Ecography 33: 103–114, 2010
doi: 10.1111/j.1600-0587.2009.06039.x

Subject Editors: Núria Roura-Pascual and Nathan K. Sanders. Accepted 30 November 2009

The uncertain nature of absences and their importance in species distribution modelling

Jorge M. Lobo, Alberto Jiménez-Valverde and Joaquín Hortal

J. M. Lobo (mcnj117@mncn.csic.es), Dept Biodiversidad y Biología Evolutiva, Museo Nacional de Ciencias Naturales, c/ José Gutiérrez Abascal 2, ES-28006, Madrid, Spain. – A. Jiménez-Valverde, Natural History Museum and Biodiversity Research Center, The Univ. of Kansas. Lawrence, KS 66045, USA. – J. Hortal, NERC Centre for Population Biology, Div. of Biology, Imperial College London, Silwood Park Campus, Ascot, Berkshire SL5 7PY, UK.

Species distribution models (SDM) are commonly used to obtain hypotheses on either the realized or the potential distribution of species. The reliability and meaning of these hypotheses depends on the kind of absences included in the training data, the variables used as predictors and the methods employed to parameterize the models. Information about the absence of species from certain localities is usually lacking, so pseudo-absences are often incorporated to the training data. We explore the effect of using different kinds of pseudo-absences on SDM results. To do this, we use presence information on *Aphodius bonvouloiri*, a dung beetle species of well-known distribution. We incorporate different types of pseudo-absences to create different sets of training data that account for absences of methodological (i.e. false absences), contingent and environmental origin. We used these datasets to calibrate SDMs with GAMs as modelling technique and climatic variables as predictors, and compare these results with geographical representations of the potential and realized distribution of the species created independently. Our results confirm the importance of the kind of absences in determining the aspect of species distribution identified through SDM. Estimations of the potential distribution require absences located farther apart in the geographic and/or environmental space than estimations of the realized distribution. Methodological absences produce overall bad models, and absences that are too far from the presence points in either the environmental or the geographic space may not be informative, yielding important overestimations. GLMs and Artificial Neural Networks yielded similar results. Synthetic discrimination measures such as the Area Under the Receiver Characteristic Curve (AUC) must be interpreted with caution, as they can produce misleading comparative results. Instead, the joint examination of ommission and comission errors provides a better understanding of the reliability of SDM results.

Estimating the different aspects of the geographical distribution of species from the fragmentary distribution data (presence/absence data) that are commonly available is of great value for both basic and applied purposes. Such an approach has been named ''niche-based modelling'', ''ecological niche modelling'', ''habitat suitability modelling'', ''climate envelope modelling'' or ''species distribution modelling'', among other denominations. For convenience, herein we use the latter in its abbreviated form, SDM. An exhaustive search in the ISI Web of Science including both these topics and the names of a number of authors publishing in the field yield 2333 studies published so far at the end of 2008. The increase in the number of publications followed an almost exponential rate since the original formulation of the theoretical framework in the 1980s by the seminal works of Busby (1986) and Austin et al. (1990), among others. From 1995 the net rate of increase in the number of published papers is 20.2 yr^{-1} (Fig. 1). Such widespread interest is an indicator of this field of research having become a hot topic in the ecological, biogeographical or conservation literature in the last two decades, and that its attraction among biologists is still increasing. However, is this interest justified by a true success of SDM in describing the distribution of species?

A great part of these modelling exercises overlook the conceptual and methodological implications of discerning between potential and realized distributions, as well as the influence of the kind and quality of the primary data used to build the models (Jiménez-Valverde et al. 2008a). Theoretically, each organism is adapted to specific tolerance zones or ''niches'' which, in a Grinellian sense, can be considered as the set of abiotic requirements in which a species can maintain a net positive rate of population increase without immigration (Grinnell 1917, Soberón and Peterson 2005, Soberón 2007, 2010, Colwell and Rangel 2009, Soberón and Nakamura 2009). It follows that if the main environmental variables that delimit such ''niche'' were known, then it would be possible to estimate the potential distribution of the species (i.e. the places environmentally suitable to maintain its populations), at

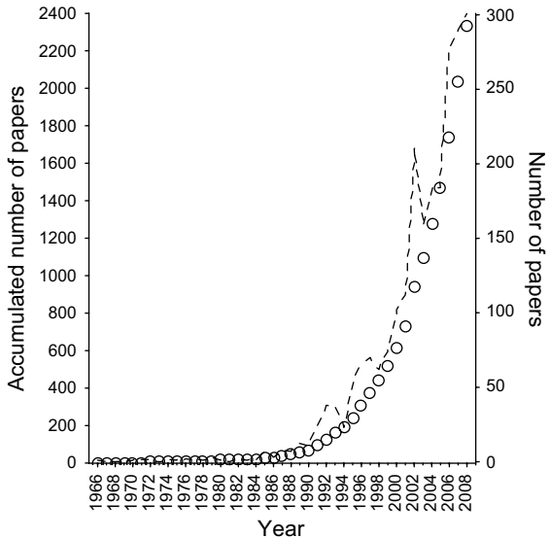

Figure 1. Number of published papers (line) and variation in the number of accumulated papers (points) on species distribution modelling found in ISI Web of Science after performing an exhaustive search by topics and authors.

least at relatively coarse geographical scales (Soberón and Nakamura 2009, Soberón 2010). However, if we were to use distributional data to generate such hypothesis, it would be necessary to assume that the environmental conditions in the localities where the species is present (herein, presences) provide a reliable description of its whole requirements. Such assumption implies that 1) the distribution of a species is an accurate geographical representation of its niche, and 2) the distributional data are not biased and recover the whole gradient of environmental conditions in which the species can inhabit.

Unfortunately, these two intrinsic requirements are never true. First, temporal changes in the environmental conditions and non-environmental processes such as dispersal limitations or historical factors constrain the potential distribution of species (Pulliam 1988, 2000, Ricklefs and Schluter 1993, Hanski 1998, Ricklefs 2007). Thus, species are never in equilibrium with the environment (Svenning and Skov 2004, Araújo and Pearson 2005) and presence points reflect their realized, not their potential, distributions. Second, most distributional data have not been collected with standardized sampling protocols. Hence, bias and lack of coverage are general characteristics of the presences known for most species (Dennis et al. 1999, Dennis and Thomas 2000, Hortal et al. 2007, 2008) and, so, they virtually never reflect the whole spectrum of conditions inhabited by the species in the time frame considered. The quality and representativeness of the distributional data are especially important because the sample from which the relationships among variables are inferred should be representative of the population described (Zar 1999). In other words, the training data used on SDM must represent the environmental gradients of the study region adequately (Kadmon et al. 2004, Hortal et al. 2008).

Our purpose is to highlight that distribution predictions are necessarily of provisional nature, especially when biased data are used. Certainly, SDM could be highly useful to study and/or represent the distributions of groups for which the available information is not only scarce but also impossible to obtain in the near future, such as insects. However, scarce data in hyperdiverse taxa usually imply environmental and geographical biases (Hortal et al. 2007, 2008, Lobo et al. 2007). This fact imposes a paradox: the more SDM are needed, the more difficult it is to apply them. Hence, a well-developed conceptual framework is needed to establish solid foundations where this emerging field of research can develop into a robust body of knowledge.

In a former work we argued that the type of distribution data (presence and absence) used for SDM determines the final results, as well as the capacity to represent the potential or the realized distribution of the species (Jiménez-Valverde et al. 2008a). Here we further explore how the type of absence information used influences the species' geographical distributions that are finally inferred from SDM, as well as the subsequent shifts from the potential to the realized domains. To do this, we examine how the different distributional predictions obtained by the use of different kinds of absence data differ in their capacity to represent the potential and/or realized distribution of a species. We use *Aphodius bonvouloiri*, an Iberian endemic dung beetle species of known distribution and present in the fossil record (see below) as a model species. Firstly, we generate maps of the probability of finding the three possible types of absences (methodological, environmental and contingent, see below). Subsequently, we use these maps in a classic presence–absence SDM procedure using climatic variables as predictors in order to identify the roles played by the two main processes responsible of the distribution of species: environmental adaptations (Brown et al. 1996) and dispersal limitations (Svenning and Skov 2005), as well as the effect of false absences caused by biases in the distributional data (Hortal et al. 2008).

