Journal for Nature Conservation 19 (2011) 1–7

Contents lists available at ScienceDirect

Journal for Nature Conservation

journal homepage: www.elsevier.de/jnc

Exploring the effects of quantity and location of pseudo-absences and sampling biases on the performance of distribution models with limited point occurrence data

Jorge M. Lobo [a,*], Marcelo F. Tognelli [b,c]

[a] Departamento de Biodiversidad y Biología Evolutiva, Museo Nacional de Ciencias Naturales (CSIC), C/José Gutiérrez Abascal, 2, 28006, Madrid, Spain
[b] IADIZA – CRICYT, CC 507, CP 5500, Mendoza, Argentina
[c] IUCN, SSC Biodiversity Assessment Unit, Science & Knowledge, Conservation International, 2011 Crystal Drive, Suite 500, Arlington, VA 22202, USA

ARTICLE INFO

Article history:
Received 15 May 2009
Accepted 22 March 2010

Keywords:
Species distribution models
Number of pseudo-absences
Location of pseudo-absences
Spatial sampling bias

SUMMARY

In the last decade, the application of predictive models of species distribution in ecology, evolution, and conservation biology has increased dramatically. However, limited available data and the lack of reliable absence data have become a major challenge to overcome. At least two approaches have been proposed to generate pseudo-absences; however it is not clear how the number of pseudo-absences created affect model performance. Moreover, the spatial bias in the collecting localities of a species (presence data) may add extra noise to the final distribution model. Here, we use a virtual species to assess the effects of spatial sampling bias, and number and location of pseudo-absences on model accuracy. We found that both number of pseudo-absences and spatial bias in sampling localities, as well as their interaction, significantly influence all accuracy measures (AUC, sensitivity, and specificity). However, location of pseudo-absences (either generated across the entire study area or only outside the environmental envelope of the species) does not affect model performance. These results provide some methodological guidelines for developing reliable distribution hypotheses when presence data are scarce.

Introduction

Currently, the general lack of reliable distributional data for biogeographical and conservation purposes is overcome by the use of model predictions (Ferrier et al. 2002; Raxworthy et al. 2003; Williams et al. 2005). Unfortunately, the power of these methods is uncertain for modelling the distribution of organisms such as invertebrates, which represent the majority of biological diversity. For these organisms, we generally only have presence data (location points of known species occurrences) that are not evenly distributed across the environmental and spatial gradient of the study area, and we do not have absence data (location points of known species absence). Therefore, the question is whether we can obtain a relatively reliable predicted distribution with this limited information. Here, we explored the effects of sampling biases of presence data, and quantity and location of pseudo-absences on the performance of distribution models, proposing some methodological guidelines for developing reliable distribution hypotheses.

To overcome the frequent inexistence of absence information, several authors have proposed two main approaches for selecting pseudo-absence points in order to use powerful group discrimination techniques that need presence–absence data. The first approach entails including background absence points across all the study area (Stockwell & Peters 1999), or within the areas with environmental characteristics similar to those containing well-sampled data for the entire group (Ferrier & Watson 1997; Zaniewski et al. 2002). The second approach includes absence points outside the environmental domain favourable for the species (Engler et al. 2004; Lobo et al. 2006). The latter method employs a profile technique (i.e. environmental envelope, ecological niche factor analysis) using only presence data to firstly calculate a suitability map for a species. With this map, pseudo-absence points are selected outside the environmental space obtained from the observed presence points of the species. Then, both presence data and pseudo-absence data are included as a binomial dependent variable in one of the available group discrimination methods that use environmental predictors (i.e. GLM or GAM; see Guisan & Zimmermann 2000) to model the distribution.

Independent of the approach used to select pseudo-absence points, there is disagreement on the number of absence data that should be used in modelling techniques. Many authors state that prevalence (i.e. the ratio of number of presences to total data used to build the model) highly influences model accuracy, although some authors disagree over its effect (Fielding & Bell 1997; Manel et al.

* Corresponding author. Tel.: +34 91 4111328; fax: +34 91 5645078.
E-mail address: mcnj117@mncn.csic.es (J.M. Lobo).

1617-1381/$ – see front matter © 2010 Elsevier GmbH. All rights reserved.
doi:10.1016/j.jnc.2010.03.002

1999; Olden et al. 2002; Vaughan & Ormerod 2003; McPherson et al. 2004; Luoto et al. 2005). The main consequence of using many pseudo-absence points (a low prevalence) is that probability values derived from predictive functions are unavoidably biased toward the highest number of absence data used (Hosmer & Lemeshow 2000; Cramer 1999). To assign these low probability scores to a true presence point, one must use an appropriate cut-off value to convert decimal fraction probabilities to a binary variable (Liu et al. 2005; Jiménez-Valverde & Lobo 2006; Jiménez-Valverde & Lobo 2007), or re-scale logistic probabilities by applying a favourability function (Real et al. 2006). Thus, although a high number of pseudo-absences affect the parameters of the models, their final reliability can remain unchanged if a correct cut-off value is applied (Jiménez-Valverde & Lobo 2006). However, because dependent variables with thousands of times more zeroes than ones can underestimate and produce imprecise probabilities for the most rare-event data (King & Zeng 2001; Dixon et al. 2005), we need to understand how the number of selected pseudo-absence points affects the performance of our models (Jiménez-Valverde et al. 2009).

