Community Ecology 10(2): 196-205, 2009
1585-8553/$20.00 © Akadémiai Kiadó, Budapest
DOI: 10.1556/ComEc.10.2009.2.9

The effect of prevalence and its interaction with sample size on the reliability of species distribution models

A. Jiménez-Valverde[1]*, J. M. Lobo[2] and J. Hortal[3]

[1] *Natural History Museum and Biodiversity Research Center, The University of Kansas, Lawrence, Kansas 66045, USA. *Corresponding author. E-mail: ajvalv@ku.edu*
[2] *Museo Nacional de Ciencias Naturales, Dpto. Biodiversidad y Biología Evolutiva. C/ José Gutiérrez Abascal 2, E-28006, Madrid, Spain*
[3] *NERC Centre for Population Biology, Division of Biology, Imperial College London, Silwood Park Campus, Ascot, Berkshire SL5 7PY, UK*

Keywords: Logistic regression, Model accuracy, Prevalence, Sample size, Species distribution modelling.

Abstract: Prevalence (the presence/absence ratio in the training data) is commonly thought to influence the reliability of the predictions of species distribution models. However, little is known about its precise impact. We studied its effects using a virtual species, avoiding the presence of unaccounted-for effects in the modeling process (false absences, non-explanatory predictors, etc.). We sampled the distribution of the virtual species to obtain several data subsets of varying sample size and prevalence, and then modeled these data subsets using logistic regressions. Our results show that model predictions can be highly accurate over a wide range of sample sizes and prevalence scores, provided that the predictors are truly related to the distribution of the species and the training data are reliable. The effect of sample size becomes apparent for datasets of less than 70 data points, and the effect of prevalence is significant only for datasets with extremely unbalanced samples (<0.01 and >0.99). There is also a strong interaction between sample size and prevalence, indicating that the most negative factor is the sample size of each event (absence and/or presence), and not biased prevalence, as previously thought. We suggest that, in the real world, an interaction must exist between the sample size of each event and the quality of the training data. We discuss that biased prevalences can be a desirable property of the data, instead of a problem to be avoided, also pointing out the importance of using the best absence data possible when modeling the distribution of species of narrow geographic range.

Abbreviations: AUC–Area Under the Receiver Operating Characteristic Curve; PCA–Principal Component Analysis; ROC–Receiver Operating Characteristic Curve.

Introduction

The prediction of species´ geographic distributions based on their known occurrences is increasingly being used in ecology, aided by both Geographic Information Systems (GIS) and statistical quantification of species-environment relationships (Guisan and Zimmermann 2000, Lehmann et al. 2002, Rushton et al. 2004). Species distribution models identify a mathematical relationship between presence/absence data and a number of predictors. As such, they are used to forecast past and future distributions, to assess the effects of environmental changes on species´ distributions, or to locate undiscovered populations and species, among many other applications. Thus, these techniques can be useful tools in biogeography, paleoecology, evolution, and conservation (e.g., Peterson et al. 1999, Anderson et al. 2002a,b, Schadt et al. 2002, Barbosa et al. 2003, Peterson and Holt 2003, Chefaoui et al. 2005, Jiménez-Valverde and Lobo 2007a, Jiménez-Valverde et al. 2007, 2008a, Nogués-Bravo et al. 2008).

The great variety of methods currently available to model the distributions of species can be divided into two main categories: group discrimination and profile techniques. Group discrimination techniques are methodologies that use both presence and absence data (Guisan and Zimmermann 2000 and Scott et al. 2002). Unlike profile techniques (i.e., those using only the available presence information), group discrimination techniques take absence data into account in the modeling process predictions, and are supposed to allow building more realistic relationships between species distribution and environmental factors (Hirzel et al. 2001, Brotons et al. 2004, Segurado and Araújo 2004). This is because by including information on the geographical locations from where the species is absent in spite of having a priori favorable environmental conditions, model results are closer to the realized distribution of the species (Jiménez-Valverde et al. 2008b, Hirzel and Le Lay 2008).

When using both presences and absences, *prevalence* is defined as the ratio of the number of presences to the total number of data points used in building the model. In general, biased training prevalences (either low or high) are expected to affect negatively model predictions (Pearce and Ferrier 2000, Vaughan and Ormerod 2003). However, the effects of prevalence have been insufficiently examined in the literature on species distribution models. Indeed, published results are inconsistent. Good models can apparently be obtained from low prevalence datasets, while the contrary has also been reported (e.g., Manel et al. 2001, Brotons et al. 2004, Luoto et al. 2005), and others found the best models at prevalence values around 0.5 (McPherson et al. 2004). Therefore, the effect of prevalence on model accuracy remains open to debate. To avoid the supposedly negative impacts of prevalence, some authors have recommended resampling the data to reduce the number of either absences or presences, in order to balance both kinds of events in the dataset (McPherson et al. 2004, Liu et al. 2005). In this sense, others have used the same number of absences and presences to fit the models, in spite of the availability of a higher numbers of absence points (e.g., Osborne et al. 2001, Seoane et al. 2006).

In spite of the common agreement on the negative impacts of unbalanced prevalences, we believe that some of these effects might come from an erroneous interpretation of published results, rather than being an effect of biased prevalences. Instead, these negative effects could appear due to three reasons:

1. The probability values obtained from regression models derived from unbalanced samples are biased towards the more prevalent category, either presences or absences (Hosmer and Lemeshow 1989, Cramer 1999). This statistical phenomenon is unavoidable, and the poorer the fit to the data the more important this effect will be (Cramer 1999). However, its effect on predictive performance is related to the probability threshold used as a cut-off to produce a presence/absence map from the continuous probability scores obtained from the regression model; such probability threshold determines the accuracy of the model obtained from a confusion matrix (Fielding and Bell 1997). For example, using the intuitive 0.5 threshold artificially increases commission errors (i.e., overprediction) in cases of high prevalences, but increases omission errors (i.e., underprediction) when working with low prevalences (see, for example, Manel et al. 1999, McPherson et al. 2004). If the probabilities attributable to the sites where the species is likely present or absent vary according to prevalence, the effect of biased presence/absence data can be avoided just by adjusting the probability threshold in accordance to the frequency of occurrence (Cramer 1999). Recent efforts have aimed at developing a cut-off criterion in accordance with the prevalence in the data (Liu et al. 2005, Jiménez-Valverde and Lobo 2007b). In addition to this, probabilities must be rescaled in order to convert the logistic probabilities into suitability values, in order to account for the prevalence bias (Jiménez-Valverde and Lobo 2006a, Real et al. 2006, Estrada et al. 2008).

