, R. G., Davies, T. J., Jones, K. E., Olson, V. A., Ridgely, R. S., Rasmussen, P. C., Ding, T. S., Bennett, P. M., Blackburn, T. M., Gaston, K. J., Gittleman, J. L. & Owens, I. P. F., 2006. Global distribution and conservation of rare and threatened vertebrates. *Nature*, 444: 93–96.

Grosso–Silva, J. M., 1999. Contribuição para o conhecimento dos lucanídeos (Coleoptera, Lucanidae) de Portugal. *Boletín de la S.E.A.*, 25: 11–15.

Grupo de Trabajo sobre Lucanidae Ibéricos, 2000. Proyecto Ciervo Volante (PCV). *Boletín de la S.E.A.*, 27: 108–109.

Hanspach, J., Kühn, I., Pompe, S. & Klotz, S., 2010. Predictive performance of plant species distribution models depends on species traits. *Perspectives in Plant Ecology, Evolution and Systematics*, 12: 219–225.

Hernández, P. A., Graham, C. H., Master, L. L. & Albert, D. L., 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, 29: 773–785.

Hirzel, A. H., Hausser, J. & Perrin, N., 2007. *Biomapper 4.0.* Division of Conservation Biology, University of Bern, Laussane.

Hortal, J., Jiménez–Valverde, A., Gómez, J. F., Lobo, J. M. & Baselga, A., 2008. Historical bias in biodiversity inventories affects the observed realized niche of the species. *Oikos*, 117: 847–858.

Hortal, J., Lobo, J. M. & Jimenez–Valverde, A., 2007. Limitations of biodiversity databases: case study on seed–plant diversity in Tenerife, Canary Islands. *Conservation Biology*, 21: 853–863.

Hortal, J., Roura–Pascual, N., Sanders, N. J. & Rahbek, C., 2010. Understanding (insect) species distributions across spatial scales. *Ecography*, 33: 51–53.

Instituto Geográfico Nacional, 1995. *Mapa de suelos del Atlas Nacional de España (Edafología)*. CSIC/IRNAS, Sevilla.

Jiménez–Valverde, A. & Lobo, J. M., 2006. The ghost of unbalanced species distribution data in geographical model predictions. *Diversity and Distributions*, 12: 521–524.