Species location data collected at different times generally results in different distribution maps for that species (Lobo et al. 2007). These different distributions are the result of, among other reasons, an increase in distributional information over time, both in a random and in an environmentally structured fashion. When the distributional information increases at random the "true" distribution of a species would be uniformly and gradually revealed across its entire range. Alternatively, if species data is incorporated over time in an environmentally structured way, an expansion of the range resulting from sociological, environmental, or sampling effort bias may be revealed. (Dennis & Hardy 1999; Dennis et al. 1999; Dennis & Thomas 2000; Dennis et al. 2006; Zaniewski et al. 2002; Anderson 2003; Reutter et al. 2003; Graham et al. 2004; Soberón & Peterson 2004; Martínez-Meyer 2005). It is widely recognised that, to produce reliable distribution models, presences and absences need to be well-distributed across the entire environmental and geographical gradient of the study area (Kadmon et al. 2004; Hortal & Lobo 2005; Reese et al. 2005; Soberón & Peterson 2005; Vaughan & Ormerod 2005; Hortal et al. 2007; Hortal et al. 2008; Lobo et al. 2010), but it is not clear how biases in the distribution of presence points influence the accuracy of distribution models.

Several issues in distributional modelling techniques need to be resolved for these models to become widely applicable (Soberón & Peterson 2005; Austin 2006). Moreover, the conditions worsen when modelling invertebrate distributions, because of the poor quality and scarce presence data, and a usually high proportion of false absences. What is the accuracy of presence–absence model predictions when only few presence points are available? We try to answer this question using a simple virtual species for which both the complete distribution and the explanatory variables are known in advance. Model distribution results are mainly conditioned by three sources of uncertainty: the quality of the dependent variable; the predictive capacity of the selected explanatory variables; and the modelling technique used to estimate the parameters of the function. The use of a virtual or artificial species allowed us: (i) to avoid the bias resulting from contingent or unknown explanatory factors; (ii) to eliminate the random noise inherent in real biological information; and (iii) to calculate true model accuracy by comparing modelled and virtual distributions. Hence, by using an artificially constructed virtual species range we control for the effects of the predictors and the biases in survey data, and are able to obtain an exact measure of the performance of the models (instead of estimating it), being able to explore how the quality of the used biological data can influence model results. Using only ten presence points selected randomly or in a spatially biased manner from the whole distribution area, and selecting different numbers of pseudo-absences both at random and within the unfavourable environmental regions, this study showed how these factors influence the accuracy of predictive models carried out with poor quality data. If we lack reliable species absence data, is it better to include pseudo-absence data? How should we choose these pseudo-absence points? How many pseudo-absence points should we select? Also, if presence data are not evenly distributed over the whole territory, how can this bias affect the reliability of our models?

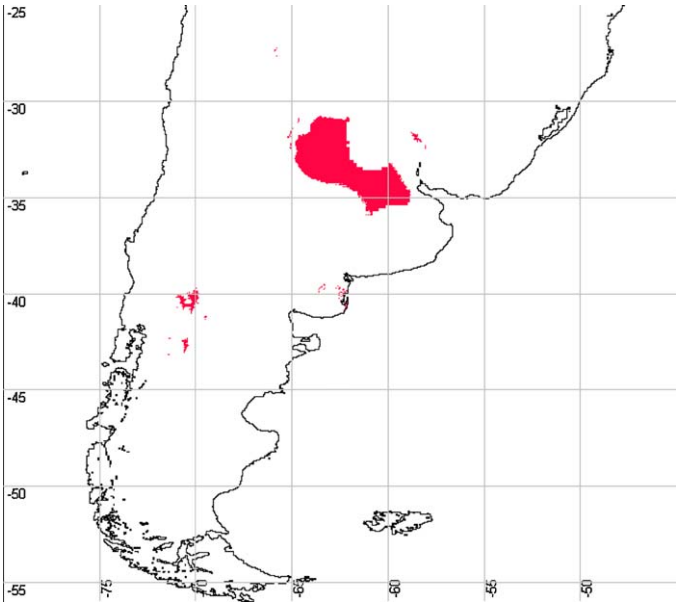

Figure 1. Mapped distribution of the virtual species in the Austral South America region (see Methods).