2. Although prevalence is a property of the data, it usually covaries with species ecology and range size, i.e., data for rare species usually show low prevalence scores, while the opposite is found for widely-distributed species. Given that the relative occurrence area of the species is also known to influence the values of model performance measures (Lobo et al. 2008, Jiménez-Valverde et al. 2008b), it has been difficult to separate its effect from the true effect of prevalence (e.g., Brotons et al. 2004).

3. Related with point 2, prevalence has been sometimes misunderstood as being a property of the species (i.e., the number of grid cells with presence records/total number of cells in the region), instead of a property of the dataset (i.e., the number of presences used/number of observations used). This conceptual error is worsened by the widespread practice of using all cells in the studied region to model presence/absence data, without separating true biological absences from those due to lack of recording effort (see, e.g., McPherson et al. 2004, Luoto et al. 2005).

Jiménez-Valverde and Lobo (2006a) hypothesized that prevalence might have little effect on the predictive accuracy of species distribution models, and that its supposed negative effects could have been confounded with the impacts of different sources of bias in poor quality distributional data, such as low sample size in any of the two possible events (presences or absences), or lack of representation of the whole environmental gradient. These sources of error are common in the information about most species (see Discussion), especially of rare ones with narrow distribution ranges and/or low population levels. Therefore, if species distribution models are going to be commonly used as a tool in ecological (e.g., Jiménez-Valverde et al. 2008a), biogeographical (e.g., Lobo et al. 2006) or paleoecological (e.g., Nogués-Bravo et al. 2008) studies as well as in Conservation Biogeography (*sensu* Whittaker et al. 2005), understanding their effects on species distribution modeling is of special concern.

The objective of this study is to examine the independent influence of prevalence on the predictive performance of the models, and its possible interaction with sample size, in the absence of other effects that could influence the modeling process. Since true species distributions are never completely known, it is quite difficult to assess the accuracy of distribution models and to determine unequivocally the effects of each particular source of uncertainty. To overcome this drawback, we build a simple virtual species whose distribution is only conditioned by known climate variables in a simple unimodal way, and then model its geographic distribution using the same climate variables as predictors. Thus, by controlling the effect of the adequacy of the explanatory variables used to build the model, some of the shortfalls of modeling species distributions in the real world are avoided, allowing the identification of the genuine effects of interest. Such simplicity in the modeling approach – which is far from the real world – is deliberate. We do so in order to understand the raw effects of prevalence, excluding all other possible sources of noise. This is a necessary first step, in order to understand the potential influences of prevalence and sample size under more realistic and complex scenarios.

Materials and methods

The virtual species

We mapped the distribution of a virtual species using actual climate data in order to build a hypothetical scenario that is representative of the situation when studying true species distribution patterns. Although the use of artificial species distributions to ascertain the influence of the data employed and model functions has generally been neglected in the literature on species distribution modeling (but see Hirzel et al. 2001, Reese et al. 2005, Real et al. 2006, Meynard and Quinn 2007), virtual species are nowadays one of the best ways to make testable experiments in this field (see Austin et al. 2006). Using such simulations allowed us to:

> i) ensure that the modeling method could be able to predict correctly species distribution, provided that all relevant predictors are used; this avoids potential biases due to contingent, unaccounted-for or unknown explanatory factors such as biotic interactions, migrations, and historical factors;
>
> ii) eliminate the random noise inherent in real biological data, and thus avoid producing overfitted models plagued by the classification errors present in real presence-absence data;
>
> iii) provide the basis for calculating true model predictive performance by comparing modeled and virtual distributions.

We mapped the distribution of the virtual species in the European region (-13^{o} to 35^{o} longitude and 34^{o} to 72^{o} latitude, Figure 1) using a spatial resolution of 0.04 degrees. The total extent of the region studied was 6,576.4 km^{2} (510,514 0.04^{o} × 0.04^{o} cells). Four environmental variables (total annual precipitation, summer precipitation, mean maximum temperature, and mean minimum temperature) were extracted from the WORLDCLIM interpolated map database (version 1.3; see http://biogeo.berkeley.edu/worldclim/worldclim.htm). These variables were Box-Cox normalized and standardized to 0 mean and 1 standard deviation, eliminating measurement-scale effects. A Principal Component Analysis (PCA) was performed to obtain two reduced non-correlated environmental factors explaining 92.6% of the environmental variation across Europe, related to temperature (Factor 1) and precipitation (Factor 2). The choice of using the factors of a PCA both to describe the environmental variations in Europe and predict the species distribution (see below) was based in the aim stated above of avoiding any effect apart from prevalence or sample size that might affect the performance of the models. To achieve this, it is necessary to work with uncorrelated variables, in order to avoid the problems of collinearity in stepwise selection procedures (Harrell 2001).

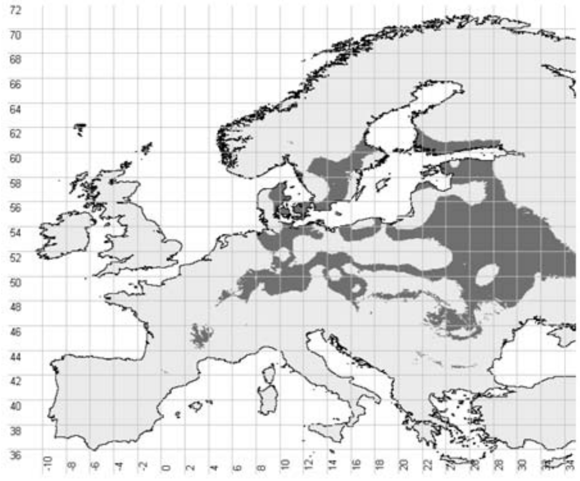


Figure 1. Spatial distribution of the virtual species in Europe. Dark grey areas are presences, and light grey absences.

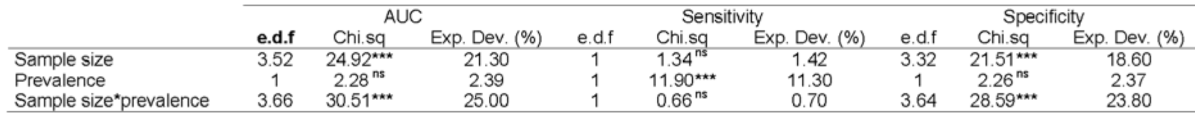
The distribution of the virtual species was assumed to be shaped only by these two factors. Therefore, the geographic range of the species was built using only these two variables, so that no unknown factors affected it (i.e., the species is in equilibrium with the environmental variables). To do this, the environmental range inhabited by the species was set to the mean ± SD of each factor. All cells falling within these intervals for both factors were selected as the true distribution range of the virtual species in Europe (presences; *n*=91,144), while the remaining cells were considered as true absences (*n*=419,296, see Figure 1). All geographic analyses were done with Idrisi Kilimanjaro GIS software (Clark Labs 2003).