– 2007. Threshold criteria for conversion of probability of species presence to either–or presence–absence. *Acta Oecologica*, 31: 361–369.
Jiménez–Valverde, A., Lobo, J. M. & Hortal, J., 2008. Not as good as they seem: the importance of concepts in species distribution modelling. *Diversity and Distributions*, 14: 885–890.
– 2009. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. *Community Ecology*, 10: 196–205.
King, G. & Zeng, L., 2000. Logistic regression in rare events data. *The Global Burden of Disease 2000 in Aging Populations, Research Paper No. 2*. Havard Burden of Disease Unit, Cambridge. <http://www.hsph.harvard.edu/burdenofdisease/publications /papers/Logistic%20Regression.pdf> 6th October 2010.
Legendre, P. & Legendre, L., 1998. *Numerical Ecology*. Elsevier, Amsterdam.
Lobo, J. M., 2008. More complex distribution models or more representative data? *Biodiversity Informatics*, 5: 14–19.
Lobo, J. M., Baselga, A., Hortal, J., Jiménez–Valverde, A. & Gómez, J. F., 2007. How does the knowledge about the spatial distribution of Iberian dung beetle species accumulate over time? *Diversity and Distributions*, 13: 772–780.
Lobo, J. M., Jiménez–Valverde, A. & Hortal, J., 2010. The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, 33: 103–114.
Lobo, J. M., Jiménez–Valverde, A. & Real, R., 2008. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17: 145–151.
Lobo, J. M., Verdú, J. R. & Numa, C., 2006. Environmental and geographical factors affecting the Iberian distribution of flightless *Jekelius* species Coleoptera: (Geotrupidae). *Diversity and Distributions*, 12:179–188.
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.
Luoto, M., Pöyry, J., Heikkinen, R. K. & Saarinen, K., 2005. Uncertainty of bioclimate envelope models based on the geographical distribution of species. *Global Ecology and Biogeography*, 14: 575–584.
Marmion, M., Luoto, M., Heikkinen, R. K. & Thuiller, W., 2008. The performance of state–of–the–art modelling techniques depends on geographical distribution of species. *Ecological Modelling*, 220: 3512–3520.
Martínez–Orti, A., 2004. Descripción de los moluscos terrestres del Valle del Najerilla. *Noticiario de la Sociedad Española de Malacología*, 41: 30–32.
Mateo, R. G., Felicísimo, Á. M. & Muñoz, J., 2010. Effects of the number of presences on reliability and stability of MARS species distribution models: the importance of regional niche variation and ecological heterogeneity. *Journal of Vegetation Science*, 21: 908–922.
McCullagh, P. & Nelder, J. A., 1989. *Generalized Linear Models*. Chapman & Hall, London.
McPherson, J. M. & Jetz, W., 2007. Effects of species' ecology on the accuracy of distribution models. *Ecography*, 30: 135–151.
McPherson, J. M., Jetz, W. & Rogers, D. J., 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology*, 41: 811–823.
Newbold, T., Reader, T., Zalat, S., El–Gabbas, A. & Gilbert, F., 2009. Effect of characteristics of butterfly species on the accuracy of distribution models in an arid environment. *Biodiversity and Conservation*, 18: 3629–3641.
Olden, J. D., Jackson, D. A. & Peres–Neto, P., 2002. Predictive models of fish species distributions: a note on proper validation and chance predictions. *Transactions of the American Fisheries Society*, 131: 329–336.
Pérez–Bote, J. L., García, J. M., Ferri, F. & Moreno, J. A., 2001. Nueva cita de *Lucanus cervus* (Linnaeus, 1758) en Extremadura (Coleoptera: Lucanidae). *Boletín de la S.E.A.*, 28: 130.
Pöyry, J., Luoto, M., Heikkinen, R. K. & Saarinen, K., 2008. Species traits are associated with the quality of bioclimatic models. *Global Ecology and Biogeography*, 17: 403–414.
R Development Core Team, 2008. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R–project.org> 6th October 2010.
Raimundo, R., Algarvio, R., Casas Novas, P. & Figuereiredo, D., 2001. Cartography of some species of xilophagous and xilomicetophagous insects in *Quercus suber* and *Quercus rotundifolia* in Alentejo using GIS. *Suplemento ao Boletim da Sociedade Portuguesa de Entomología*, 6: 459–468.
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. *Journal of Biogeography*, 30: 581–590.
Ripley, B. D., 1996. *Pattern Recognition and Neural Networks.* Cambridge University Press.
Rocchini, D., Hortal, J., Lengyel, S., Lobo, J. M., Jiménez–Valverde, A., Ricotta, C., Bacaro, G. & Chiarucci, A., 2011. Uncertainty in species distribution mapping and the need for maps of ignorance. *Progress in Physical Geography*, 35: 211–226.
Rosas, G., Ramos, M. A. & García Valdecasas, A., 1992. *Invertebrados españoles protegidos por convenios internacionales*. ICONA, Madrid.
Santika, T., 2010. Assessing the effect of prevalence on the predictive performance of species distribution models using simulated data. *Global Ecology and Biogeography*, 20: 181–192.
Schröder, B., 2004. *ROC Plotting and AUC Calculation Transferability Test*. Potsdam University, Potsdam. <http://brandenburg.geoecology.uni–potsdam.de/users/schroeder/download.html> 6th October 2010.
Segurado, P. & Araújo, M. B., 2004. An evaluation of methods for modelling species distributions.

Journal of Biogeography, 31: 1555–1568.

Seoane, J., Carrascal, L. M., Alonso, C. L. & Palomino, D., 2005. Species–specific traits associated to prediction errors in bird habitat suitability modelling. *Ecological Modelling*, 185: 299–308.

Soria, S., Abos, F. & Martín, E., 1986. Influencia de los tratamientos con diflubenzurón ODC 45% sobre pinares en las poblaciones de *Graellsia isabelae* (Graells) (Lep. Syssphingidae) y reseña de su biología. *Boletín de sanidad vegetal*, 12: 29–50.

StatSoft, I., 2001. *Statistica.* Data analysis software system, Tulsa.

Stockwell, D. R. B. & Peterson, A. T., 2002. Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, 148: 1–13.

Syphard, A. D. & Franklin, J., 2010. Species traits affect the performance of species distribution models for plants in southern California. *Journal of Vegetation Science*, 21: 177–189.

VanDerWal, J., Shoo, L. P., Graham, C. & Williams, S. E., 2009. Selecting pseudo–absence data for presence–only distribution modeling: How far should you stray from what you know? *Ecological Modelling*, 220: 589–594.

Veloz, S. D., 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence–only niche models. *Journal of Biogeography*, 36: 2290–2299.

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. *Diversity and Distributions*, 8: 259–274.

– 2006. *Libro Rojo de los Invertebrados de España*. Ministerio de Medio Ambiente, Madrid.