Methods

The virtual species

We mapped the distribution of a virtual species using current climate data, and then we used these same climatic data as predictor variables. The distribution of the virtual species was mapped in the Austral South American region (80–44°W longitude, and 56–24°S latitude), using a spatial resolution of ~0.04° (see Fig. 1). The total extent of the terrestrial studied region was 4,232,951 km^2 (246,013 0.04° × 0.04° cells). Twenty-one climatic variables were extracted from the WorldClim (Hijmans et al. 2005) interpolated map database (Table 1). Additionally, we used four other variables (wind speed, sunshine duration, frost day frequency, and relative humidity) extracted from the Climate Research Unit (New et al. 2002). These variables were standardised to eliminate measurement-scale effects (0 mean and 1 standard deviation). To select the minimum number of variables able to represent the environment of the considered region we used the so-called Jolliffe's principal component method (Rencher 2002). First, a Principal Component Analysis (PCA) was carried out with all of the considered variables, and five non-correlated factors with eigenvalues ≥1 were obtained that explained 87.13% of the climatic variation across the region. For each one of the five PCA factors, the variable with the highest factor loadings (which measure the correlations between the original variables and the factor axes) was selected (>0.8). The five selected variables were annual mean temperature, isothermality, mean diurnal range, precipitation of the driest month, and precipitation of the wettest quarter. Relative humidity was also included as a predictor variable because it was the only one that was not significantly correlated with any of the PCA factors. In total, these six variables have been considered as the most representative of the Austral South American

Table 1
Climatic variables from which a subset (marked with an asterisk) was selected to map the virtual species distribution (see text for details).

Annual mean temperature*
Annual precipitation
Frost days frequency
Isothermality*
Maximum annual temperature
Maximum temperature of warmest month
Mean diurnal range*
Mean temperature of coldest quarter
Mean temperature of driest quarter
Mean temperature of warmest quarter
Mean temperature of wettest quarter
Minimum annual temperature
Minimum temperature of coldest month
Precipitation of coldest quarter
Precipitation of driest month*
Precipitation of driest quarter
Precipitation of warmest quarter
Precipitation of wettest month
Precipitation of wettest quarter*
Precipitation seasonality
Relative humidity*
Sunshine duration
Temperature annual range
Temperature seasonality
Wind speed

climate. The distribution of the virtual species was determined only by these six climatic variables. After calculating the quartiles of each variable, the true unimodal distribution range of the virtual species was constructed by including all cells falling within the two central quartiles of the six climatic variables (Fig. 1). In total, the virtual species had 8616 presence (150,194 km^2, around 3.5% of total terrestrial area), and 237,397 $0.04^\circ \times 0.04^\circ$ absence cells. All geographic analyses were done with IDRISI Kilimanjaro GIS software (Clark Labs 2003).

Predictive models

The statistical relationship between the binomial dependent variables and the environmental predictors was established using generalised additive models (GAM), a non-parametric regression method that captures complex response curves and that is traditionally considered as having very good model performance (see Segurado & Araújo 2004 and references therein). GAMs models were fitted in *R* (www.r-project.org) using the *mgcv* package.

Models were calibrated using always only ten presences (0.01% of total presences) and 10, 100, or 1000 pseudo-absences. Presences were selected at random or in a spatially structured manner (arranging firstly the cells at random to subsequently select the 10 with higher latitudinal scores). Pseudo-absences were selected in two different ways: (1) randomly across the whole study area; or (2) outside the environmental domain previously defined by the available presence points. The simplest bioclimatic envelope model (BIOCLIM; Nix 1986), which involves intersecting the ranges inhabited by the species along each environmental variable, was used to define the range outside of which pseudo-absences were selected (see Lobo et al. 2006). The same climatic variables used to delimit the distribution of the virtual species were used to carry out the bioclimatic envelope model. The selection of presences and pseudo-absences were replicated ten times.

To measure how well the predicted map matched the virtual species distribution, derived probabilities or suitability scores were transformed into presence–absence data by applying the threshold which minimises the difference between sensitivity and specificity (Jiménez-Valverde & Lobo 2006; Jiménez-Valverde & Lobo 2007). The model distributions that were obtained from this process were then projected onto the whole study area, and the number of true presence and absence cases was compared against the predicted presences and absences (Fielding & Bell 1997). Three main accuracy measures were calculated from this confusion matrix: (1) specificity (the proportion of correctly predicted absent cells to the total number of absences or commission error); (2) sensitivity (the proportion of correctly predicted presences to their total number or omission error); and (3) the total area that falls under the receiver operating characteristic (ROC) curve (AUC). Although AUC values should not be used indiscriminately to compare the accuracy of models between different species (Lobo et al. 2008) it is an adequate threshold-independent accuracy measure for within species comparisons (Zweig & Campbell 1993; Fielding & Bell 1997; Fielding 2002). To calculate AUC, sensitivity is plotted against 1-specificity over a number of thresholds (100 in this case), and the area under the curve (AUC) calculated. AUC, ranges from 0.5 for models that have no discrimination ability, to 1 for models that have perfect discrimination.

