MORE COMPLEX DISTRIBUTION MODELS OR MORE REPRESENTATIVE DATA?

JORGE M. LOBO

Dpto. de Biodiversidad y Biología Evolutiva, Museo Nacional de Ciencias Naturales (CSIC), c/José Gutiérrez Abascal, 2 – 28006, Madrid, Spain. e-mail: mcnj117@mncn.csic.es

Abstract.— Distribution models for species are increasingly used to summarize species' geography in conservation analyses. These models use increasingly sophisticated modeling techniques, but often lack detailed examination of the quality of the biological occurrence data on which they are based. I analyze the results of the best comparative study of the performance of different modeling techniques, which used pseudo-absence data selected at random. I provide an example of variation in model accuracy depending on the type of absence information used, showing that good model predictions depend most critically on better biological data.

Key words.—distribution models, model reliability, pseudo-absences, conservation usefulness.

Recently, many efforts have focused on creation of models able to predict species' distributions from partial data. These distributional models use known distribution records of a species, as well as environmental and spatial explanatory variables, to build statistical functions for interpolating species' distributions across the environmental spectrum (Guisan & Zimmermann 2000). Models may also extrapolate species' distributions to sets of environmental conditions outside those used to build the models (Peterson 2003). The reliability of these predictions depends on many factors, but the three main ones are (1) quality of data used for model calibration, (2) predictive power of the explanatory variables, and (3) the modeling technique chosen to produce predictions from such variables. In spite of recent theoretical opinions on the errors that may result from the first two sources (Soberón & Peterson 2005, Barry & Elith 2006, Araújo & Guisan 2006), little effort has been devoted to testing experimentally the effects of these factors on the reliability of model outputs. Instead, much effort has been devoted to comparisons of the different available modeling techniques (Brotons et al. 2004, Segurado & Araújo 2004, Pearson et al. 2006, and references therein). Some recent papers (Drake et al. 2006, Elith et al. 2006) suggest that certain newly developed modeling techniques are better able to parametrize complex relationships, producing better distributional hypotheses for conservation purposes.

The study by Elith and collaborators (2006) drew especially interesting conclusions. This research paper is undoubtedly the best comparative study of the relative performance of different modeling techniques. Comparing the reliability of 16 techniques, and modeling 226 species from six world regions, the researchers validated the predicted distributions with "independent" and reliable species presence/absence data that were withheld from model building. As distribution models were derived from both presence and presence-pseudo absence data, with absences randomly distributed throughout the territory considered, results illustrate the potential of existing techniques applied to widely available information. This short paper is designed to illustrate the shortcomings of the model comparison approach in improving the results of predictive models of species' distributions: improvements in the biological occurrence data may provide more important advances than a more complex modeling approach.

ENOUGH ACCURACY FOR CONSERVATION?

Unfortunately, in the study by Elith and collaborators (Elith et al. 2006), maximum mean scores of the area under the Receiver Operating Characteristic curve (AUC), a measure of predictive accuracy, do not surpass 0.82 (mean AUC score around 0.70 for most species and types of models). Average AUC score for the modeling technique with the best predictions for all regions was 0.73. The average AUC score in the study of Drake and collaborators is similar (0.79).

Let us suppose that we obtain an AUC score of 0.82 for a model accomplished at a 100 x 100 m resolution in Switzerland (41.290 km^2 or 4.129.000 pixels), using for that 6000 presence points (similar conditions to those of the best model in the Elith et al. study). For an AUC score like this it is exceptional to obtain an outstanding specificity score of 0.99 (99% of absences correctly predicted), but even in this case 41,230 pixels were erroneously ascribed as presences (4.123.000 absence pixels x 0.01); a predicted area almost 7-fold larger (413 km^2) than the observed one (60 km^2). Hence, the usefulness for conservation of even the best models identified by these studies is questionable. Should we prioritize the use of such techniques, or search for others that are still more sophisticated?

AN EXAMPLE

Although biologists may know the places in which a species is unlikely to be observed (e.g., species not detected at a locality after intense sampling), such data are not usually published. Thus, despite the potential usefulness of relatively reliable absence data, such information is generally not available. Random selection of absences, a crude approach, may introduce an indeterminate number of false absences into models owing to the all-too-frequent sampling biases in biological information (see Dennis et al. 1999, Dennis & Thomas 2000, Zaniewski et al. 2002, Reutter et al. 2003, Graham et al. 2004, Martínez-Meyer 2005).