Three kinds of absences

If presences inform about the places that are environmentally suitable for a species (with some noise due to source/sink population dynamics and/or high dispersal capacity, Pulliam 2000), absences do the opposite. However, absences provide a more diverse source of information that the mere lack of suitability of some places. Some of the localities from where a species is absent can in fact be environmentally favourable places where dispersal limitations (Svenning and Skov 2005), historical factors (McGlone 1996), local extinctions (Hanski 1998), biotic interactions or other factors such as the size of the patches of suitable habitat (Hirzel and Le Lay 2008) have prevented the presence of the species. From now on we define contingent absences as those caused by these restrictive forces on the pool of a priori climatically or environmentally favourable areas. In contrast, we define environmental absences as those caused solely by the lack of environmentally or climatically favourable conditions in these places. Both kinds of absences are the outcome of the processes shaping species distributions (see Soberón and Nakamura 2009 and Soberón 2010 for a detailed framework). Apart

from these two types, there is also a third kind of absences that derives from the very nature of distributional information, which is frequently (if not always) incomplete and biased (see Gaston 1991, Gaston and Blackburn 1994, Dennis et al. 1999, Dennis and Thomas 2000, Graham et al. 2004, Soberón and Peterson 2004, Hortal et al. 2007, 2008, Lobo et al. 2007 and references therein). We call these absences methodological absences because they are the result of the bias and scarceness in the survey information. This type of absences might constitute the most important source of uncertainty for the study on the patterns and processes underlying the geographic distribution of biodiversity (the so-called Wallacean shortfall; Whittaker et al. 2005). Following the niche framework proposed by Soberón (2007, 2010, Soberón and Nakamura 2009), contingent absences would be outside of the realized but inside the fundamental niche, environmental absences would be outside of both the realized and fundamental niche, and methodological absences would be inside both the realized and the fundamental niche.

Although knowing the location of absences and their type (i.e. their origin) would be highly informative, it is perhaps a naïve goal. However, we argue that the probability of occurrence of each type of absences varies across the territory, according to the spatial and environmental distance of each locality from the conditions prevailing in the known presence points. More precisely, environmental absences are more probable in those localities showing environmental conditions far away from the environmental universe defined by the presence localities. Conversely, contingent absences will be more probable in spatially distant localities with favourable environmental conditions, while the probability of finding methodological absences will be higher in the environmentally favourable localities placed nearest to the known presence points. Here we assume that is more likely that these absences correspond to lack of knowledge than to the actual absence of populations, although this is also scale-dependent (Hortal 2008, Kriticos and Leriche 2010). Below a certain grain threshold (different for each species), presence will largely depend on micro-habitat selection, short-distance dispersal and metapopulational processes (Wilson et al. 2010), thus increasing the number of truly contingent absences that are placed spatially near the observed presences. However, if the scale of analysis is appropriate, by including subsets of absences selected according to these probabilities in the calibration data (together with the known presences), we will be able to identify their effects on the results of SDM, as well as to study the different aspects of the distribution of a species.

A case study: *Aphodius bonvouloiri*

Study species

Aphodius bonvouloiri is a dung beetle species (Coleoptera, Scarabaeoidea, Aphodiinae), currently endemic to the medium to high mountain areas of the north and central Iberian Peninsula (Fig. 2A). However, this species was one of the most abundant beetles in the south of Great Britain during a temperate interlude in the middle of the last glaciation around 43 000 yr BP (Coope and Angus 1975, Coope 1990). During such interstadial period summers were warmer than at present and *A. bonvouloiri* was associated with other temperate species, hence being a clear example of an insect species which current distribution is not in equilibrium with current climatic conditions.

Origin of presence data

We compiled all the available information on the distribution of *A. bonvouloiri* from natural history collections and bibliographic sources. In total, this species has been recorded at 47 UTM cells of 100 km^2 (Fig. 2) (i.e. 47 presences), being the resolution determined by the spatial precision that we could obtain from most of the specimens deposited in natural history collections. Apart from the Iberian Peninsula, we considered most of France and parts of Great Britain and the north of Africa to estimate the potential distribution of the species. The whole extent of the area considered is 1 806 100 km^2, from $-14°3'$ to $12°15'$ in longitude, and from $34°45'$ to $52°43'$ in latitude, approximately (Fig. 2).

Selection of predictors

A necessary first step in the use of SDM is the selection of the predictor variables that are most likely to be relevant for the distribution of the species, especially when the information is scarce and/or biased. When both presences and reliable absences are available, the most important predictors can be recognized by means of, e.g. variance partitioning or hierarchical partitioning methods (Legendre and Legendre 1998, MacNally 2002). However, when the only reliable data available are presences, some exploratory analyses are needed. The Ecological-Niche Factor Analysis (ENFA) provides a mean of making such exploration, since it allows identifying the major environmental requirements of a species (Hirzel et al. 2002, Basille et al. 2008, Calenge and Basille 2008, Calenge et al. 2008). By assuming that the variables with a lower level of variability in the presence locations are those having a higher likelihood of limiting the distribution (Rotenberry et al. 2006), ENFA allows to identify the response of the species to the main environmental variations in the study area. ENFA transforms the original ecogeographical variables into new orthogonal axes. The first axis accounts for the marginality of the species, i.e. differences between the conditions inhabited by the species and the regional average conditions. The other axes (specialisation axes) account for the tolerance of the species to other secondary environmental gradients in the study area (see Hirzel et al. 2002 for a detailed explanation of the method).

Climatic factors are among the most important agents limiting the demography and colonization ability of insect species, due to their general physiological dependence of environmental temperatures (Chown and Terblanche 2007). Hence, we used the nineteen bioclimatic variables (Table 1) derived from monthly temperature and rainfall values provided by the WorldClim ver. 1.4 interpolated map database (Hijmans et al. 2005; <www.worldclim.org/>) as

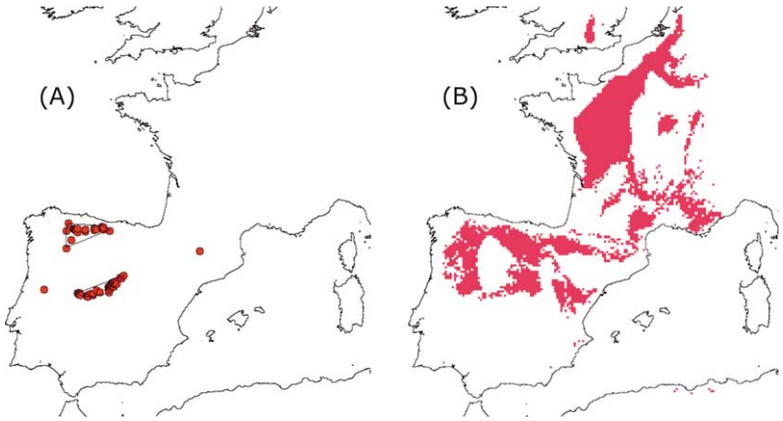

Figure 2. (A) Localities with information about the presence *Aphodius bonvouloiri* (red circles) and the convex-hull polygons which joint all known observations, used to characterize the realized distribution for evaluation purposes (see text). (B) Model of the potential distribution of the species, generated by first estimating the most relevant climatic variables through an ENFA analysis (Table 1), and subsequently use these variables to build a simple multidimensional envelope model which includes all the 100 km^2 UTM cells with climatic values within the range of climatic conditions in which the species was observed (see text).

input ecogeographical variables in ENFA. ENFA was performed using Biomapper (Hirzel et al. 2004).