The effect of the three main factors (number of pseudo-absences, location of presences and selection of pseudo-absences) on the three obtained accuracy parameters (i.e. AUC, sensitivity and specificity) were examined using a multi-factor ANOVA procedure to test the influence of each factor while controlling for all others, as well as to detect interaction effects among factors. ANOVA test assume that the dependent variable is normally distributed and that variances in the different groups are similar. Although the *F* statistic is quite robust against violations of these assumptions, we also estimated a rank transformation test because accuracy parameters show slight departures from normality (Kolmogorov–Smirnov one-sample test scores with probabilities between 0.05 and 0.01). For this test, all data are ranked from 1 to *N*, and a new ANOVA computed on ranks (Conover & Iman 1981). This procedure is far more robust to departures from the assumptions of normality and constant variance allowing multiple comparison procedures (Helsel & Hirsch 2002).

Results

Both the number of pseudo-absences and the location of presences (randomly or spatially structured) significantly influenced all accuracy measures. The interaction between these two factors was also statistically significant (Table 2) in the case of AUC and specificity. In contrast, the method applied to select pseudo-absences does not affect model performance. However, a slightly significant interaction occurred between this factor and the number of pseudo-absences (Table 2), which would indicate that the choice of pseudo-absences beyond the environmental envelope provided by the presence data generated better AUC scores only when the smallest number of pseudo-absences (10) was used.

Models generated with the largest number of pseudo-absences provided higher AUC scores. However, the percentage of well-predicted absences did not differ when 1000 and 100 pseudo-absences were used (85% and 86%, respectively; Fig. 2), and differences in the percentage of well-predicted presences can be considered negligible (Fig. 2).

Using spatial bias in the selection of presences resulted in worse AUC scores, mainly because the unavoidable lack of accuracy in the prediction of true presences (36% are incorrectly predicted as absences). Interestingly, this bias in the selection of used presences produces a low level of overprediction (Fig. 2). If presences are selected at random, 27% of absences are incorrectly predicted as presences (i.e., more than seven times the area of distribution).

The significant interaction between the number of pseudo-absences and the location of presences suggest that, if a high number of pseudo-absences are used, it is possible to obtain high AUC scores even when presence data are spatially biased. True pres-

Table 2
ANOVA results (*F* values) to test variations in model accuracy measures according to the number of pseudo-absences (10, 100 or 1000), the type of pseudo-absences (at random or outside the environmental envelope defined by presence localities), the location of used presences (at random or spatially structured), and all possible interactions among these three factors. Scores in brackets are the *F* ANOVA results computed on ranks (rank transformation test, see Conover & Iman 1981). *≤0.05, **≤0.01, ***≤0.001.

Factors	AUC	Sensitivity	Specificity
Number of pseudo-absences	71.23*** (78.82***)	4.90** (3.99*)	103.66*** (133.36***)
Location of presences	14.92*** (7.11**)	87.99*** (83.84***)	72.91*** (100.00 ***)
Selection of pseudo-absences	0.16 (0.10)	0.31 (0.04)	0.20 (0.21)
Number of pseudo-absences × location of presences	16.98*** (20.28***)	3.07* (0.50)	19.38*** (27.36***)
Selection of pseudo-absences × number of pseudo-absences	4.28* (2.29)	1.23 (0.88)	1.53 (1.81)
Location of presences × selection of pseudo-absences	2.72 (1.56)	1.26 (0.45)	0.63 (0.95)
Location of presences × selection of pseudo-absences × number of absences	0.55 (0.28)	0.08 (0.22)	0.18 (0.12)

ences were always relatively well predicted when presence data were chosen at random, independent of the number of pseudo-absences. The advantage of using a high number of absences in model training is that overprediction occurs less frequently when presence information is not randomly selected (Fig. 2). In contrast, overprediction in the species distribution may occur more frequently when a few randomly selected pseudo-absences are used.