The modeling process

We simulated several sampling processes on the virtual distribution range generated above. To do this, we extracted presences randomly within the occurrence range of the virtual species, and absences outside its range. This ensures that no spurious effects due to false absences are included in our analyses, and hence that the measured effects can be attributed either to prevalence or sample size, or both. Nine sets of increasing numbers of presence plots (n = 91, 456, 911, 4,557, 9,114, 22,786, 45,572, 68,358, and 91,144) were selected from the species distribution using a random stratified procedure in a GIS environment (Idrisi Kilimanjaro, Clark Labs 2003). Nine sets of absences of the same sizes as those of presences were also randomly selected. All possible combinations of presences and absences were combined into presence/absence datasets (81 datasets, with *n* ranging from 182 to 182,288, and prevalence ranging from 0.001 to 0.999). Fourteen additional datasets were used in order to detect a possible interaction with sample size (*n*=20 and 50, with 7 prevalence classes each: 0.9, 0.75, 0.6, 0.5, 0.4, 0.25, and 0.1). Thus, a total of 95 datasets were modeled, varying both in the number of observations (from 20 to 182,288) and in the prevalence or proportion of presences (from 0.001 to 0.999). Since we were mainly interested in detecting the effect of prevalence, most of our samples were of a great sample size in order to avoid this possible confounding factor.

Species distribution models were based on the same environmental variables used to build the distribution of the virtual species (i.e., the two environmental factors, see above). Therefore, all variables entering the models are truly explanatory, and no potentially explanatory factor is missing. All models were built using logistic regression (Generalized Linear Models with binomial distribution and logit-link function; McCullagh and Nelder 1989), which is commonly used to develop models from existing records of species distribution (Guisan et al. 2002, Lehmann et al. 2002, Reineking

Table 1. Simple regressions showing the effect of sample size, prevalence, and their interaction factor on prediction accuracy scores estimated by three statistics (columns; e.d.f, estimated degrees of freedom; Exp. Dev., explained deviance) (^{ns} not significant; * p < 0.05; ** p < 0.01; *** p < 0.001) tested using penalized regression splines with 5 initial degrees of freedom (Wood and Augustin 2002).



	AUC			Sensitivity			Specificity		
	e.d.f	Chi.sq	Exp. Dev. (%)	e.d.f	Chi.sq	Exp. Dev. (%)	e.d.f	Chi.sq	Exp. Dev. (%)
Sample size	3.52	24.92***	21.30	1	1.34^{ns}	1.42	3.32	21.51***	18.60
Prevalence	1	2.28^{ns}	2.39	1	11.90***	11.30	1	2.26^{ns}	2.37
Sample size*prevalence	3.66	30.51***	25.00	1	0.66^{ns}	0.70	3.64	28.59***	23.80

and Schröder 2003). The linear, quadratic, and cubic functions of the two environmental factors, together with their interaction, were used as explanatory variables; variables were selected by a backward stepwise procedure (Harrell 2001), eliminating the non-significant terms (p<0.05) from the model. All model building calculations were done in STATISTICA (StatSoft 2001).

Validation

The obtained model distributions were projected onto the whole European territory. The accuracy measures used to compare true and predicted maps were derived from a confusion matrix (i.e., a cross-tabulated matrix of the number of observed presence and absence cases against the predicted presences and absences; Fielding and Bell 1997). First of all, a cut-off was established for the logistic predictions, and all cases with predicted values higher than that threshold were accepted as predicted presences. To do this, we calculated specificity (ratio of correctly predicted absences to the total number of absences) and sensitivity (ratio of correctly predicted presences to their total number) from the training data over a range of 100 thresholds, and selected the cut-off which minimized their difference for each one of the 95 models. Such a criterion yields better results than others widely used, as it accounts for prevalence (Liu et al. 2005, Jiménez-Valverde and Lobo 2007b). The confusion matrix was set up after applying the threshold criterion (calculated with the training data) to the model probabilities, and predicted and virtual maps were compared (note that the entire study area is used for validation, i.e., including the training points, and that prevalence remains constant) by calculating sensitivity, specificity, and the area under the Receiver operating characteristic Curve (AUC). All three measures are independent of prevalence (McPherson et al. 2004, Allouche et al. 2006). The Receiver operating characteristic (ROC) curve is widely used as a threshold-independent accuracy measure (Zweig and Campbell 1993, Fielding and Bell 1997), as is commonly accepted as the best to assess model accuracy (Fielding 2002, but see Lobo et al. 2008). Here, sensitivity is plotted against 1–specificity over a number of thresholds (100 in this study), and the area under the curve (AUC) calculated. AUC ranges from 0 to 1; values under 0.5 indicate discrimination worse than chance, 0.5 implies no discrimination (i.e., random predictions), and 1.0 indicates perfect discrimination.

Validation was done using the entire distribution of the species, which was completely known, contrary to the situation in real-world species distribution modeling. In real situations, modelers try to predict a distribution that is always unknown. Then, samples of reality are used for training and validating models, and the best option for validation is to use samples that are as independent from the training dataset as possible. We, on the contrary, know completely the distribution range of the virtual species, thus being able to compare model predictions with its whole distribution. To ensure that we did not add any spurious effect due to such differences with real-world validation datasets, we also calculated sensitivity and specificity for the data not used in the training process. The correlations between sensitivities and specificities of the whole dataset and the independent dataset were 0.97 and 1.00 (p < 0.01), respectively, showing that the results would not have been different if validation was done on an independent dataset.

Testing for prevalence and sample size effects

We drew scatterplots to examine the effect of prevalence and sample size on AUC, sensitivity, and specificity, and calculated penalized regression splines with 5 initial degrees of freedom to estimate the variation explained by each factor (Wood and Augustin 2002). We also tested for significance of the interactions between the studied effects. We fitted the splines in R (R Development Core Team 2006) using the mgcv package (Wood 2004). We estimated the break-points in the scatterplots by fitting regression models with segmented relationships between the dependent (accuracy measures) and independent variables (Muggeo 2003), fitting segmented regressions in R using the segmented package (Muggeo 2004).