After normalizing the climatic variables by a box-cox procedure, ENFA results showed a marginality of 1.08 (i.e. the optimum for *A. bonvouloiri* is relatively far from the mean available conditions in the region). Further, specialisation was 7.87 indicating that the species inhabit on a narrow interval of climatic conditions (almost eight times smaller than the whole range of variation available). Three factors allow explaining 85% of total information and 69% of specialisation (Table 1). The first axis (marginality) was negatively related with temperature variables (mainly mean temperature of the wettest quarter, annual mean temperature and minimum temperature of the coldest month), and positively with seasonal variables (mean diurnal range, annual temperature range and isothermality). In other words, *A. bonvouloiri* occupies localities with low temperatures during winter and autumn, and high seasonal and/or diurnal temperature oscillations. Temperature variables contributed the most to the first and second specialization axes, either negatively (maximum temperature of the warmest month) or positively (temperature annual range and minimum temperature of the coldest month), respectively. Among the precipitation variables, only rainfall during summer was positively related with these specialization factors. Thus, this species seems to be restricted from areas with high temperatures and low precipitations during the summer, but favoured by the existence of low winter temperatures and wide annual temperature ranges. In total, we selected for the SDM analysis seven climatic variables that showed ENFA factor scores higher than 0.20, but were not highly correlated between them (absolute Pearson correlation values lower than 0.80) (Table 1).

Table 1. WorldClim bioclimatic variables with ENFA factor scores higher than 0.20 for the marginality first axis (F1) and the two main specialisation axis (F2 and F3). Variables in bold are those selected for the multidimensional envelope model (see text). V1 and V3 were rejected due to their high correlation with V2 ($r = 0.91$ and 0.98, respectively), while V4 was rejected due to its correlation with V7 ($r = 0.97$).

		F1	F2	F3
V1	Annual mean temperature	−0.39		
V2	**Minimum temperature coldest month**	−0.38	0.49	0.58
V3	Mean temperature coldest quarter	−0.32		
V4	Mean temperature warmest quarter	−0.31		
V5	**Mean temperature wettest quarter**	−0.46		
V6	Mean temperature driest quarter			
V7	**Maximum temperature warmest month**		−0.55	−0.52
V8	Annual mean precipitation			
V9	Precipitation coldest quarter			
V10	Precipitation driest month			
V11	**Precipitation driest quarter**		0.31	0.23
V12	Precipitation warmest quarter			
V13	Precipitation wettest month			
V14	Precipitation wettest quarter			
V15	**Temperature annual range**	0.23	0.51	0.45
V16	**Mean diurnal range**	0.31		
V17	**Isothermality**	0.24		
V18	Temperature seasonality			
V19	Precipitation seasonality			

Creating maps of probability for each kind of absences

The probability of each locality where the species has not been recorded to be one of the three kinds of absences relies on both climatic favourability and distance to the presence localities (see above). Thus, in a first step we used Mahalanobis Distance (MD, Farber and Kadmon 2003) to calculate the climatic favourability (i.e. the distance between each cell and the conditions prevailing in the known presence cells according to the seven predictors selected before, *envdist*). This measure differs from the Euclidean distance in that it takes into account the dependence among the variables being also scale-invariant (i.e. the variables have the same weight independently of their variance). Mahalanobis distances oscillate between 0 and 660; only 1.1% of all 100 km^2 UTM cells (n = 18061) showed values higher than 100, and 3.7% values higher than 50. Thus, we decided to use the logarithm of the Mahalanobis distance as a measure of climatic favourability. Apart from that, we created a map of the spatial distance to the presence localities (using Euclidean Distances), and rescaled it to vary between 0 and 1 (*spdist*).

We calculated the probability of occurrence of environmental absences (p_{EA}) by rescaling climatic favourability to vary between 0 and 1. Similarly, we estimated the probability of occurrence of methodological absences (p_{MA}) as the complementary probability (i.e. $1 - p$) of the product of the rescaled spatial distances (*spdist*) by the probability of environmental absences: $p_{MA} = 1 - (spdist \times p_{EA})$. This ensures that, climatically favourable cells placed near the presences possess a higher likelihood of being false absences by methodological reasons. Given that absences resulting from contingent effects are more probable in geographically distant but environmentally favourable cells, we calculated the probability of occurrence of contingent absences (p_{CA}) as the product of the complementary probability of environmental absences by the rescaled spatial distances: $p_{CA} = (1 - p_{EA}) \times spdist$. The probability of occurrence of environmental absences increases constantly with distance until 500 km (the approximate limit of the Iberian Peninsula), while those of methodological or contingent absences either diminish or increase almost regularly with the distance to the presence cells (Fig. 3).

Hypothesizing the potential and realized distribution of the species

To estimate the role of selecting absences with different probabilities of being caused by environmental, contingent or methodological processes in the variation of SDM results we first need maps of the realized and the potential distributions of the considered species to evaluate the models. The knowledge on the distribution of *A. bonvouloiri* in Europe is reasonably complete; this species is

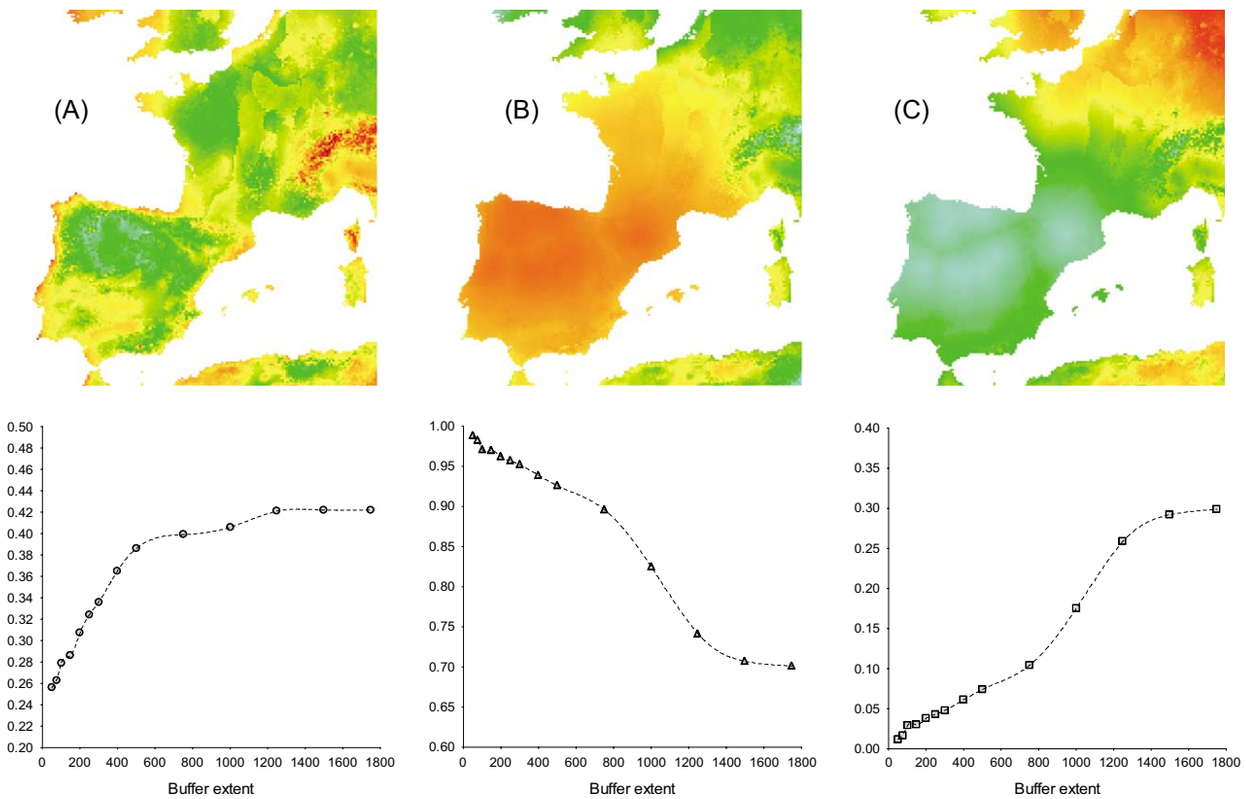

Figure 3. Variation in the probability of occurrence of environmental (A), methodological (B) and contingent (C) absences of *A. bonvouloiri* (Fig. 2). Probability values oscillate from cold (blue; low values) to hot colours (red; high values). The bottom scatterplots show the variation in the mean probabilities of the cells according to fourteen buffer extents located at increasingly higher distances apart from the centroid of presence known localities (50, 75, 100, 150, 200, 250, 300, 400, 500, 750, 1000, 1250, 1500 and 1750 km).

unknown from the most exhaustively surveyed countries in central and northern Europe (Löbl and Smetana 2006). Thus, the realized distribution of the species was assumed to be represented by a minimum convex polygon (i.e. the smallest polygon in which no internal angle exceeds 180 degrees and contains all presence sites). In order to reduce the frequent overprediction error of this method (Burgman and Fox 2003), we excluded discontinuities within the species range by performing separate convex-hulls for regionally aggregated data and also maintaining unconnected the two most distant presence localities (Fig. 2A).