The influence of random selection of absences on distribution models was illustrated for a large Iberian dung beetle species (*Copris hispanus*) that occurs mainly in the southern half of the Iberian Peninsula. Using an exhaustive compilation of all available occurrence information regarding Iberian dung beetles (54 species, 15,924 database records), I first used accumulation curves to select reliably inventoried 50 x 50 km UTM cells. For each cell, I examined the number of species accumulated with the increase in the number of database records (an effective surrogate of the sampling effort carried out in each cell, see Hortal et al. 2006). Each curve was estimated 100 times, randomizing the entry order of the database records to smooth the curve, and subsequently fitted to the Clench function (Colwell & Coddington 1994, Soberón & Llorente 1993) to estimate the asymptotic value (i.e. the estimated total richness score for an unlimited number of samples). The adequately-inventoried UTM cells were defined as those with observed species richness of $\geq$80% of the asymptotic predicted scores. All of the 100 km^2 UTM cells belonging to the 2500 km^2 well-surveyed cells at which *C. hispanus* had not been detected were considered as true absences.

Forty-seven presence points and an equal number of absences were selected: (1) at random from all cells lacking presence or (2) from the cells considered to be true absences, and were modeled via a widely-accepted prediction technique (GAMs). The model used 9 climatic and lithological variables as predictors (total annual precipitation, rainfall during summer months, yearly mean temperature, minimum annual temperature, an aridity index, area with stony siliceous soils, calcareous soils, siliceous sediments and calcareous sediments). All models were repeated 10 times, and predictions were validated using information from 205 cells (158 presences and 47 sound absences) not used in model calibration. Beside the AUC scores for the validation data, percentages of presences and absences correctly predicted (sensitivity and specificity scores) were also calculated after applying to model probabilities the threshold that minimizes the difference between sensitivity and specificity (see Jiménez-Valverde & Lobo 2006 and 2007). Model predictions were obtained for both the entire Iberian Peninsula and only the southern half of the Peninsula to illustrate the effects of the geographic extent at which these types of models are developed on output probabilities and validation scores.

PREDICTING WITH SOUND ABSENCE DATA

Inclusion of reliable absence data significantly improved model predictions, especially for smaller territories with a less variable environment (Fig. 1). As anticipated, average AUC scores for the entire Iberian Peninsula (± 95% confidence intervals) using random absences are significantly lower (0.951 ± 0.011) than in the case of absences selected among well-surveyed cells (0.977 ± 0.007; $F_{(2,\ 27)}$ = 42.14, $p <$ 0.0001). Interestingly, this difference is still greater when only the southern

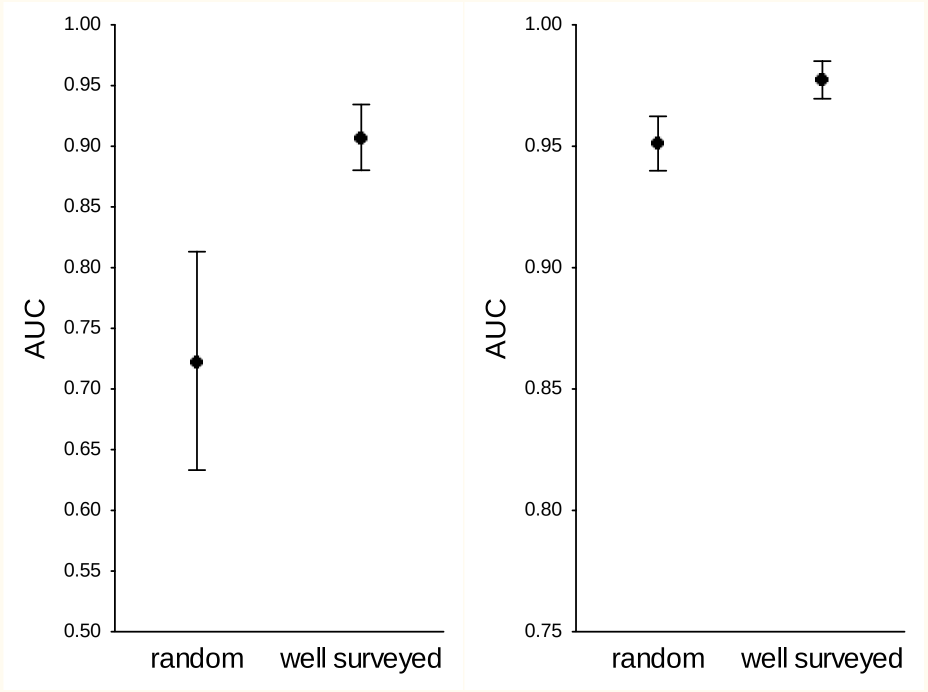

Figure 1. Mean AUC scores (± 95% confidence intervals) with respect to method of selection of absence information (at random or from among well-surveyed cells). Left: model based on southern half of the Iberian Peninsula; right: model based on the entire Iberian Peninsula.

half of the Iberian territory is considered: AUC scores derived from random absences are even lower and more variable (0.723 ± 0.090) than with well-surveyed cells (0.907 ± 0.027; $F_{(2, 27)} = 10.87$, $p = 0.0003$). Thus, the inclusion of reliable absence data significantly improves model predictions, especially for smaller ranges with less variable environments (see Fig. 2). Put another way, absences randomly distributed in a larger area lead to better predictions through reduced possibility of including false absences.