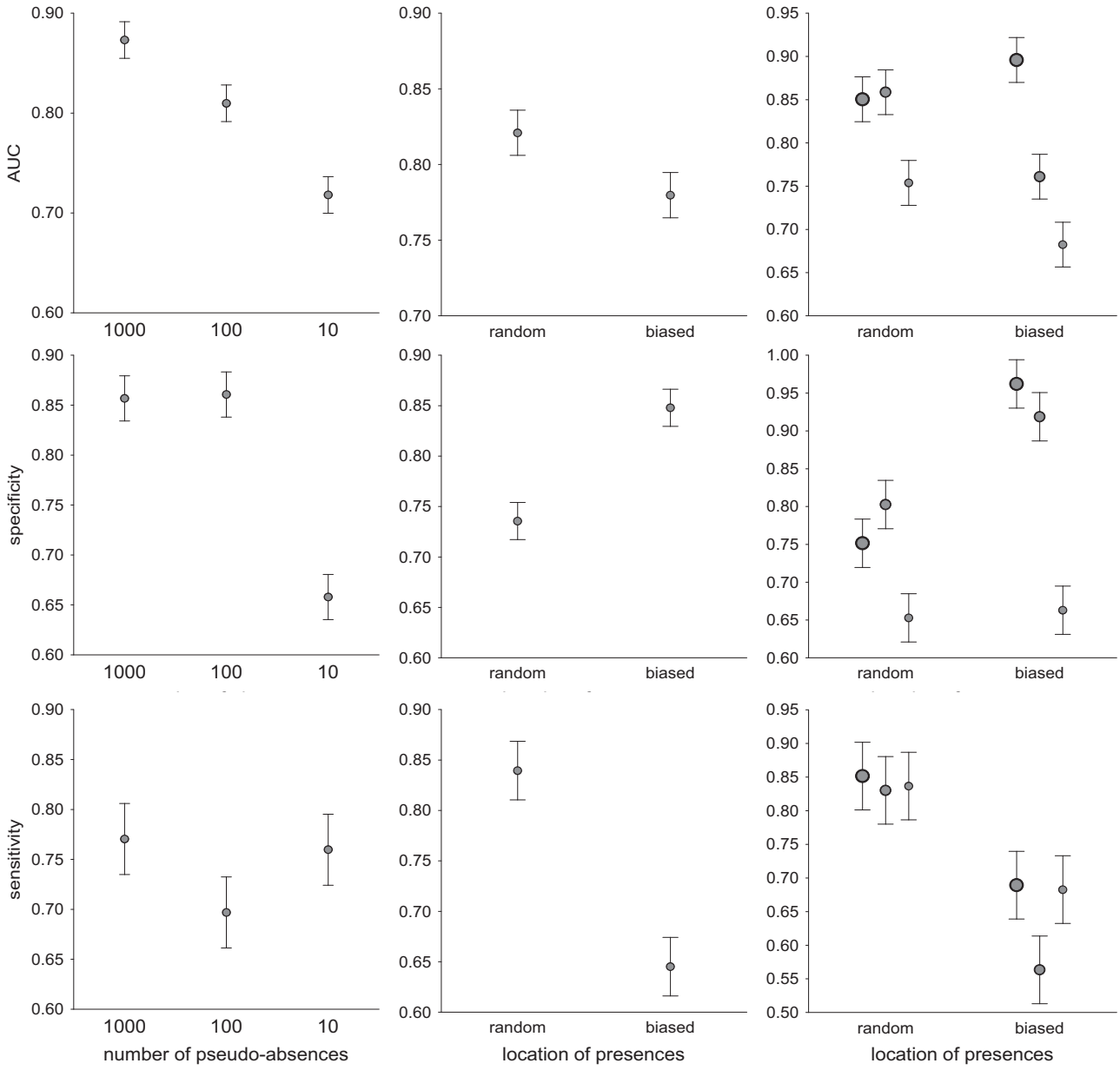

Figure 2. Variation in AUC, specificity and sensitivity scores (mean ± 95% confidence intervals; $n = 10$) according to the number of pseudo-absences (10, 100 or 1000) and the location of used presences (at random or spatially structured). Graphs in the right column represent significant interactions among the above-mentioned factors (see Table 1) in which the number of pseudo-absences are represented by circles of different sizes.

Discussion

Sampling bias in presence data and number of pseudo-absence points

Our results showed that, although the reliability of presence predictions does not increase with the number of pseudo-absences, the general accuracy of distribution models increases when 1000 pseudo-absences are used because of the low degree of overprediction of these models (less absence localities are erroneously predicted as presences). The advantage of using a high number of pseudo-absences is more evident when the presence data in the training are not homogeneously distributed across the environmental and geographical gradient of the whole study area, because the percentage of well-predicted absences surpasses 95% when 1000 pseudo-absences are used. In contrast, when the training data contains evenly distributed presences, the model tends to overpredict the distribution, especially if a low number of pseudo-absences are used. Overprediction of species distributions is a common shortcoming of these kind of models, particularly when limited distribution localities are used (Fielding & Haworth 1995; Araújo & Williams 2000; Stockwell & Peterson 2002; Brotons et al. 2004; Segurado & Araújo 2004; Stockman et al. 2006). But, why is overprediction less common when available presence data are spatially structured?

Available presence data may be frequently spatially or environmentally biased (Kadmon et al. 2004; Lobo et al. 2007; Hortal et al. 2007). Under these circumstances, used presences come from an area with lower environmental variability, and the predictive function restricts the range of environmental conditions in which the species can be found. Overprediction in species distribution models results from the lack of relevant explanatory variables such as biotic interactions (Austin & Meyers 1996; Fielding & Bell 1997; Parra et al. 2004; Peterson et al. 1999; Raxworthy et al. 2003), or the role of spatial autocorrelation (Segurado et al. 2006). However, we hypothesise that overpredictions are difficult to avoid in species distribution models, occurring more frequently when the available presence data come from the whole spectrum of the environmental conditions that the target species inhabits. Many studies advocate the use of large and evenly distributed data in predictive distribution models (see for example Hirzel et al. 2001; Stockwell & Peterson 2002; Zaniewski et al. 2002; Engler et al. 2004 or Reese et al. 2005). We do not suggest otherwise, however, we believe that there may be a greater tendency for modelling methods to overpredict species ranges when unreliable absence data is used (Lobo et al. 2010) and the available presence data used to train the model are scarce and at the same time homogeneously distributed along the environmental space.