The potential distribution of the species was estimated by selecting a multidimensional envelope on the seven climatic variables previously selected by the ENFA exploratory analysis. The limits of such envelop correspond to the maximum and minimum values of each climatic variable in the known presence localities. All the localities within such range of conditions in the studied region were considered to represent the potential distribution of the species (i.e. BIOCLIM, Busby 1986; Fig. 2B). Due to its simplicity, this procedure probably allows generating the most reliable hypotheses on the potential distribution of species when only few and biased presence data are available, provided that a set of variables known to be relevant are used. If the data available is biased, new-found presences will likely be located outside the observed environmental domain. Such lack of completeness is quite common in distributional data (Lobo et al. 2007, Hortal et al. 2008), so the reliability of the estimate of the potential distribution for a species that has not been exhaustively surveyed will increase when the predicted distribution is the widest possible according to the environmental conditions of presence localities. Hence, a simple technique not prone to generate restricted distributions through a tight fit to the data would be the most appropriate to describe the potential distribution of the species.

Assessing the effect of the different kinds of absences

The values of each one of the three probability of absence maps were divided in five categories: below the 10% percentile, between 10% percentile and 25% quartile, between 25 and 75% quartiles, between 75% quartile and 90% percentile, and above 90% percentile. Calibration datasets of prevalence equal to 0.1 (Jiménez-Valverde et al. 2009a) were obtained for the 15 combinations of kind of absence (3 kinds) and percentile category (5 categories), by selecting at random ten times more absences than the observed number of presences. The random selection of absences was repeated 50 times for each one of these combinations. Presence/absence data from these calibration datasets were modelled using Generalized Additive Models (GAMs; Hastie and Tibshirani 1990) with a logit link function. This technique was chosen because is traditionally considered to show very good model performance (see Segurado and Araújo 2004 and references therein). Here, the seven variables previously selected by the ENFA exploratory analysis with three degrees of freedom were submitted to a backward stepwise selection procedure (Harrell 2001). In total, 750 models were run ($50 \times 3 \times 5$). For each combination of kind and percentile of absence, the probabilities of occurrence obtained with the 50 models were averaged in each pixel, to obtain a probability map for each one of these 15 percentile-kinds of absence combinations (5×3). The prevalence of the training data (0.1) was used as the threshold to convert these continuous maps into Boolean presence/absence maps (Jiménez-Valverde and Lobo 2006, 2007a). These maps were compared with the potential and real distribution maps by means of their sensitivity (proportion of presences correctly predicted as presences) and specificity (proportion of absences correctly predicted as absences). The AUC (Area under the Receiver Characteristic Curve) was also computed as a measure independent of a threshold value (but see Lobo et al. 2008) and because, at present, it is the most widely used statistic for model evaluation. In addition, we conducted the same analyses using other two widely applied SDM techniques: General Linear Models (GLMs; McCullagh and Nelder 1989) and Artificial Neural Networks (ANNs; Özesmi et al. 2006). GLMs were allowed for cubic terms, using a logit link function and a backward stepwise selection of variables. In the case of ANNs, the hidden layer was set to contain 15 neurons, initial weights of connections were set to 0.1 and the maximum number of iterations was 2000. Given that the results are consistent regardless of the SDM technique used (compare Fig. 4 with Fig. S1 in the Supplementary material), for clarity we will base the Results section on the GAM results, unless noted otherwise. All analyses were made using the gam (Hastie 2008), nnet (Venables and Ripley 2002) and PresenceAbsence (Freeman 2008) packages for R (ver. 2.7.2; R Development Core Team 2008).

Results

When the aim is to predict the potential distribution, species presences are well predicted when absences are selected among the areas far away from the known presences, both geographically and environmentally (Fig. 4). In these cases, GAM models show high percentages of explained deviance (Fig. 5B). These results show that to improve the estimates of the potential distribution it is advisable to avoid using absence data from localities that are either environmentally suitable or placed geographically near to the presence data used. The models carried out with these absences show important rates of omission and commission errors, predicting potentially suitable places in areas that are not suitable. The high rate of false absences results in bad-performance models, with low percentages of explained deviance (Fig. 5B). Interestingly, high sensitivity also implies a high commission error rate (i.e. overprediction, low specificity) and absences placed too far away from the presences in both the environmental and geographical spaces decrease model performance. Thus, the best specificity values ($\approx$90%) are obtained when the absences are selected within the regions with probability values between the lower (25%) and upper quartiles (75%), independently of the type of absences considered. The best model (AUC = 0.938) was obtained with contingent absences in the $>$90% percentile. However, it is worth noting that this model overpredicts the potential

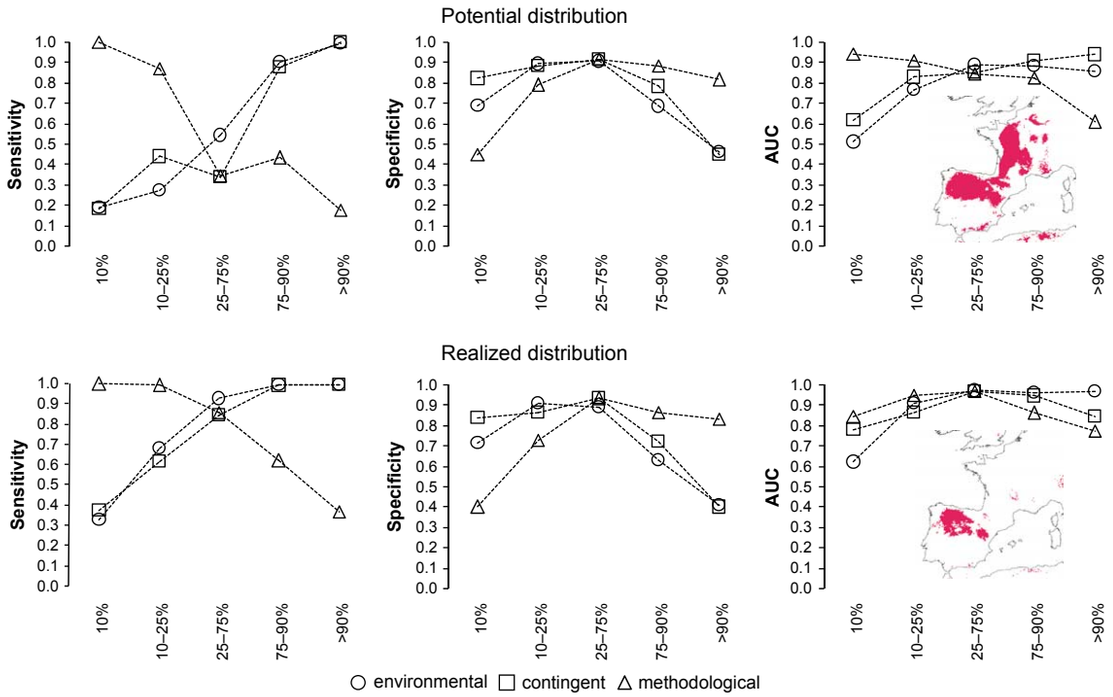

Figure 4. Variation in the rate of accurate predictions of presence (sensitivity), accurate predictions of absence (specificity) and AUC values according to the location and type of the absences used to model either the potential or the realized distribution of *Aphodius bonvouloiri* using GAMs. The used presences are the available observations of this species (Fig. 2), while ten times more environmental (circles), contingent (squares) or methodological absences (triangles) were randomly selected among those present in five probability categories (Fig. 3): below 10% percentile, between 10% percentile and 25% quartile, between 25 and 75% quartile, between 75% quartile and 90% percentile, and above 90% percentile. The maps provided are those in which the sum sensitivity and specificity values are higher and the absolute difference between the two accuracy measures are lower [maximization of: (sensitivity + specificity) − |sensitivity − specificity|], for both the potential and realized distributions.

distribution of the species in 118 500 km^2 (Fig. 4), increasing the original potential area in about a third (36%, Fig. 2), and also fails to correctly predict around a fourth of the potential area of presence (23%, 75 100 km^2).