The study by Elith and collaborators (Elith et al. 2006) is, without a doubt, the most comprehensive of all to date. Unlike preceding studies, the authors validated model predictions with independent data and good absence information. What would have been the result if the reliable absence information used to validate their models had been used for calibration? Interestingly, when good data and predictors were used by Elith and collaborators (Elith et al. 2006; see, e.g., the case of Switzerland), accuracy differences among modeling methods seemed to diminish.

Unfortunately, most such modeling exercises do not use reliable absence information either to calibrate models or to validate them. This undesirable practice highly compromises the conservation usefulness of distribution model results. If one wants to generate a distributional simulation able to reflect the realized distribution of the species, good absence data need to be incorporated. These data should be located in climatically suitable localities in which the species does not occur due to historical factors, biotic interactions or dispersal limitation processes (Pulliam 1988, Ricklefts & Schluter 1993, Hanski 1998, Pulliam 2000). Including absences from *a priori* favorable environmental localities will inevitably diminish the predicted range size, so that the modeled distribution approaches the realized one (see Chefaoui & Lobo 2008). On the other hand, including absences from environmentally unsuitable places generates simulations which approach the potential distribution (all the environmentally suitable places in which a species could occur according to a group of environmental variables; see Soberón & Peterson 2005, Peterson 2006).

The probability of including false absences when absences are selected at random increases at

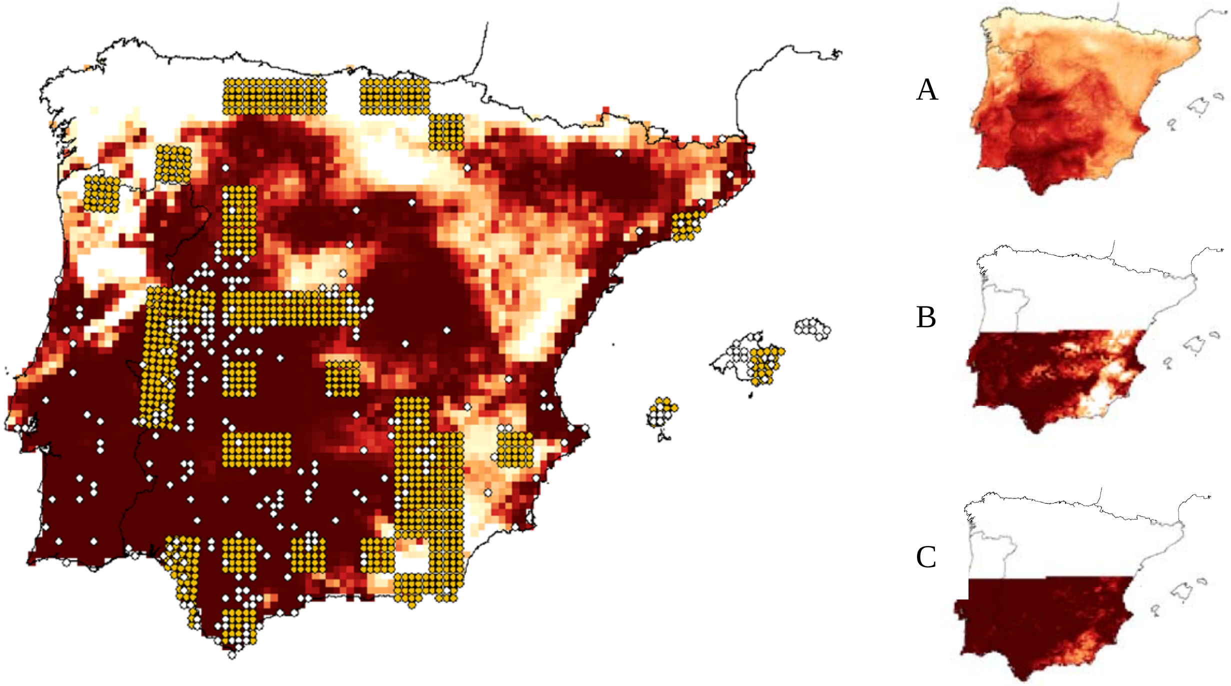

Figure 2. Left: GAM-derived distributional prediction for the Iberian dung beetle species *Copris hispanus* based on absence information derived from adequately-inventoried 50x50 km UTM cells (yellow squares). White dots represent known presence localities (100 km^2 UTM cells). The three figures at right are distributional predictions based on absences randomly selected from the whole Iberian peninsula (A); analyzing only the southern half of the Iberian Peninsula based on reliable (B) and randomly selected (C) absence information. The different shades represent probabilities from 0 (white) to 1 (dark red) that are averages of 10 replicate model predictions. Note that use of randomly selected absences generates higher probability scores in regions of absence.