Additionally, because most of the data for rare, endangered, or poorly surveyed species is generally environmentally and geographically biased, we also recommend that, regardless of the quantity and quality of available presence data, a high number of pseudo-absences should be used in modelling the distribution of these species (Jiménez-Valverde & Lobo 2006). Thus, when presence data is limited, more pseudo-absences should be incorporated to obtain more accurate predictive models. Minority-class predictions have a higher error rate than majority-class predictions (if there is no evidence favouring one classification over another, modelling methods tend to predict the majority class; see Weiss & Provost 2003). Thus, highly unbalanced designs (such as those having many pseudo-absences) facilitate the correct classification of the absence zone, but increase the misclassification of the presence zone, which is a desirable property when models are used for conservation purposes (see below). Austin & Meyers (1996) suggested restricting model calculations within the species' known environmental range, excluding the so-called "naughty noughts'. In contrast, Thuiller et al. (2004) proposed including a large part of the environmental combinations where the species currently occur or not. We agree with this last proposal; including many pseudo-absences in the training data: (i) increases the likelihood of representing all truly negative environmental regions; (ii) results in more complete response curves; (iii) maximises the explanatory capacity of the used environmental variables; and (iv) generates models with a lower rate of overprediction. Austin (2006) recently stated, "conclusions about the response curve of species can only be unambiguously determined if the sampled environmental gradient clearly exceeds the upper and lower limits of the species occurrence". More recently, Lobo et al. (2010) show that those absences from outside the environmental conditions determined by the presences are useful provided they are not excessively extreme. Hence, we recommend the use of a high number of environmentally weighted pseudo-absences able to cover all the environmental conditions in which the species is absent, but mainly those present at the environmental boundary between presences and absences. Of course, in this case, there is no issue regarding the use of the percentage of explained variability as a measure of model performance, which must be assessed by a correct validation procedure with independent data (Vaughan & Ormerod 2003, 2005).

How many pseudo-absences should we select to carry out predictive distribution models with limited presence data? It is widely recognised among statisticians that the recommendable sample size in logistic regressions must be at least ten times the number of explanatory variables (Peduzzi et al. 1996). Therefore, when presence data is limited, we suggest incorporating many absences in order to increase the sample size. Monte Carlo simulations based on species with more than 200 observations have indicated that the mean accuracy of prediction reaches its maximum asymptotic value at about 100 observations (Kadmon et al. 2003). However, if presence data is limited, sample size should be substantially larger. Dixon et al. (2005) have estimated that, when an event is truly rare, its probability estimate has reasonable precision only if the sample size exceeds 1000 total observations. In our case, increasing the number of presences from 10 to 100 and to 1000 decreased the coefficient of variation of these probability estimates from 117% to 50% and 16%, respectively (see Dixon et al. 2005 for calculations). This last precision score is considered acceptable. Therefore, to avoid excessively unbalanced designs (King & Zeng 2001) when we have limited presence data, we suggest selecting 100 times more pseudo-absences than presences. Again, unbalanced training data does not necessarily always result in inaccurate models (see Jiménez-Valverde et al. 2009). Machine learning algorithms perform well in very unbalanced designs (Prati et al. 2004), and an appropriate threshold to convert derived probabilities into a binomial variable can be used to avoid inaccuracies (Jiménez-Valverde & Lobo 2006).

Pseudo-absences and extent

In our case, the virtual species only inhabited 3.5% of the total study area. Because of this low ratio between the area of the species distribution and the whole extent of the modelled territory (the ROA or relative occurrence area; Lobo et al. 2008), there was little difference between randomly and environmentally selected pseudo-absences. Our results suggested that, when a small number of pseudo-absences are used, they should be selected in areas falling outside the environmental envelope defined by presence localities. Thus, as this type of pseudo-absence selection improves the results at lower relative occurrence areas, we suggest using this strategy for selecting pseudo-absence when there are no better alternatives for obtaining reliable absence data (see for example Lütolf et al. 2006). The influence of ROA in model distribution performance requires further investigation (but see Bulluck et al. 2006 and

VanDerWal et al. 2009). It is generally assumed that distribution models of restricted-range species generally perform better than those of widespread species (Araújo & Williams 2000; Segurado & Araújo 2004; Brotons et al. 2004). We suspect that this improved performance of the model may be influenced by the higher probability of selecting reliable absence data at random when there is a low relative occurrence area.

The recommendations provided by our study should be considered with caution when the target species do not inhabit all the territory with suitable environmental conditions (not in equilibrium with environmental conditions; see Araújo & Pearson 2005). In this case, models using environmentally weighted pseudo-absences will result in high rates of overprediction. Only an adequate selection of absence data together with the inclusion of predictors able to represent the role played by the processes that hinder the presence of a species under environmentally favourable conditions can improve the results of the models under these circumstances (Kadmon et al. 2003; Guo et al. 2005; Chefaoui & Lobo 2008; Lobo et al. 2010).