Results

AUC values were quite high in almost all cases (mean ± SD; 0.961 ± 0.050), being higher than 0.90 in 87 out of the 95 models, and higher than 0.80 in 6 of the remaining models (Figure 2). Only two models showed AUC values under 0.80 (0.689 and 0.706), corresponding to the cases with smaller sample size and higher prevalence (n=20 and 50; prevalence=0.9). Variations in AUC were significantly related to both sample size and the interaction between sample size and prevalence, accounting for 21.3% and 25.0% of the variability, respectively (Table 1). AUC values were consistently high until very low sample sizes were reached; segmented regressions yielded a break-point of 72.4 (Figure 2). With 20 observations, the higher the prevalence the lower the AUC value, while with 50 observations the effect of prevalence disappeared at prevalence values lower than 0.75 (see Figure 2). When both sample size and prevalence were included together in a model in order to explain AUC variation, only the interaction term was significant.

Sensitivity values were also high and stable, with percentages of success always higher than 80% (0.929 ± 0.024) (Figure 2). These slight variations in sensitivity were significantly related to prevalence, accounting for 11.3% of variability (Table 1); the higher the prevalence, the higher the sensitivity (Figure 2). If low sample size cases (*n*=20 and 50) were omitted, sensitivity showed a slight increment at high prevalence values (break-point=0.99) and a decrease at low values (break-point < 0.01).

In general, specificity values were also high (0.893 ± 0.103). Specificity values were highly correlated with AUC values ($r = 0.99$, $p < 0.05$), so the pattern of variation with sample size and prevalence was similar for both accuracy measures (Figure 2). The break-point for the relationship with sample size was estimated as 65.44. Variations in specificity were significantly related to both sample size and the

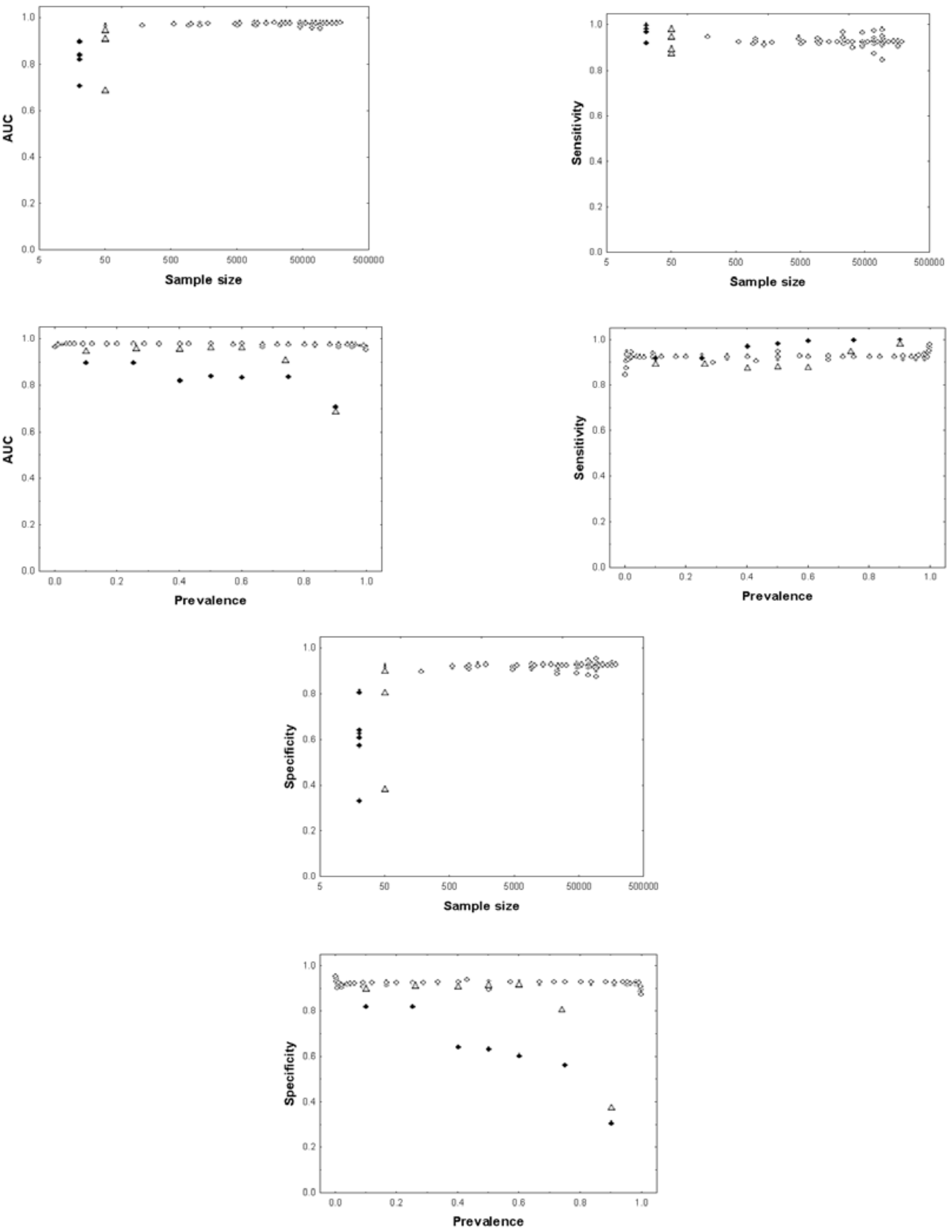


Figure 2. Relationships between the three accuracy measures (AUC, sensitivity, and specificity) and sample size and prevalence (triangles, cases with *n*=50; black dots, cases with *n*=20). Note that sample size axes are in logarithmic scale.

interaction between sample size and prevalence, accounting for 18.6% and 23.8% of variability, respectively (see Table 1). Samples with 20 observations yielded the lowest specificity values, which negatively correlated with prevalence. With 50 observations, specificity was negatively affected at prevalences higher than 0.75 (Figure 2). Again, when both sample size and prevalence were included together in a model in order to explain specificity variation, only the interaction term remained significant. Specificity was also negatively correlated with sensitivity ($r = -0.66$, $p < 0.05$). If low sample size cases (n = 20 and 50) were omitted, specificity showed a slight increment at low prevalence values (break-point < 0.01) and a decrease at high values (break-point = 0.99).