Viejo Montesinos, J. L., 1992. Biografía de un naturalista y biología del lepidóptero por él descrito. Graells y la *Graellsia*. *Quercus*, 74: 22–30.

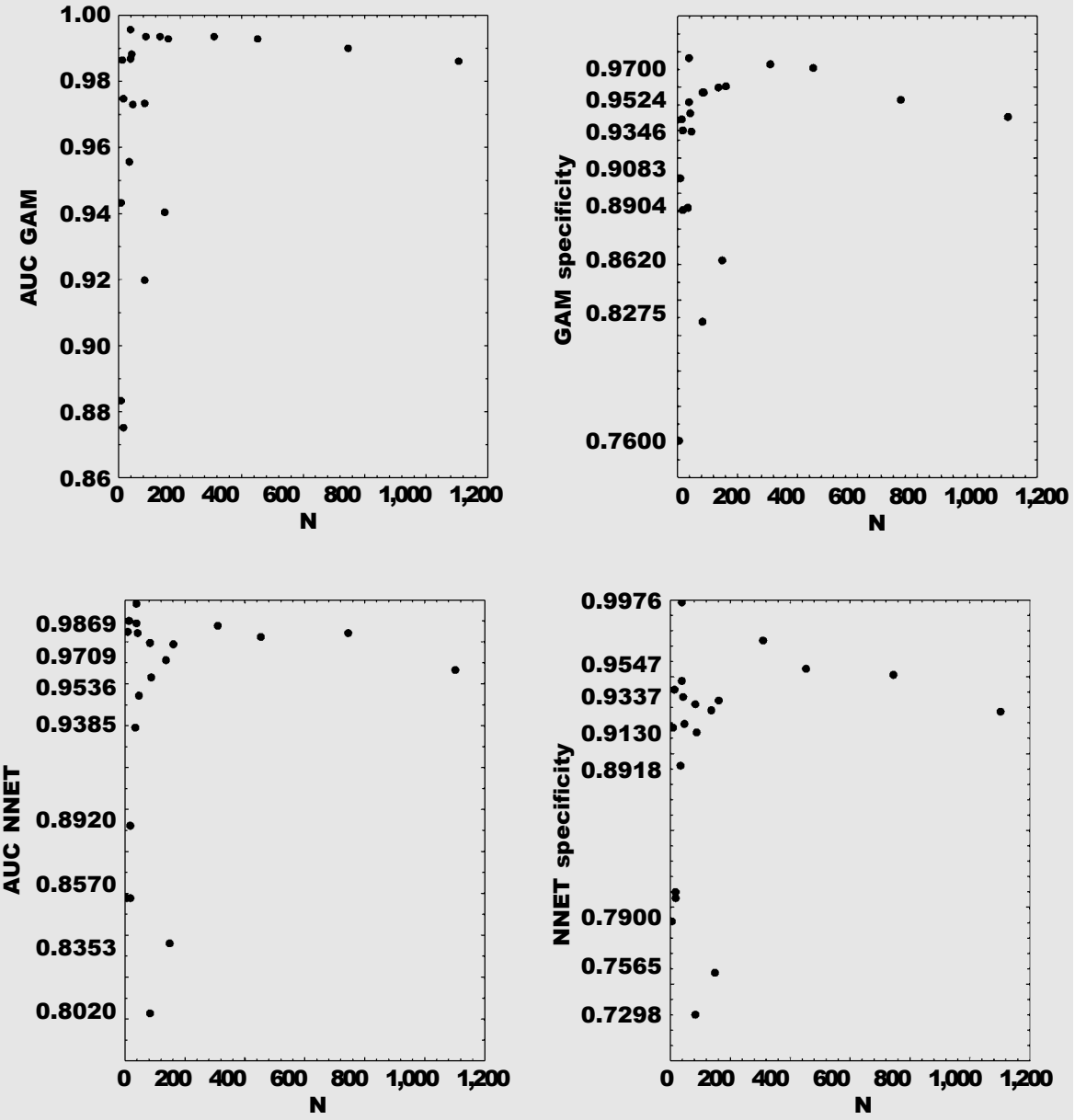
Wisz, M. S., Hijmans, R. J., Li J., Peterson, A. T., Graham, C. H., Guisan, A. & N. P. S. D. W. Group, 2008. Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, 14: 763–773.

Wood, S. N. & Augustin, N. H., 2002. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling*, 157: 157–177.