In the case of the realized distribution, the accurate prediction of presences also requires avoiding the absences from nearest localities which are contaminated with methodological absences. Similar to the former, including absences from those environmentally suitable and geographically close localities to the presences results in poorly-performing models. However, as the realized distribution is smaller than the potential one, high sensitivity

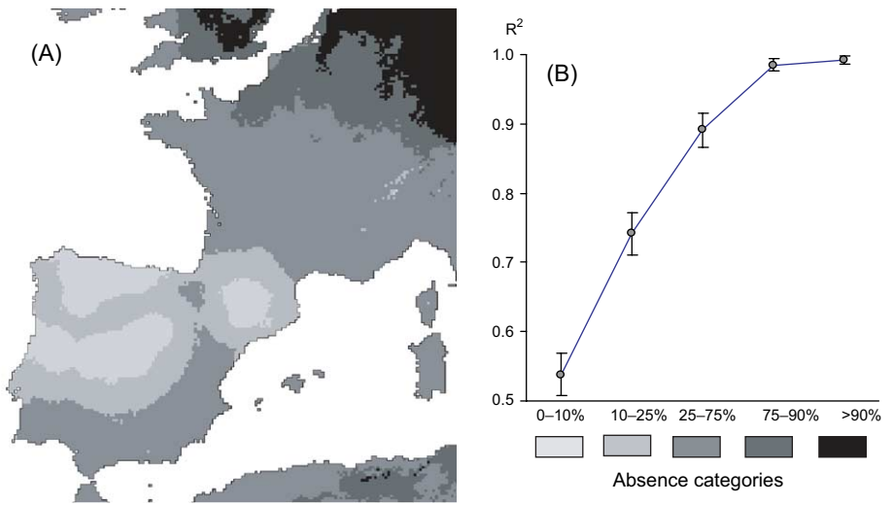

Figure 5. (A) Location of absence zones (from light grey to black) according to the five categories of the contingent probability map (Fig. 3). (B) Variation in the values of the coefficient of determination (±SD) of the models according to the five categories from which absences were selected. The shades of grey below these categories correspond to the greyscale in the map.

values are reached sooner than in the case of the potential distribution (Fig. 4 and Fig. S1 of the Supplementary material). As before, including the most geographic or climatically distant places in the training data increases the rate of commission errors, because they generate too-wide distribution simulations with a high overprediction rate. Both sensitivity and AUC values are significantly higher in the case of real distribution simulations than for potential distribution representations (Mann-Whitney U test, $Z = 1.94$, $n1 = n2 = 15$, $p = 0.05$ and $Z = 1.97$, $n1 = n2 = 15$, $p = 0.05$, respectively). However, specificity values are not significantly different among potential and realized models ($Z = 0.19$, $n1 = n2 = 15$, $p = 0.85$) showing that the gain in predictive power is mainly due to the higher success in predicting correctly the smaller-sized realized distribution. However, as in the case of the potential distribution and despite the higher values of discrimination capacity, model results are far from being accurate. The best model (AUC = 0.969, sensitivity = 0.930, specificity = 0.886) was obtained with absences from the 25–75 percentile (Fig. 5A), and overpredicts the distribution range of the species nearly 5 times (119 900 km^2 from a realized distribution of 20 600 km^2, a 482% more), incorrectly predicting as absences only 8% of the original area of presence (1700 km^2) (compare Fig. 2 and Fig. 4).

Our results clearly suggest that the change in model reliability among geographical estimates of both realized and potential distributions is basically due to the success in predicting species presences. However, obtaining a high rate of success in predicting presences is negatively correlated with the success in absences (Pearson correlation coefficient between sensitivity and speciality values $r = -0.57$, $n = 30$, $p < 0.01$). This evidences that there is a trade-off between maximizing the rate of predicting presences or absences (Fielding and Bell 1997), so when a high percentage of presences are correctly predicted it is unavoidable to commit overpredictions (Fig. 4 and Fig. S1 in the Supplementary material). This trade-off is also evident when the predictions for potential and realized distributions are compared. Maximizing the success in predicting the potential distribution inevitably generates a high rate of overprediction in the realized one; hence, there is also a negative correlation between the sensitivity of the potential distribution and the specificity of the realized one (Pearson $r = -0.80$, $n = 15$, $p = 0.0003$).

Discussion

A few studies have stressed the negative effects of false absences for the results of species distribution models (Tyre et al. 2003, Gu and Swihart 2004, Pearce and Boyce 2006), and it is also well known that a proper selection of absences within the calibration dataset enhances such results (Zaniewski et al. 2002, Brotons et al. 2004, Engler et al. 2004, Elith and Leathwick 2007). In spite of this, species distributions are often modelled using either pseudoabsences extracted at random from the sites where the species has not been recorded, or methods based solely on presence data. Is such common practice correct?

Our results show that the kind of absences included in the calibration dataset conditions SDM results, as suggested before (Jiménez-Valverde et al. 2008a). The best hypotheses about the potential distribution are obtained using absences placed farther apart than the ones needed for the best hypotheses on the realized distribution (Chefaoui and Lobo 2008, Fig. 5A). Thus, it is of outmost importance that modellers decide the type of distributional hypothesis they are interested in modelling while designing their analyses, and use one or another type of data accordingly (Jiménez-Valverde et al. 2009b). However, the distributional hypotheses generated here are far from being accurate, in spite of the use of an adequate selection of absence data and the high evaluation scores obtained by the models (which would be considered very reliable according to SDM literature). Specially striking is the fact that, being the hypotheses on the realized distribution the ones that show the highest discrimination values, they are less accurate than the potential ones, yielding much higher overprediction rates (see also Jiménez-Valverde et al. 2008a). This phenomenon is a matter of the relative occurrence area (ROA, the ratio between the species extent of occurrence and the whole extent of the region of study; Lobo et al. 2008), and casts for some caution when comparing results between species that differ in the proportion of the total extent of the region they occupy (Jiménez-Valverde et al. 2008a, Lobo et al. 2008) or, like in this case, distributions that vary in size such as the potential and realized ones (see also Soberón and Nakamura 2009). Besides, obtaining reliable estimations of the realized distribution would require also taking into account variables that constrain the potential distribution and/or techniques able to fit interactions and more complex relationships between dependent and independent variables (Soberón 2007, 2010, Jiménez-Valverde et al. 2008a).

Another point that arises from our results is that the most distant absences (the ''naughty noughts'' sensu Austin and Meyers 1996) provide little information, yielding considerable overpredictions of both potential and realized distributions. Although there are contrasted opinions on the use of absence data outside the environmental domain known to be used by the species (Austin and Myers 1996, Thuiller et al. 2004), these works do not consider the key distinction between potential and realized distributions made here. Austin (2006) states that ''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''. Due to this, several authors have suggested to include in the calibration datasets absences from outside the climatic envelope determined by the presences, in order to either avoid including false absences (Zaniewski et al. 2002, Engler et al. 2004), or build estimates of the potential distribution (Lobo et al. 2006, Jiménez-Valverde et al. 2007). However, according to our results the inclusion of absences of the extreme type ''there are no elephants in Antarctica'' must be avoided. Although SDM calibrated using such kind of absences will often show high percentages of explained variance (Jiménez-Valverde and Lobo 2007a), they may lead to overestimated predictions because of an exaggerated distortion of the response functions (Austin and Meyers 1996, VanDerWal et al. 2009).