Discussion

In general, quite accurate predictions can be obtained from a wide range of sample sizes and prevalences. According to AUC values, the predictions of 87 models were highly accurate, while six were qualified as good and useful models (following Swets 1988). Using this criterion, only the two models with 20 and 50 observations and prevalences of 0.9 could be considered as having a poor discrimination capacity. Sensitivity values were always higher than 80%, so presences were relatively well-predicted in all cases. Specificity values were higher than 80% in most cases, except at a sample size of 20 observations and prevalences higher than 0.25, as well as at a sample size of 50 observations and prevalences higher than 0.6. In these cases, there was a substantial overprediction.

Prevalence effects

In the absence of noise, the effect of prevalence by itself is not important (except in the small sample sizes of 20 and 50 data points). Only at extreme prevalence values (lower than 0.01 and higher than 0.99), sensitivity and specificity were affected, reducing their values at very low and high values, respectively; even in those cases, models would be considered as good, with sensitivity and specificity values higher than 0.80 and AUC scores higher than 0.90 (see Figure 2). In any case, increasing sensitivity implies a reduction in specificity and *vice versa*. Cramer (1999) stated that there is no reason for the rarest events to be badly predicted. Prevalence affects the tests of model performance of the logistic regressions due to the mean probability biases (Fielding and Bell 1997, Manel et al. 1999, Olden et al. 2002). Therefore, the use of an appropriate cut-off to convert the probability map into a presence/absence map (Cramer 1999, Liu et al. 1995, Jiménez-Valverde and Lobo 2007b) is central to avoiding the drawbacks that could be associated with prevalence. If the frequency of occurrence is accounted for in the selection of an appropriate cut-off threshold, our results show that prevalence does not have a great impact on model reliability. This is true except for cases of extremely low and high prevalence values; there, biases in the estimation of the parameters (King and Zeng 2001) may have implications in the predictive performance. Values of 0.01 and 0.99 seem to be the lowest and highest workable thresholds; prevalences within these two extreme values will not present negative effects because of being unbalanced samples (see also Dixon et al. 2005).

Sample size effects

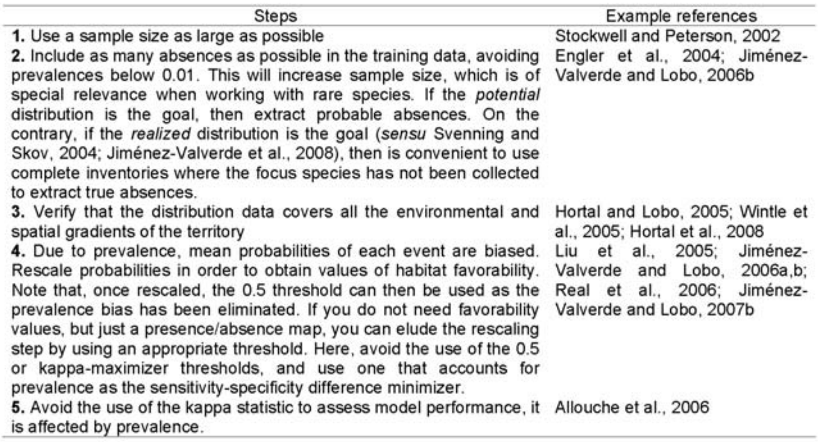
It is well-known that sample size influences the results of species distribution models (Wisz et al. 2008) and different minimum sample sizes have been suggested in the species distribution modeling literature (Pearce and Ferrier 2000, Stockwell and Peterson 2002, McPherson et al. 2004). Moreover, there is no established rule to decide the minimum sample sizes that can be used for logistic regression (Peng et al. 2002). In the absence of a general rule, it is generally assumed that the greater the sample size, the more accurate the model (Cumming 2000, Olden and Jackson 2000, McPherson et al. 2004, Martínez-Meyer 2005, Reese et al. 2005). Our results seem to show that quite small sample sizes significantly reduce the reliability of model predictions, even in the absence of noise. In our extremely simplified modeling approach, once sample size reached a value of around 70, model reliability became independent of sample size.

The interaction between prevalence and sample size

There is a strong interaction between prevalence and sample size; at small sample sizes, the higher the prevalence, the more the distribution of the species is overpredicted. Thus, it is possible to obtain moderately accurate models with sample sizes even as small as 20 data points, provided that the number of absences in the training data is higher than the number of presences. Our virtual species can be considered "central", i.e., a species with its optimum at intermediate levels of each environmental factor. Thus, while presences are extremely climate dependent, absences can be found in a greater variety of environmental conditions. Therefore, it is likely that in our particular example the number of absences is not enough to restrict model predictions at small sample sizes and high prevalence scores. High (or low) prevalences would not have any effect in restricted (or widespread) species if the quantity of absence (or presence) points was large enough to cover all the environmental variation.

In sum, unbalanced samples are not the source of any problem; rather, the sample size of each event might cause the problems previously attributed to prevalence. These results support the suggestions of Jiménez-Valverde and Lobo (2006a) about the importance of the sample size of each event (presences or absences) in order to represent the environmental gradients, regardless of their relative size. Similar results were obtained by Coudun and Gégout (2006) using virtual species with Gaussian responses to environmental gradients. It seems evident that, most times, the probability of sampling spatial and environmental variation accurately increases with increasing sample size. Thus, the lower the sample size, the greater the relevance of well-designed sampling protocols designed to select training points across the whole spatial and environmental gradient (Wessels et al.

Table 2. Protocol for species distribution modeling, including recommendations to avoid the problems related to prevalence, and to take advantage of it in the modeling process if needed.



Steps	Example references
1. Use a sample size as large as possible	Stockwell and Peterson, 2002
2. Include as many absences as possible in the training data, avoiding prevalences below 0.01. This will increase sample size, which is of special relevance when working with rare species. If the *potential* distribution is the goal, then extract probable absences. On the contrary, if the *realized* distribution is the goal (*sensu* Svenning and Skov, 2004; Jiménez-Valverde et al., 2008), then is convenient to use complete inventories where the focus species has not been collected to extract true absences.	Engler et al., 2004; Jiménez-Valverde and Lobo, 2006b
3. Verify that the distribution data covers all the environmental and spatial gradients of the territory	Hortal and Lobo, 2005; Wintle et al., 2005; Hortal et al., 2008
4. Due to prevalence, mean probabilities of each event are biased. Rescale probabilities in order to obtain values of habitat favorability. Note that, once rescaled, the 0.5 threshold can then be used as the prevalence bias has been eliminated. If you do not need favorability values, but just a presence/absence map, you can elude the rescaling step by using an appropriate threshold. Here, avoid the use of the 0.5 or kappa-maximizer thresholds, and use one that accounts for prevalence as the sensitivity-specificity difference minimizer.	Liu et al., 2005; Jiménez-Valverde and Lobo, 2006a,b; Real et al., 2006; Jiménez-Valverde and Lobo, 2007b
5. Avoid the use of the kappa statistic to assess model performance, it is affected by prevalence.	Allouche et al., 2006

1998, Jiménez-Valverde and Lobo 2004, Kadmon et al. 2004, Hortal and Lobo 2005, Funk et al. 2005).