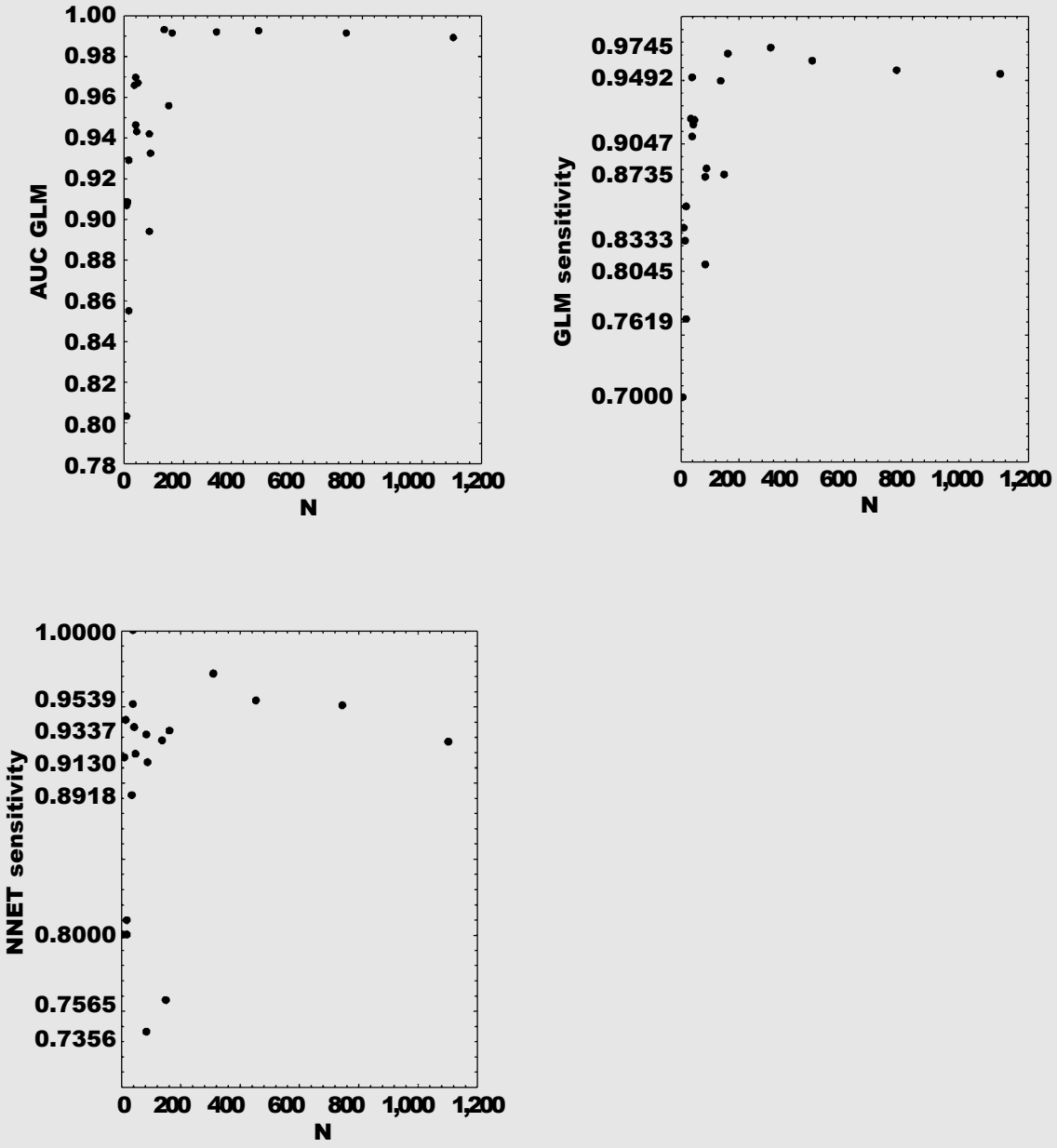
Zaniewski, A. E., Lehmann, A. & Overton, J. M., 2002. Predicting species spatial distributions using presence–only data: a case study of native New Zealand ferns. *Ecological Modelling*, 157: 261–280.

Zweig, M. H. & Campbell, G., 1993. Receiver–operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39: 561–577.