Perhaps the most striking of our results is the inadequacy of including the less distant absences within the calibration

dataset. These are more likely to include several false absences, in the places where unknown populations the species will be discovered as the survey process continues (Lobo et al. 2007, Hortal et al. 2008). In addition, the environmental closeness to the domain occupied by presences might also hamper the parameterization of the models, as shown by their low percentages of explained deviance and low sensitivity and specificity values.

All the problems identified above point out to the adequacy of including absences well distributed along the spatial gradient under consideration, but well outside the environmental domain where the presences lay. As commented before, the amount and the actual probability of a point with no information being an absence of any kind will depend on the ROA. Due to this, indiscriminate use of background or pseudo-absence data, randomly selected from the whole territory regardless on the environmental location of the presences and with an unknown level of error in their assignment (Ferrier and Watson 1997, Stockwell and Peters 1999, Zaniewski et al. 2002, Engler et al. 2004, Elith et al. 2006, Lobo et al. 2006, Lütolf et al. 2006, Pearce and Boyce 2006), remains a potentially unreliable and difficult-to-evaluate procedure. This will be especially true when the results of the models for several species that differ in their ROA are directly compared, given that the location of the estimated distribution within the potential-realized gradient defined by Jiménez-Valverde et al. (2008a) will vary from one species to another. Here, it is important to take into account that 1) the probability of selecting each one of the three types of absences depends on the considered extent (Fig. 3), and also that 2) the proportion of these types of absences used conditions the obtained geographic representation. Based on these facts, SDMs developed from calibration data where pseudoabsences are selected at random will render predictions that approximate either potential or realized distributions only as a consequence of differences in species' ROAs. Due to this, we argue that the widely used practice of selecting pseudoabsences merely at random should be seriously questioned.

In contrast, we suggest that the best way to obtain a reliable representation of the potential distribution of a species should be using absences located relatively near the external boundary of the environmental domain occupied by the presences. To model the realized distribution of the species, these absences should be located also relatively near to the known presences in the geographical space as well. However, more research is needed on this point (Soberón 2010). Nevertheless, the selection of the most appropriate absences is context dependent; the actual probability that a locality with no presence data pertains to any one of the three kinds of absences defined here will depend on the spatial extent under consideration and the size of the distribution range of the species being studied. Thus, applying the same procedure to obtain absences may have different effects depending on the species and the spatial scale of the study.

The inclusion of absences on a reliable way within model evaluation processes remains as an open issue. A proper evaluation of model results is basic to estimate the degree of confidence of the distributional hypotheses generated through SDM (Vaughan and Ormerod 2003, 2005). However, the most used evaluation measures, such as AUC or kappa, can yield high discrimination values (i.e. attributed to good-performing models) in cases when model predictions show high rates of commission and/or omission errors (Jiménez-Valverde et al. 2008a and Lobo et al. 2008, but also Raes and Ter Steege 2007). Moreover, similar AUC scores can be obtained with predictions of the distribution in the geographical space very different one from another. Hence, these measures do not provide reliable estimates of SDM performance. Rather, it is advisable to conduct separate analyses for commission and omission errors, in order to obtain a more accurate picture of the predictive behaviour of the models. Here, the most appropriate evaluation procedure will depend on the pursued purpose. In the case of the potential distribution, validation is only partially possible by examining the success in the prediction of presences in other spatial or time scenarios (as provided by e.g. biological invasions or fossil data, Sax et al. 2007 and Nogués-Bravo et al. 2008, respectively), and by examining the agreement among physiological and distributional data (Dormann 2007, Kearney et al. 2008). For the realized distribution, reliable absences are essential both for training and evaluation processes, being also highly recommendable to always estimate the degree of overprediction.

Concluding remarks

While presences are usually free of doubt about their reliability, absence data always have an associated degree of uncertainty. Confirmed absences are very difficult to obtain, and require higher levels of sampling effort to ensure their reliability than those required for the presences (Mackenzie and Royle 2005). The absences coming from lack of adequate survey effort (as is often the case) should be handled with caution, avoiding the indiscriminate inclusion of zeros from badly surveyed localities within the dataset used for SDM. Instead, in these cases absences should be selected by means of expert opinion and/or conceptual designs such as the one developed here. The compilation of exhaustive databases and the study of survey completeness can also help to identify the location of these methodological absences to some extent (Hortal and Lobo 2005). However, when the goal is to obtain a highly accurate description of the distribution of a particular species, additional fieldwork is quite likely to be needed, either for the empirical validation of SDM results or to confirm the absence of the species from some key localities. This is of particular importance for invasive species, where some of the absences of contingent origin can in turn become presences within short periods of time.

Apart from the need to identify methodological absences in order to avoid using false absences, we have shown that environmental and contingent absences can be of outmost importance for the study of species distributions. Environmental absences are required if the goal is to produce a hypothesis of the potential distribution of the species, (see above), and a few methods have been already proposed for such task (Engler et al. 2004, Lobo et al. 2006, Jiménez-Valverde and Lobo 2007b). However, as pointed out before, extremely distant and

uninformative absences should be discarded, and perhaps the way of deciding when to start discarding deserves further investigation. On the other hand, non-environmental or contingent absences are mandatory for the study of the realized distribution of the species. However, they are much more difficult to obtain, because their identification would require a previous knowledge on the areas with environmentally suitable conditions that are not inhabited by the species of interest (Jiménez-Valverde et al. 2008b). Hence, we suggest that a previous recognition of well-surveyed territories and a delimitation of the potential distribution can partially help in the recognition of these absences. These steps should be part of a continuous process, where the information obtained from previous steps (either through SDM, sampling effort assessment or additional field work) is used to inform the forthcoming ones, thus emphasizing the preliminary nature of the distributional hypotheses developed from species distribution modelling.

Acknowledgements – We thank Núria Roura, Darren Kriticos and Robin Engler for their useful comments and suggestions to this manuscript. AJ-V is supported by a MEC (Ministerio de Educación y Ciencia, Spain) postdoctoral fellowship (Ref.: EX-2007-0381), and JH by the U.K. Natural Environment Research Council.

References

Araújo, M. B. and Pearson, R. G. 2005. Equilibrium of species' distributions with climate. – Ecography 28: 693–695.

Austin, M. 2006. Species distribution models and ecological theory: a critical assessment and some possible new approaches. – Ecol. Model. 200: 1–19.

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

Austin, M. P. et al. 1990. Measurement of the realized qualitative niche: environmental niches of five *Eucalyptus* species. – Ecol. Monogr. 60: 161–177.

Basille, M. et al. 2008. Assessing habitat selection using multivariate statistics: some refinements of the ecological-niche factors analysis. – Ecol. Model. 211: 233–240.

Brotons, L. et al. 2004. Presence–absence versus presence–only based habitat suitability models for bird atlas data: the role of species ecology and prevalence. – Ecography 27: 285–298.

Brown, J. H. et al. 1996. The geographic range: size, shape, boundaries and internal structure. – Annu. Rev. Ecol. Syst. 27: 597–623.

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

Busby, J. R. 1986. A biogeographical analysis of *Notophagus cunninghamii* (Hook.) in south-eastern Australia. – Aust. J. Ecol. 11: 1–7.

Calenge, C. and Basille, M. 2008. A general framework for the statistical exploration of the ecological niche. – J. Theor. Biol. 252: 674–685.

Calenge, C. et al. 2008. The factorial decomposition of the Mahalanobis distances in habitat selection studies. – Ecology 89: 555–566.

Chefaoui, R. M. and Lobo, J. M. 2008. Assessing the effects of pseudo-absences on predictive distribution model performance. – Ecol. Model. 210: 478–486.

Chown, S. L. and Terblanche, J. S. 2007. Physiological diversity in insects: ecological and evolutionary contexts. – Adv. Insect Physiol. 33: 50–152.