smaller extents; at larger extents, it is more likely that random absence data are environmentally distant from the presence domain. Thus, the drawback of selecting random absences is higher when the ratio between the extent of species occurrence and the extent of the entire studied territory increases (the relative occurrence area). This point is exemplified by the model based on the southern half of the Iberian Peninsula (Fig. 2). For the same species, a model built at a smaller extent will produce inferior results if the absence data used are not reliable. As species modeled across the same region will frequently differ in relative occurrence area, the accuracy of models results cannot be compared among species (Lobo et al. 2008), particularly when random selection of absences implies the choice of a high number of false absences.

While recognizing the relevance of the search for improved modeling techniques, researchers must not forget that model prediction quality depends on data quality, and that species' absences input into such models should be as reliable as species' presences. Among other steps, collaboration between modelers and taxonomists in designing data selection, and databases compiling all available information can allow assessment of inventory completeness; both of these points offer strategies towards better distribution hypotheses for conservation purposes.

ACKNOWLEDGMENTS

This paper was supported by a Fundación BBVA project (Diseño de una red de reservas para la protección de la Biodiversidad en América Austral), and two MEC Projects (CGL2004-04309 and CGL2006-09567/BOS). Special thanks to Alberto Jiménez Valverde for his valuable comments

REFERENCES

Araújo, M. B. and A. Guisan. 2006. Five (or so) challenges for species distribution modelling. Journal of Biogeography 33:1677-1688.

Barry, S. and J. Elith. 2006. Error and uncertainty in habitat models. Journal of Applied Ecology 43:413-423.

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

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

Colwell, R. K. and J. A. Coddington. 1994. Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society of London B 345**:**101-118.

Dennis, R. L. H., T. H. Sparks, and P. Hardy. 1999. Bias in butterfly distribution maps: the effects of sampling effort. Journal of Insect Conservation 3:33-42.

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

Drake, J. M., C. Randin, and A. Guisan. 2006. Modelling ecological niches with support vector machines. Journal of Applied Ecology 43:424-432.

Elith, J., C. H. Graham, R. P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, J. Li, L. G. Lohmann, B. A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. M. Overton, A. T. Peterson, S. J. Phillips, K. Richardson, R. Scachetti-Pereira, R. E. Schapire, J. Soberón, S. Williams, M. S. Wisz, and N. E. Zimmermann. 2006. Novel methods improve prediction of species’ distributions from occurrence data. Ecography 29:129-151.

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

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

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

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

Hortal, J., Borges, P. A., and C. Gaspar. 2006. Evaluating the performance of species richness estimators: sensitivity to sample grain size. Journal of Animal Ecology 75**:**274-287.

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

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

Lobo, J. M., A. Jiménez-Valverde, and R. Real. 2008. AUC: A misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography 17: 145-151.

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

Pearson, R. G., W. Thuiller, M. B. Araújo, E. Martinez-Meyer, L. Brotons, C. McClean, L. Miles, P. Segurado, T. P. Dawson, and D. C. Lees. 2006. Model-based uncertainty in species range prediction. Journal of Biogeography 33:1704-1711.

Peterson, A. T. 2003. Predicting the geography of species’ invasions vias ecological niche modelling. Quarterly Review of Biology 78:419-433.

Peterson, A. T. 2006. Uses and requirements of ecological niche models and related distributional models. Biodiversity Informatics 3:59-72.

Pulliam, H. R. 1988. Sources, sinks and population regulation. American Naturalist 132**:**652-661.

Pulliam, H. R. 2000. On the relationship between niche and distribution. Ecology Letters 3**:**349-361.

Reutter, B. A., V. Helfer, A. H. Hirzel, and P. Vogel. 2003. Modelling habitat-suitability using museum collections: an example with three sympatric *Apodemus* species from the Alps. Journal of Biogeography 30:581–590.

Ricklefs, R. E. and D. Schluter. 1993. Species Diversity in Ecological Communities. Historical and Geographical Perspectives. University Chicago Press, Chicago.

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

Soberón, J. and J. Llorente. 1993. The use of species accumulation functions for the prediction of species richness. Conservation Biology 7:480-488.

Soberón, J. and A. T. Peterson. 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. Biodiversity Informatics 2:1-10.

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