In our study, the lack of spatial bias in presences and absences is guaranteed by a random stratified survey. This is commonly assumed but often unverified when using real biodiversity data. However, spatial and environmental bias could be the rule rather than the exception (see Hortal et al. 2007, 2008, Lobo et al. 2007), compromising the reliability of model results (Jiménez-Valverde et al. 2008b). The undesirable effects of sample size and its interaction with prevalence can be promoted by spatial aggregation in the training data; the larger the spatial bias, the greater the sample size should be in order to represent the environmental gradient. Additionally, the lower the number of data points in the training data, the more unstable the models are likely to be. Instability will arise because small perturbations in the training data will induce models to select different predictors. In real situations, where variables are often collinear and a great number of unknown factors control the distribution of the species, models are more prone to instability. All these facts are among the possible explanations for the differences in the minimum sample size required to build optimal models reported in the literature (Pearce and Ferrier 2000, Stockwell and Peterson 2002, McPherson et al. 2004). Other noise sources, such as unaccounted-for environmental variables or even simple stochasticity, make real data even more difficult to model.

The potentially confounding factors in species distribution modeling are usually unknown. However, they may be omnipresent in the real world, and are also likely to be specific for each modeling exercise. Therefore, no simple rule can be given to decide the optimal sample size; the best strategy is to gather sample sizes as big as possible (but see Stockwell and Peterson 2002). However, this is not always affordable. In fact, sample sizes in the interval of uncertainty (N < 500; Long 1997) are more the rule than the exception. When working with presence data extracted from bibliography or biodiversity databases, there are several procedures to determine sites that are likely to correspond to true absences. This allows including a great number of reliable absences in the training data, improving the likelihood of discriminating the most adequate predictors during the modeling process (see, for example, Zaniewski et al. 2002, Engler et al. 2004, Lobo et al. 2006, Jiménez-Valverde and Lobo 2006b, Jiménez-Valverde et al. 2007). Whenever possible, reliable absences can also be obtained from places with well-known inventories (see, e.g., Hortal et al. 2004, 2007) where the absence of a non-reported species can be determined with higher certainty.

It could be argued that the position of our virtual species along the PCA axes (centred in the means) could be determining the results, and that species with different responses to the environment (i.e., more marginal) would show different results. However, the design of our virtual experiment consists in sampling fixed numbers of presences and absences, and combining them to change the prevalence of the training data. Species with responses to the PCA axis shifted from the mean would be sampled in the same way, including absences from all outside the virtual species range. Hence, no significant change on the results would be expected, because the geographical distribution of the species and its environmental response (including the unsuitable areas) would be equally sampled no matter its position in the environmental axes. As explained before, effects in the results could be expected only if the virtual species is so marginal that the number of absences needs to be big enough to account for the unsuitable environmental conditions. However, several authors have shown that the inclusion of absences that are much away (in the environmental space) from the presences (the so called "naughty noughts", *sensu* Austin and Meyers 1996) do not improve the models (Austin and Meyers 1996, Thuiller et el. 2004, J. M. Lobo, A. Jiménez-Valverde and J. Hortal, submitted), so we do not believe that the position of the species along the PCA axes would significantly change our results. On the other hand, the effects on model results of a more realistic sampling strategy , i.e., reflecting the spatial and environmental biases that characterize most of the data available (see above), would be due to a lack of sampling of the full environmental response of the species, rather to effects related to prevalence. Our apparently unrealistic study design and the simplicity of the virtual species are necessary to tear apart any potentially confounding factor other than the one we study, the pure effect of prevalence.

Concluding remarks

Our results demonstrate that biased prevalences in training data are unimportant to obtain accurate results, except with extremely unbalanced samples. Our results are consistent with the proposals of Jiménez-Valverde and Lobo (2006a) about the low relevance of prevalence on the performance of predictive models, and the implication of other confounding factors associated which usually are correlated with prevalence. The sample size of each event (presences or absences) and the completeness of the representation of the environmental gradient provided by the training data arise as more important factors affecting the likelihood of obtaining reliable models. The strategy of resampling the data available to obtain training data with prevalences of 0.5 must be discarded, since it would yield only a loss of information, that can be especially relevant when rare species are the focus of the research. In addition, training data should be tested for spatial and environmental biases (see, e.g., Wintle et al. 2005, Hortal et al. 2008). Although the number of presences cannot be increased, the number of absences can be increased by identifying sites that are likely to be true absences. These techniques are of special importance to model the potential distribution of rare species, which generally occupy a small proportion of a study's spatial extent, so absences are necessary to include the restrictions within model predictions. In Table 2, we list the basic steps needed to avoid the biases due to prevalence and to take advantage of it in the modeling process, if needed.

Although they are not directly based in a real case study, understanding our results is essential to ascertain the true effects of prevalence, sample size and their interaction in the more complex universe of the real world. Noise is an apparently random signal in the data that cannot be modelled because the factors that cause it are unknown, or even impossible to include within the predictors. In this sense, noise in distributional data is quite common and can have multiple sources. Its effects depend on the amount of noise, its spatial structure, the different allocation in either presences or absences, its interaction with scale, etc. In order to understand these interactions, we first need to understand the effects of sample size and prevalence in the absence noise, in order to depict the particular effects of the latter. So, further investigation is desirable to understand the effects of data-driven sources of error in species distribution modeling, as well as their interactions with the sample size of each event (presence and absence). The virtual experiment approach applied in this work could help to estimate the effect of these error sources.

Acknowledgements. This paper was 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). AJ-V was supported by a MEC (Ministerio de Educación y Ciencia, Spain) postdoctoral fellowship (Ref.: EX-2007-0381). JH was supported by the U.K. Natural Environment Research Council.

References

Allouche, O., A. Tsoar and R. Kadmon. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *J. Appl. Ecol.* 43: 1223-1232.

Anderson, R. P., M. Gómez-Laverde and A. T. Peterson. 2002a. Geographical distributions of spiny pocket mice in South America: insights from predictive models. *Global Ecol. Biogeogr.* 11**:** 131-141.

Anderson, R. P., A. T. Peterson and M. Gómez-Laverde. 2002b. Using niche-based GIS modeling to test geographic predictions of competitive exclusion and competitive release in South American pocket mice. *Oikos* 98: 3-16.