Colwell, R. K. and Rangel, T. F. 2009. Hutchinson's duality: the once and future niche. – Proc. Nat. Acad. Sci. USA 106: 19651–19658.

Coope, G. R. 1990. The invasion of northern Europe during the Pleistocene by Mediterranean species of Coleoptera. – In: di Castri, F. et al. (eds), Biological invasions in Europe and the Mediterranean Basin. Kluwer, pp. 203–215.

Coope, G. R. and Angus, R. B. 1975. An ecological study of a temperate interlude in the middle of the Last Glaciation, based on fossil Coleoptera from Isleworth, Middlesex. – J. Anim. Ecol. 44: 365–391.

Dennis, R. L. H. and Thomas, C. D. 2000. Bias in butterfly distribution maps: the influence of hot spots and recorder's home range. – J. Insect Conserv. 4: 73–77.

Dennis, R. L. H. et al. 1999. Bias in butterfly distribution maps: the effects of sampling effort. – J. Insect Conserv. 3: 33–42.

Dormann, C. F. 2007. Promising the future? Global change projections of species distributions. – Basic Appl. Ecol. 8: 387–397.

Elith, J. and Leathwick, J. 2007. Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. – Divers. Distrib. 13: 265–275.

Elith, J. et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. – Ecography 29: 129–151.

Engler, R. et al. 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.

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

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

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

Freeman, E. 2008. PresenceAbsence: presence–absence model evaluation. Version 1.1.2. – R Foundation for Statistical Computing, Vienna, Austria.

Gaston, K. J. 1991. Body size and probability of description: the beetle fauna of Britain. – Ecol. Entomol. 16: 505–508.

Gaston, K. J. and Blackburn, T. M. 1994. Are newly described bird species small-bodied? – Biodivers. Lett. 2: 16–20.

Graham, C. H. et al. 2004. New developments in museum-based informatics and applications in biodiversity analysis. – Trends Ecol. Evol. 19: 497–503.

Grinnell, J. 1917. Field tests of theories concerning distributional control. – Am. Nat. 51: 115–128.

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

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

Harrell, F. E. J. 2001. Regression modelling strategies: with applications to linear models, logistic regression, and survival analysis. – Springer.

Hastie, T. J. 2008. gam: generalized additive models. R package ver. 1.0. – R Foundation for Statistical Computing, Vienna, Austria.

Hastie, T. J. and Tibshirani, R. J. 1990. Generalized additive models. – Chapman and Hall.
Hijmans, R. J. et al. 2005. Very high resolution interpolated climate surfaces for global land areas. – Int. J. Climatol. 25: 1965–1978.
Hirzel, A. H. and Le Lay, G. 2008. Habitat suitability modelling and niche theory. – J. Appl. Ecol. 45: 1272–1381.
Hirzel, A. H. et al. 2002. Ecological-niche factors analysis: how to compute habitat-suitability maps without absence data? – Ecology 83: 2027–2036.
Hirzel, A. H. et al. 2004. Biomapper 3.0. Laboratory for conservation biology. – Univ. of Lausanne, Lausanne.
Hortal, J. 2008. Uncertainty and the measurement of terrestrial biodiversity gradients. – J. Biogeogr. 35: 1355–1356.
Hortal, J. and Lobo, J. M. 2005. An ED-based protocol for the optimal sampling of biodiversity. – Biodivers. Conserv. 14: 2913–2947.
Hortal, J. et al. 2007. Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife (Canary Islands). – Conserv. Biol. 21: 853–863.
Hortal, J. et al. 2008. Historical bias in biodiversity inventories affects the observed realized niche of the species. – Oikos 117: 847–858.
Jiménez-Valverde, A. and Lobo, J. M. 2006. The ghost of unbalanced species distribution data in geographic predictive models. – Divers. Distrib. 12: 521–524.
Jiménez-Valverde, A. and Lobo, J. M. 2007a. Threshold criteria for conversion of probability of species presence to either-or presence–absence. – Acta Oecol. 31: 361–554.
Jiménez-Valverde, A. and Lobo, J. M. 2007b. Potential distribution of the endangered spider *Macrothele calpeiana* (Walckenaer, 1805) (Araneae, Hexathelidae) and the impact of climate warming. – Acta Zool. Sinica 53: 865–876.
Jiménez-Valverde, A. et al. 2007. Exploring the distribution of *Sterocorax* Ortuño, 1990 (Coleoptera, Carabidae) species in the Iberian Peninsula. – J. Biogeogr. 34: 1426–1438.
Jiménez-Valverde, A. et al. 2008a. Not as good as they seem: the importance of concepts in species distribution modelling. – Divers. Distrib. 14: 885–890.
Jiménez-Valverde, A. et al. 2008b. Challenging species distribution models: the case of *Maculinea nausithous* in the Iberian Peninsula. – Ann. Zool. Fenn. 45: 200–210.
Jiménez-Valverde, A. et al. 2009a. The effect of prevalence and its interaction with sample size on the reliability of species distribution models. – Community Ecol. 10: 196–205.
Jiménez-Valverde, A. et al. 2009b. Species distribution models do not account for abundance: the case of arthropods on Terceira Island. – Ann. Zool. Fenn. 46: 451–464.
Kadmon, R. et al. 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. – Ecol. Appl. 14: 401–413.
Kearney, M. et al. 2008. Modelling species distributions without using species distributions: the cane toad in Australia under current and future climates. – Ecography 31: 423–434.
Kriticos, D. J. and Leriche, A. 2010. The effects of climate data scale on fitting and projecting species niche models. – Ecography 33: 115–127.
Legendre, P. and Legendre, L. 1998. Numerical ecology, 2nd ed. – Elsevier.
Löbl, I. and Smetana, A. 2006. Catalogue of Palaearctic Coleoptera, Vol. 3: Scarabaeoidea, Scirtoidea, Dascilloidea, Buprestoidea, Byrrhoidea. – Apollo Books.
Lobo, J. M. et al. 2006. Environmental and geographical factors affecting the Iberian distribution of flightless *Jekelius* species (Coleoptera: Geotrupidae). – Divers. Distrib. 12: 179–188.
Lobo, J. M. et al. 2007. How does the knowledge on the spatial distribution of species increase? – Divers. Distrib. 13: 772–780.
Lobo, J. M. et al. 2008. AUC: a misleading measure of the performance of predictive distribution models. – Global Ecol. Biogeogr. 17: 145–151.
Lütolf, M. et al. 2006. The ghost of past species occurrence: improving species distribution models for presence–only data. – J. Appl. Ecol. 43: 802–815.
Mackenzie, D. I. and Royle, A. 2005. Designing occupancy studies: general advice and allocating survey effort. – J. Appl. Ecol. 42: 1105–1114.
MacNally, R. 2002. Multiple regression and inference in ecology and conservation biology: further comments on retention of independent variables. – Biodivers. Conserv. 11: 1397–1401.
McCullagh, P. and Nelder, J. A. 1989. Generalized linear models. – Chapman and Hall.
McGlone, M. 1996. When history matters: scale, time, climate and tree diversity. – Global Ecol. Biogeogr. 5: 309–314.
Nogués-Bravo, D. et al. 2008. Climate change, humans and the extinction of the woolly mammoth. – PLoS Biol. 6: e79.
Özesmi, U. et al. 2006. Methodological issues in building, training, and testing artificial neural networks in ecological applications. – Ecol. Model. 195: 83–93.
Pearce, J. and Boyce, M. 2006. Modelling distribution and abundance with presence–only data. – J. Appl. Ecol. 43: 405–412.
Pulliam, H. R. 1988. Sources, sinks and population regulation. – Am. Nat. 132: 652–661.
Pulliam, H. R. 2000. On the relationship between niche and distribution. – Ecol. Lett. 3: 349–361.
R Development Core Team 2008. R: a language and environment for statistical computing. – R Foundation for Statistical Computing, Vienna, Austria, <www.R-project.org>.
Raes, N. and Ter Steege, H. 2007. A null-model for significance testing of presence–only species distribution models. – Ecography 30: 727–736.
Ricklefs, R. E. 2007. History and diversity: explorations at the intersection of ecology and evolution. – Am. Nat. 170: S56–S70.
Ricklefs, R. E. and Schluter, D. (eds) 1993. Species diversity in ecological communities. Historical and geographical perspectives. – Univ. Chicago Press.
Rotenberry, J. T. et al. 2006. GIS-based niche modelling for mapping species' habitat. – Ecology 87: 1458–1464.
Sax, D. F. et al. 2007. Ecological and evolutionary insights from species invasions. – Trends Ecol. Evol. 22: 465–471.
Segurado, P. and Araújo, M. B. 2004. An evaluation of methods for modelling species distributions. – J. Biogeogr. 31: 1555–1569.
Soberón, J. 2007. Grinnellian and Eltonian niches and geographic distributions of species. – Ecol. Lett. 10: 1115–1123.
Soberón, J. 2010. Niche and distributional range: a population ecology perspective. – Ecography 33: 159–167.
Soberón, J. and Peterson, T. 2004. Biodiversity informatics: managing and applying primary biodiversity data. – Phil. Trans. R. Soc. B 359: 689–698.
Soberón, J. and Peterson, A. T. 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. – Biodivers. Inform. 2: 1–10.
Soberón, J. and Nakamura, M. 2009. Niches and distributional areas: concepts, methods, and assumptions. – Proc. Nat. Acad. Sci. USA 106: 19644–19650.
Stockwell, D. R. B. and Peters, D. 1999. The GARP modelling system: problems and solutions to automated spatial prediction. – Int. J. Geogr. Inform. Sci. 13: 143–158.