Acknowledgements

This study 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). Thanks to Judy Boshoven for helping with the English version of the manuscript.

References

Anderson, R. P. (2003). Real vs. artefactual absences in species distributions: Tests for *Oryzomys albigularis* (Rodentia: Muridae) in Venezuela. *Journal of Biogeography*, *30*, 591–605.

Araújo, M. B., & Williams, P. H. (2000). Selecting areas for species persistence using occurrence data. *Biological Conservation*, *96*, 331–345.

Araújo, M. B., & Pearson, R. G. (2005). Equilibrium of species' distributions with climate. *Ecography*, *28*, 693–695.

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.

Austin, M. (2006). Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling*, *200*, 1–19.

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.

Bulluck, L., Fleishman, E., Betrus, C., & Blair, R. (2006). Spatial and temporal variations in species occurrence rate affect the accuracy of occurrence models. *Global Ecology and Biogeography*, *15*, 27–38.

Chefaoui, R., & Lobo, J. M. (2008). Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecological Modelling*, *210*, 478–486.

Clark Labs. (2003). *Idrisi Kilimanjaro. GIS software package*. Worcester, MA: Clark Labs.

Conover, W. J., & Iman, R. L. (1981). Rank transformation as a bridge between parametric and nonparametric statistics. *The American Statistician*, *35*, 124–129.

Cramer, J. S. (1999). Predictive Performance of the binary logit model in unbalanced samples. *Journal of the Royal Statistical Society, Series D (The Statistician)*, *48*, 85–94.

Dennis, R. L. H., & Hardy, P. B. (1999). Targeting squares for survey: Predicting species richness and incidence of species for a butterfly atlas. *Global Ecology and Biogeography*, *8*, 443–454.

Dennis, R. L. H., & Thomas, C. D. (2000). Bias in butterfly distribution maps: The influence of hot spots and recorder's home range. *Journal of Insect Conservation*, *4*, 73–77.

Dennis, R. L. H., Sparks, T. H., & Hardy, P. B. (1999). Bias in butterfly distribution maps: The effects of sampling effort. *Journal of Insect Conservation*, *3*, 33–42.

Dennis, R. L. H., Shreeve, T. G., Isaac, N. J. B., Roy, D. B., Hardy, P. B., Fox, R., et al. (2006). The effects of visual apparency on bias in butterfly recording and monitoring. *Biological Conservation*, *128*, 486–492.

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

Engler, R., Guisan, A., & Rechsteiner, L. (2004). An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, *41*, 263–274.

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

Ferrier, S., Watson, G., Pearce, J., & Drielsma, M. (2002). Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling. *Biodiversity and Conservation*, *11*, 2275–2307.

Fielding, A. H. (2002). What are the appropriate characteristics of an accuracy measure? In J. M. Scott, P. J. Heglund, J. B. Haufler, M. Mprrison, M. G. Raphael, W. B. Wall, & F. Samson (Eds.), *Predicting species occurrences. Issues of accuracy and scale* (pp. 271–280). Covelo, CA: Island Press.

Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, *24*, 38–49.

Fielding, A. H., & Haworth, P. F. (1995). Testing the generality of Bird-Habitat Models. *Conservation Biology*, *9*, 1466–1481.

Graham, C. H., Ferrier, S., Huettman, F., Moritz, C., & Peterson, A. T. (2004). New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution*, *19*, 497–503.

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, *135*, 147–186.

Guo, Q., Kelly, M., & Graham, C. C. (2005). Support vector machines for predicting distribution of Sudden Oak Death in California. *Ecological Modelling*, *182*, 75–90.

Helsel, D. R., & Hirsch, R.M. (2002). *Statistical methods in water resources* (524 pp.). U.S. Geological Survey, Techniques of Water-Resources Investigations Book 4, http://water.usgs.gov/pubs/twri/twri4a3/.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, *25*, 1965–1978.

Hirzel, A. H., Helfer, V., & Metral, F. (2001). Assessing habitat-suitability models with a virtual species. *Ecological Modelling*, *145*, 111–121.

Hortal, J., & Lobo, J. M. (2005). An ED-based protocol for optimal sampling of biodiversity. *Biodiversity and Conservation*, *14*, 2013–2947.

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

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

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (second ed.). New York: John Wiley & Sons.

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

Jiménez-Valverde, A., & Lobo, J. M. (2007). Threshold criteria for conversion of probability of species presence to either-or presence–absence. *Acta Oecologica*, *31*, 361–369.

Jiménez-Valverde, A., Lobo, J. M., & Hortal, J. (2009). The effect of prevalence and its intercation with sample size on the reliability of species distribution models. *Community Ecology*, *10*, 196–205.

Kadmon, R., Farber, O., & Danin, A. (2003). A systematic analysis of factors affecting the performance of climatic envelope models. *Ecological Applications*, *13*, 853–867.

Kadmon, R., Oren, F., & Avinoam, D. (2004). Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, *14*, 401–413.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*, 137–163.