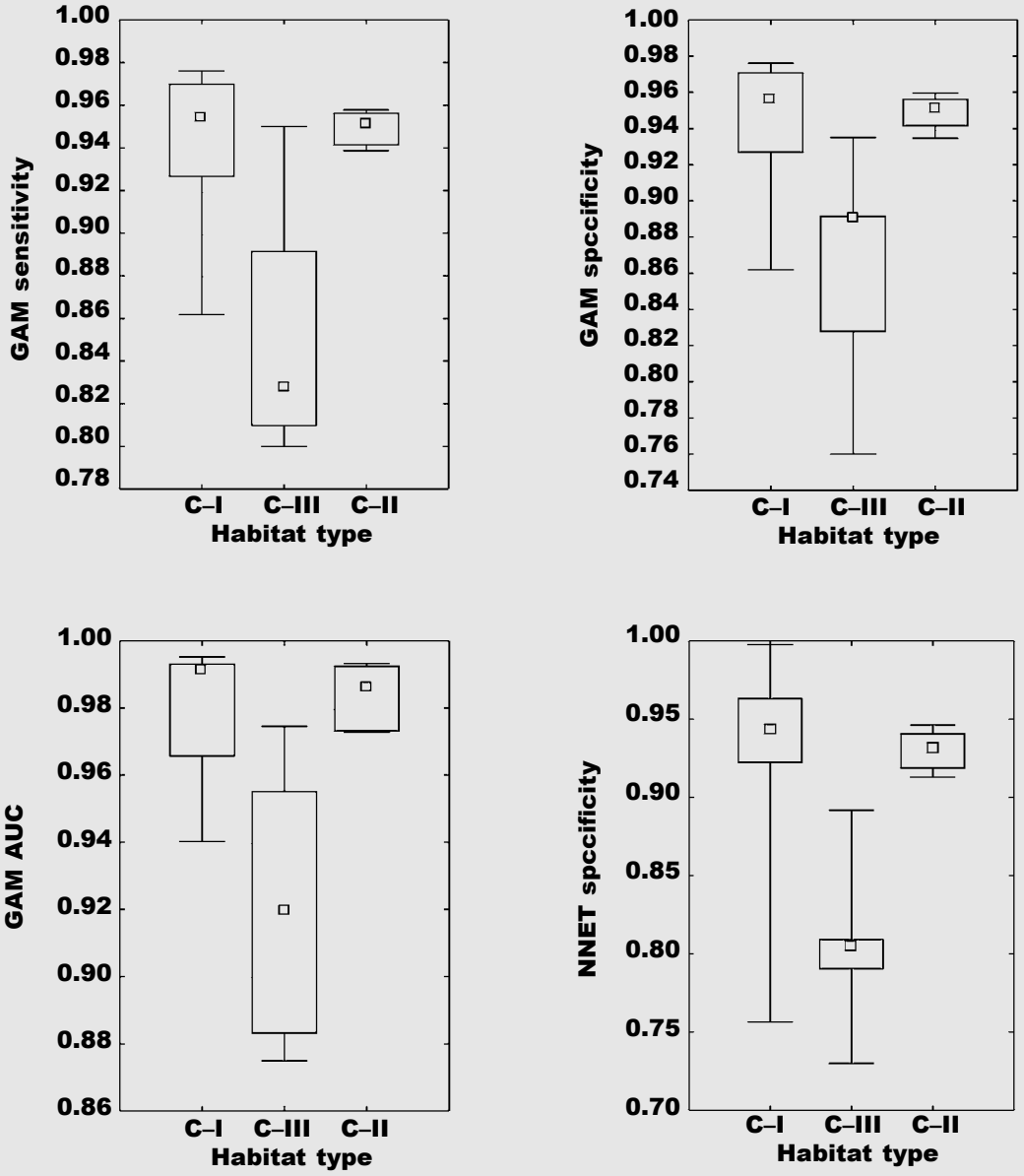
Appendix 1. Scatterplots of significant correlation analyses between accuracy measures (AUC, sensitivity and specificity) and data size (N).



Apéndice 1. Gráfico de dispersión de las corelaciones significativas entre las medidas de precisión (AUC, sensitividad y especificidad) y el tamaño de muestra (N).



Appendix 2. Accuracy measures results by habitat: C–I. Woods and Mountainous habitats; C–II. Grasslands and varied habitats; C–III. Riparian and humid habitats. Less accurate models are obtained for species associated to riparian and humid habitats. The middle point shows the median response, the bottom and top of the box show the 25 and 75 percentiles respectively. The whiskers show minimum and maximum vaues).



Apéndice 2. Precisión obtenida por los modelos en función del tipo de hábitat: C–I. Bosque y hábitats montañosos; C–II. Praderas y hábitats mixtos; C–III. Hábitats húmedos y riparios. Las especies asociadas a hábitats húmedos y riparios obtienen modelos menos precisos. El punto central representa la mediana, los límites inferiores y superiores de la caja muestran los percentiles 25 y 75 respectivamente. Los bigotes señalan el valor máximo y mínimo.