Svenning, J. C. and Skov, F. 2004. Limited filling of the potential range in European tree species. – Ecol. Lett. 7: 565–573.
Svenning, J. C. and Skov, F. 2005. The relative roles of environment and history as controls of tree species composition and richness in Europe. – J. Biogeogr. 32: 1019–1033.
Thuiller, W. et al. 2004. Effects of restricting environmental range of data to project current and future species distributions. – Ecography 27: 165–172.
Tyre, A. J. et al. 2003. Improving precision and reducing bias in biological surveys by estimating false negative error rates in presence–absence data. – Ecol. Appl. 13: 1790–1801.
VanDerWal, J. et al. 2009. Selecting pseudo-absence data for presence–only distribution modeling: how far should you stray from what you know? – Ecol. Model. 220: 589–594.
Vaughan, I. P. and Ormerod, S. J. 2003. Improving the quality of distribution models for conservation by addressing shortcomings in the field collection of training data. – Conserv. Biol. 17: 1601–1611.
Vaughan, I. P. and Ormerod, S. J. 2005. The continuing challenges of testing species distribution models. – J. Appl. Ecol. 42: 720–730.
Venables, W. N. and Ripley, B. D. 2002. Modern applied statistics with S. – Springer.
Whittaker, R. J. et al. 2005. Conservation biogeography: assessment and prospect. – Divers. Distrib. 11: 3–23.
Wilson, R. J. et al. 2010. Linking habitat use to range expansion rates in fragmented landscapes: a metapopulation approach. – Ecography 33: 73–82.
Zaniewski, A. E. et al. 2002. Predicting species spatial distributions using presence–only data: a case study of native New Zealand ferns. – Ecol. Model. 157: 261–280.
Zar, J. H. 1999. Biostatistical analysis. – Prentice Hall.

Download the Supplementary material as file E6039 from <www.oikos.ekol.lu.se/appendix>.

Ecography

E6039

Lobo, J. M., Jiménez-Valverde, A. and Hortal, J. 2010. The uncertain nature of absences and their importance in species distribution modelling. – Ecography 33: 103–114.

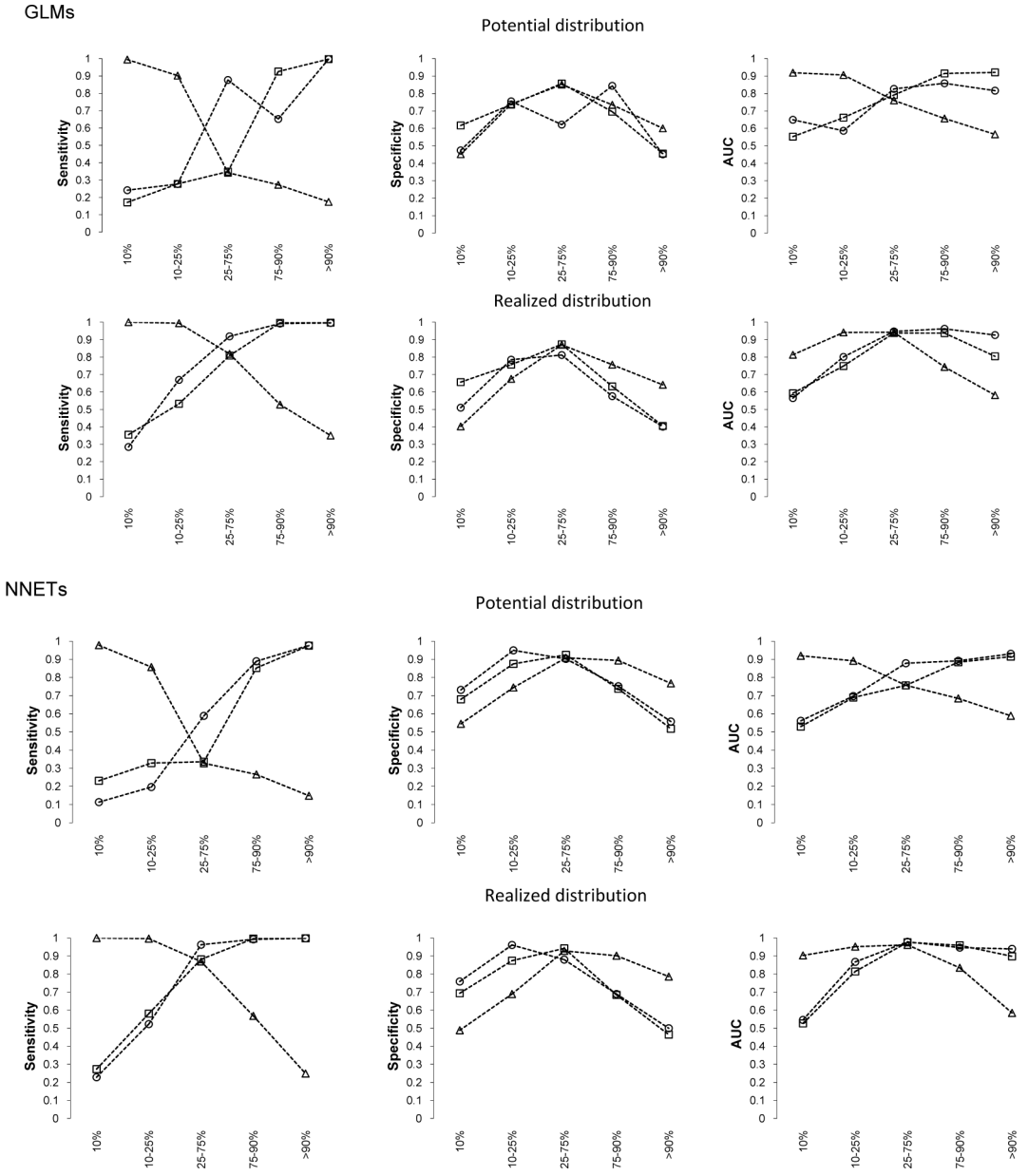

Figure S1. Variation in the rate of accurate predictions of presence (sensitivity), accurate predictions of absence (specificity) and AUC values according to the location and type of the absences used to model either the potential or the realized distribution of *Aphodius bonvouloiri* using GLMs or Artificial Neural Networks (NNETs). The used presences are the available observations of this species (Fig. 2), while ten times mores environmental (circles), contingent (squares) or methodological absences (triangles) were randomly selected among those present in five probability categories (Fig. 3): below 10% percentile, between 10% percentile and 25% quartile, between 25% and 75% quartile, between 75% quartile and 90% percentile, and above 90% percentile.