Austin, M.P. and J.A. Meyers. 1996. Current approaches to modeling the environmental niche of eucalyptus: implications for the management of forest biodiversity. *For. Ecol. Manage*. 85: 95-106.

Austin, M.P., L. Belbin, J.A. Meyers, M.D. Doherty and M. Luoto. 2006. Evaluation of statistical models used for predicting plant species distributions: Role of artificial data and theory. *Ecol. Model.* 199: 197-216.

Barbosa, A. M., R. Real, J. Olivero and J. M. Vargas. 2003. Otter (*Lutra lutra*) distribution modeling at two resolution scales suited to conservation planning in the Iberian Peninsula. *Biol. Conserv.* 114**:** 377-387.

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

Chefaoui, R. M., J. Hortal and J. M. Lobo. 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. 2003. *Idrisi Kilimanjaro. GIS software package.* Clark Labs, Worcester, MA.

Coudun, C. and J.-C. Gégout. 2006. The derivation of species response curves with Gaussian logistic regression is sensitive to sampling intensity and curve characteristics. *Ecol. Model.* 199: 164-175.

Cramer, J. S. 1999. Predictive performance of binary logit model in unbalanced samples. *J. Royal Statistical Soc., Series D* 48: 85-94.

Cumming, G. S. 2000. Using between-model comparisons to fine-tune linear models of species ranges. *J. Biogeogr.* 27: 441-455.

Dixon, P. M., A. M. Ellison and N. J. Gotelli. 2005. Improving the precision of estimates of the frequency of rare events. *Ecology* 85: 1114-1123.

Engler, R., A. Guisan and L. Rechsteiner. 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.

Estrada, A., R. Real and J.M. Vargas. 2008. Using crisp and fuzzy modelling to identify favourability hotspots useful to perform gap analysis. *Biodivers. Conserv*. 17: 857-871.

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

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

Funk, V.A., K.S. Richardson and S. Ferrier. 2005. Survey-gap analysis in expeditionary research: where do we go from here? *Biol. J. Linn. Soc.* 85: 549-567.

Guisan, A. and N. E. Zimmermann. 2000. Predictive habitat distribution models in ecology. *Ecol. Model.* 135: 147-186.

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

Harrell, F. E. J. 2001. *Regression Modelling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis.* Springer, New York.

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

Hirzel, A.H. and G. Le Lay. 2008. Habitat suitability and niche theory. *J. Appl. Ecol.* 45: 1371-1381.

Hortal, J. and J. M. Lobo. 2005. An ED-based protocol for optimal sampling of biodiversity. *Biodiv. Conserv.* 14: 2913-2947.

Hortal, J., P. Garcia-Pereira and E. García-Barros. 2004. Butterfly species richness in mainland Portugal: Predictive models of geographic distribution patterns. *Ecography* 27: 68-82.

Hortal, J., J.M. Lobo and A. Jiménez-Valverde. 2007. Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife (Canary Islands). *Conserv. Biol.* 21: 853-863.

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

Hosmer D. W. and S. Lemeshow. 1989. *Applied Logistic Regression.* Wiley, New York.

Jiménez-Valverde, A. and J. M. Lobo. 2004. Un método sencillo para seleccionar puntos de muestreo con el objetivo de inventariar taxones hiperdiversos: el caso práctico de las familias *Araneidae* y *Thomisidae* (*Araneae*) en la Comunidad de Madrid, España. *Ecología* 18: 297-308.

Jiménez-Valverde, A. and J. M. Lobo. 2006a. The ghost of unbalanced species distribution data in geographic model predictions. *Divers. Distrib.* 12: 521-524.

Jiménez-Valverde, A. and J. M. Lobo. 2006b. Distribution determinants of endangered Iberian spider *Macrothele calpeiana* (Araneae, Hexathelidae). *Environ. Entomol.* 35: 1491-1499.

Jiménez-Valverde, A. and J. M. Lobo. 2007a. Potential distribution of the endangered spider *Macrothele calpeiana* (Walckenaer, 1805) (Araneae, Hexathelidae) and the impact of climate warming. *Acta Zool. Sin.* 53: 865-876.

Jiménez-Valverde, A. and J. M. Lobo. 2007b. Threshold criteria for conversion of probability of species presence to either-or presence-absence. *Acta Oecol.* 31: 361-369.

Jiménez-Valverde, A., V. M. Ortuño and J. M. Lobo. 2007. Exploring the distribution of *Sterocorax* Ortuño, 1990 (Coleoptera, Carabidae) species in the Iberian Peninsula. *J. Biogeogr.* 34: 1426-1438.

Jiménez-Valverde, A., J.F. Gómez, J.M. Lobo, A. Baselga and J. Hortal. 2008a. Challenging species distribution models: the case of *Maculinea nausithous* in the Iberian Peninsula. *Ann. Zool. Fenn.* 45: 200-210.

Jiménez-Valverde, A., J. M. Lobo and J. Hortal. 2008b. Not as good as they seem: the importance of concepts in species distribution modelling. *Divers. Distrib.* 14: 885-890.

King, G. and L. Zeng. 2001. Logistic regression in rare events data. *Political Analysis* 9: 137-163.

Kadmon, R., O. Farber and A. Danin. 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecol. Appl.* 14: 401-413.

Lehmann, A., J. M. Overton and M. P. Austin. 2002. Regression models for spatial prediction: their role for biodiversity and conservation. *Biodiv. Conserv.* 11: 2085-2092.

Liu, C., P. M. Berry, T. P. Dawson and R. G. Pearson. 2005. Selecting thresholds of occurrence in the prediction of species distributions. *Ecography* 28: 385-393.

Lobo, J. M., J. R. Verdú and C. Numa. 2006. Environmental and geographical factors affecting the Iberian distribution of flightless *Jekelius* species (Coleoptera: Geotrupidae). *Divers. Distrib.* 12: 179-188.

Lobo, J.M., A. Baselga, J. Hortal, A. Jiménez-Valverde and J. F. Gómez. 2007. How does the knowledge about the spatial distribution of Iberian dung beetle species accumulate over time? *Divers. Distrib.* 13: 772-780.

Lobo, J.M., A. Jimenez-Valverde and R. Real. 2008. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecol. Biogeogr.* 17: 145-151.

Long, J. S. 1997. *Regression Models for Categorical and Limited Dependent Variables.* Sage Publications, Thousand Oaks, CA.