Liu, C., Berry, P. M., Dawson, T. P., & Pearson, R. G. (2005). Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, *28*, 385–393.

Lobo, J. M., 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., 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.

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., Jiménez-Valverde, A., & Hortal, J. (2010). The uncertain nature of absences and their importance in species distribution Modelling. *Ecography*, *33*, 103–114.

Luoto, M., Pöyry, J., Heikkinen, K., & Saarinen, K. (2005). Uncertainty of bioclimate envelope models based o geographical distribution of species. *Global Ecology and Biogeography*, *14*, 575–584.

Lütolf, M., Kienast, F., & Guisan, A. (2006). The ghost of past species occurrence: Improving species distribution models for presence-only data. *Journal of Applied Ecology*, *43*, 802–815.

Manel, S., Dias, J. M., Buckton, S. T., & Ormerod, S. J. (1999). Alternative methods for predicting species distribution: An illustration with Himalayan river birds. *Journal of Applied Ecology*, *36*, 734–747.

Martínez-Meyer, E. (2005). Climate change and biodiversity: Some considerations in forecasting shifts in species potential distributions. *Biodiversity Informatics*, *2*, 42–55.

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.

New, M., Lister, D., Hulme, M., & Makin, I. (2002). A high resolution data set of surface climate over global land areas. *Climatic Research*, *21*, 1–25.

Nix, H. A. (1986). A biogeographic analysis of Australian elapid snakes. In *Atlas of Australian Elapid Snakes*. Canberra, Australia: Bureau of Flora Fauna., pp. 4–15

Olden, J. D., Jackson, D. A., & Peres-Neto, P. R. (2002). Predictive models of fish species distributions: A comment on proper validation and chance predictions. *Transactions of the American Fisheries Society*, *131*, 329–336.

Parra, J. L., Graham, C. C., & Freile, J. F. (2004). Evaluating alternative data sets for ecological niche models of birds in the Andes. *Ecography*, *27*, 350–360.

Peduzzi, P., Concato, J., Kemper, R., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, *49*, 1373–1379.

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

Prati, R. C., Batista, G. E. A. P. A., & Monard, M. C. (2004). *Class imbalances versus class overlapping: An analysis of a learning system behavior* III Mexican International Conference on Artificial Intelligence. Lecture notes in computer science, vol. 2972, (pp. 312–321). Mexico City, Mexico: Springer-Verlag.

Raxworthy, C. J., Martínez-Meyer, E., Horning, N., Nussbaum, R. A., Schneider, G. E., Ortega-Huerta, M. A., et al. (2003). Predicting distributions of known and unknown reptile species in Madagascar. *Nature*, *426*, 837–841.

Real, R., Barbosa, A. M., & Vargas, J. M. (2006). Obtaining environmental favourability functions from logistic regression. *Environmental and Ecological Statistics*, *13*, 237–245.

Reese, G. C., Wilson, K. R., Hoeting, J. A., & Flather, C. H. (2005). Factors affecting species distribution predictions: A simulation model experiment. *Ecological Applications*, *15*, 554–564.

Rencher, A. C. (2002). *Methods of multivariate analysis* (second edition). New York: John Wiley & Sons.

Reutter, B. A., 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.

Segurado, P., & Araújo, M. B. (2004). An evaluation of methods for modelling species distributions. *Journal of Biogeography*, *31*, 1555–1569.

Segurado, P., Araújo, M. B., & Kunin, W. E. (2006). Consequences of spatial autocorrelation for niche-based models. *Journal of Applied Ecology*, *43*, 433–444.

Soberón, J., & Peterson, A. T. (2004). Biodiversity informatics: Managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London, Series B*, *359*, 689–698.

Soberón, J., & Peterson, A. T. (2005). Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity Informatics*, *2*, 1–10.

Stockman, A. K., Beamer, D. A., & Bond, J. E. (2006). An evaluation of a GARP model as an approach to predicting the spatial distribution of non-vagile invertebrate species. *Diversity and Distributions*, *12*, 81–89.

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

Stockwell, D. R. B., & Peters, D. (1999). The GARP modelling system: Problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science*, *13*, 143–158.

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

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.

Vaughan, I. P., & Ormerod, S. J. (2003). Improving the quality of distribution models for conservation by addressing shortcomings in the field collection of training data. *Conservation Biology*, *17*, 1601–1611.

Vaughan, I. P., & Ormerod, S. J. (2005). The continuing challenges of testing species distribution models. *Journal of Applied Ecology*, *42*, 720–730.

Weiss, G. M., & Provost, F. (2003). Learning when training data are costly: The effects of class distribution on tree induction. *Journal of Artificial Intelligence Research*, *19*, 315–354.

Williams, P. H., Hannah, L., Andelman, S., Midgley, G., Araújo, M. B., Hughes, G., et al. (2005). Planning for climate change: Identifying minimum-dispersal corridors for the Cape Proteaceae. *Conservation Biology*, *19*, 1063–1074.

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.