Luoto, M., J. Poyry, R. K. Heikkinen and K. Saarinen. 2005. Uncertainty of bioclimate envelope models based on the geographical distribution of species. *Global Ecol. Biogeogr.* 14: 575-584.

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

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

Martínez-Meyer, E. 2005. Climate change and biodiversity: some considerations in forecasting shifts in species´ potential distributions. *Biodiv. Informatics* 2: 42-55.

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

McPherson, J. M., W. Jetz and D. J. Rogers. 2004. The effects of species´ range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *J. Appl. Ecol.* 41: 811-823.

Meynard, C. N. and J. F. Quinn. 2007. Predicting species distributions: a critical comparison of the most common statistical models using artificial species. *J. Biogeogr.* 34: 1455–1469.

Muggeo, V. M. R.. 2003. Estimating regression models with unknown break-points. *Stat. Med.* 22: 3055–3071.

Muggeo, V. M. R. 2004. segmented: segmented relationships in regression models. R package version 0.1-4.

Nogués-Bravo, D., J. Rodríguez, J. Hortal, P. Batra and M. B. Araújo. 2008. Climate change, humans and the extinction of the woolly mammoth. *PLoS Biol.* 6: e79.

Olden, J. D. and D. A. Jackson. 2000. Torturing data for the sake of generality: How valid are our regression models? *Écoscience* 7: 501-510.

Olden, J. D., D. A. Jackson and P. R. Peres-Neto. 2002. Predictive models of fish species distributions: A note on proper validation and chance predictions. *Trans. Am. Fish. Soc.* 131: 329-336.

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

Pearce, J. and S. Ferrier. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. *Ecol. Model.* 133: 225-245.

Peng, C.-Y. J., K. L. Lee and G. M. Ingersoll. 2002. An introduction to logistic regression analysis and reporting. *J. Educational Res.* 96: 3-14.

Peterson, A. T. and R. D. Holt. 2003. Niche differentiation in Mexican birds: using point occurrences to detect ecological innovation. *Ecol. Lett.* 6: 774-782.

Peterson, A. T., J. Soberón and V. Sánchez-Cordero. 1999. Conservatism of ecological niches in evolutionary time. *Science* 285: 1265-1267.

R Development Core Team. 2006. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org

Real, R., A. M. Barbosa and J. M. Vargas. 2006. Obtaining environmental favourability functions from logistic regression. *Environ. Ecol. Stat.* 13: 237-245.

Reese, G. C., K. R. Wilson, J. A. Hoeting and C. H. Flather. 2005. Factors affecting species distribution predictions: a simulation modeling experiment. *Ecol. Appl.* 15: 554-564.

Reineking, B. and B. Schröder. 2003. Computer-intensive methods in the analysis of species-habitat relationships. In: H. Reuter, B. Breckling and A. Mittwollen (eds), *GfÖ Arbeitskreis Theorie in der Ökologie.* P. Lang Verlag, Frankfurt, pp. 100-117.

Rushton, S. P., S. J. Ormerod and G. Kerby. 2004. New paradigms for modelling species distributions? *J. Appl. Ecol.* 41: 193-200.

Schadt, S., E. Revilla, T. Wiegand, F. Knauer, P. Kaczensky, U. Breitenmoser, L. Bufka, J. Červený, P. Koubek, T. Huber, C. Staniša and L. Trepl. 2002. Assessing the suitability of central European landscapes for the reintroduction of Eurasian lynx. *J. Appl. Ecol.* 39: 189-203.

Scott, J. M., P. J. Heglund, J. B. Haufler, M. Morrison, M. G. Raphael, W. B. Wall and F. Samson (eds.). 2002. *Predicting Species Occurrences. Issues of Accuracy and Scale*. Island Press, Covelo, CA.

Segurado, P. and M. B. Araújo. 2004. An evaluation of methods for modelling species distributions. *J. Biogeogr.* 31: 1555-1568.

Seoane, J., J. H. Justribó, F. García, J. Retamar, C. Rabadán and J. C. Atienza. 2006. Habitat-suitability modelling to assess the effects of land-use changes on Dupont´s lark *Chersophilus duponti*: A case study in the Layna Important Bird Area. *Biol. Conserv.* 128: 241-252.

StatSoft. 2001. *STATISTICA (data analysis software system and user´s manual)*, version 6. StatSoft, Inc., Tulsa, OK.

Stockwell, D. R. B. and A. T. Peterson. 2002. Effects of sample size on accuracy of species distribution models. *Ecol. Model.* 148: 1-13.

Svenning, J.C. and F. Skov. 2004. Limited filling of the potential range in European tree species. *Ecol. Lett.* 7: 565-573.

Swets, J. A. 1988. Measuring the accuracy of diagnostic systems. *Science* 240: 1285-1293.

Thuiller, W., L. Brotons, M.B. Araújo and S. Lavorel. 2004 Effects of restricting environmental range of data to project current and future species distributions. *Ecography* 27: 165-172.

Vaughan, I. P. and S. J. Ormerod. 2003. Improving the quality of distribution models for conservation by addressing shortcomings in the field collection of training data. *Conserv. Biol.* 17: 1601-1611.

Wessels, K. J., A. S. Van Jaarsveld, J. D. Grimbeek and M. J. Van der Linde. 1998. An evaluation of the gradsect biological survey method. *Biodiv. Conserv.* 7: 1093-1121.

Whittaker, R. J., M. B. Araújo, P. Jepson, R. J. Ladle, J. E. M. Watson and K. J. Willis. 2005. Conservation Biogeography: assessment and prospect. *Divers. Distrib.* 11: 3-23.

Wintle, B. A., J. Elith and J. M. Potts. 2005. Fauna habitat modelling and mapping: a review and case study in the Lower Hunter Central Coast region of NSW. *Austral Ecol.* 30: 719-738.

Wisz, M. S., R. J. Hijmans, J. Li, A. T. Peterson, C. H. Graham, A. Guisan and NCEAS Predicting Species Distributions Working Group. 2008. Effects of sample size on the performance of species distribution models. *Divers. Distrib.* 14: 763-773.

Wood, S. N. 2004. mgcv: GAMs with GCV smoothness estimation and GAMMs by REML/PQL. R package version 1.1-8.

Wood, S. N. and N. H. Augustin. 2002. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecol. Model.* 157: 157-177.

Zaniewski, A. E., A. Lehmann and J. M. Overton. 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecol. Model.* 157: 261-280.

Zweig, M. H. and G. Campbell. 1993. Receiver-operating characteristics (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.* 39: 561-577.

Received June 27, 2008
Revised March 9, 2009; June 23, 2009
Accepted September 4, 2009