Diversity and Distributions, (Diversity Distrib.) (2006) **12**, 521–524

BIODIVERSITY RESEARCH

The ghost of unbalanced species distribution data in geographical model predictions

A. Jiménez-Valverde* and J. M. Lobo

Departamento Biodiversidad y Biología Evolutiva, Museo Nacional de Ciencias Naturales (CSIC), c/José Gutiérrez Abascal 2, 28006 Madrid, Spain

*Correspondence: A. Jiménez-Valverde, Departamento Biodiversidad y Biología Evolutiva, Museo Nacional de Ciencias Naturales (CSIC), c/José Gutiérrez Abascal 2, 28006 Madrid, Spain. Tel.: +34 +914111328 ext, 1212, E-mail: mcnaj651@mncn.csic.es

ABSTRACT

Unbalanced samples are considered a drawback in predictive modelling of species' potential habitats, and a prevalence of 0.5 has been extensively recommended. We argue that unbalanced species distribution data are not such a problem from a statistical point of view, and that good models can be obtained provided that the right predictors and cut-off to convert probabilities into presence/absence are chosen. The effects of unbalanced prevalence should not be confused with those of low-quality data affected by false absences, low sample size, or unrepresentativeness of the environmental and spatial gradient. Finally, we point out the necessity of greater research effort aimed at improving both the quality of training data sets, and the processes of validating and testing of models.

Keywords

Species distribution models, unbalanced samples, predictive reliability, threshold cut-off.

INTRODUCTION

Species distribution modelling is now in wide use to develop analytical and prediction tools for ecology and conservation biology (Guisan & Zimmermann, 2000; Guisan & Thuiller, 2005), to locate previously unknown populations of rare and endangered species (Raxworthy *et al.*, 2003; Guisan *et al.*, 2006), to study the effect of climate warming on species distribution (Peterson, 2003; Thuiller *et al.*, 2005a), to assess the possible impact of biological invasions (Rouget *et al.*, 2004; Thuiller *et al.*, 2005b), and to aid management in taking decisions (Schadt *et al.*, 2002; Barbosa *et al.*, 2003; Russell *et al.*, 2004; Chefaoui *et al.*, 2005). Quantified species–environment relationships, obtained through the development of a mathematical function linking species distribution information (usually presence/absence) to environmental predictors, are used to map decimal fraction probabilities. These probabilities are usually taken as probabilities of presence and, so, as a measure of habitat adequacy. However, probability values are highly dependent on the relative proportion of each event in the sample, being biased toward the highest number of either presences or absences, where they differ. This inherent and unavoidable bias has long been recognized by statisticians under the name of the unbalanced sample effect (Hosmer & Lemeshow, 1989). This has some important consequences for the prediction of species distributions using models and has generated confused debate in the ecological literature that is not yet resolved.

Statistical effects of unbalanced samples

The influence of prevalence on the performance of model predictions has repeatedly been judged to be of major importance (Vaughan & Ormerod, 2003; McPherson *et al.*, 2004), leading to the supposition that the more unbalanced the samples, the less reliable the model predictions. In principle, there is no reason why the rarest events should necessarily be badly predicted, provided that models fit the data well (Cramer, 1999). Good fits can be obtained when good predictors are used and the dependent variable reflects all environmental variability. However, even in such circumstances, mean estimated probabilities of each event will be biased as a consequence of prevalence. This bias could be especially noticeable in the case of models that do not fit the data well (Cramer, 1999), typical of those derived from field studies where the most adequate predictors are usually unknown. This interaction between model fit and prevalence bias is a question that deserves further attention.

The apparently negative effect of prevalence on prediction reliability is mediated by the cut-off value selected to convert decimal fraction probabilities to a binary variable. This cut-off should be selected appropriately to account for unbalanced samples in the conversion of the decimal fraction probabilities to presence/absence, and to evaluate the model correctly when such measures as sensitivity, specificity, or the Kappa statistic, derived from a confusion matrix, are used (Fielding & Bell, 1997). As this conversion will determine model output, it will condition the

DOI: 10.1111/j.1366-9516.2006.00267.x

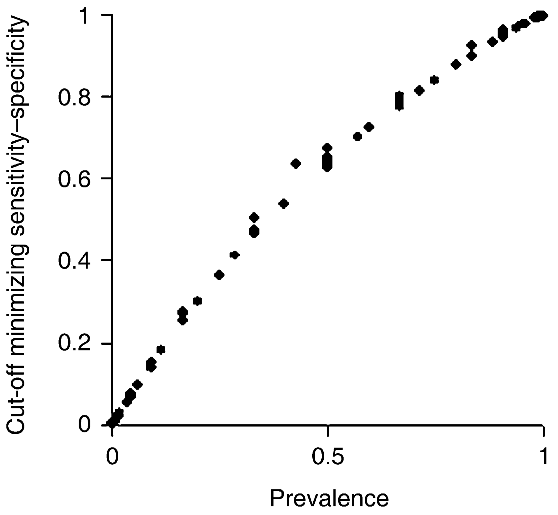

Figure 1 Relationship between the cut-off that minimized the sensitivity–specificity difference and prevalence, using data from a simulated species and from randomly resampling different training data sets varying in prevalence. Data were modelled using logistic regressions.

cases assigned to each of the four categories of the matrix (true and false predicted presence, true and false predicted absences). The intuitively appealing 0.5 cut-off (e.g. Li *et al*., 1997; Berg *et al*., 2004; Meggs *et al*., 2004) makes no sense, as each model has its own characteristics related to prevalence and fit. For example, in the case of rare species data, a 0.5 cut-off would convert presences to absences and would yield a false sensitivity value (true predicted presences) of zero in the most extreme case. In a recently published paper (Liu *et al*., 2005), the optimum cut-off is sought through comparison of numerous criteria. Therein, the fixed 0.5 cut-off, or the widely used one that maximizes the Kappa value, was found to be among those that produced the worst results. The best presence/absence models were derived from cut-offs that maximize the sum, or minimize the difference, between sensitivity and specificity (true predicted absences), among others. Interestingly, cut-offs selected by these two criteria are highly and positively correlated with prevalence. Figure 1 shows the relationship between prevalence and the cut-off which minimizes the difference between sensitivity and specificity (see also Fig. 5 in Manel *et al*., 2001), using data from a simulated species and randomly resampling different training data sets varying in prevalence. Data were modelled using logistic regressions. These results suggest that the prevalence value itself could be used as a cut-off (Liu *et al*., 2005), as formerly recognized and suggested by statisticians (Cramer, 1999).

CONFOUNDING FACTORS

Prevalence is a characteristic *of the data* that may sometimes correlate with species ecology, such as marginality, rarity, or specialization; these species are generally those of higher conservation concern. Bearing this in mind, caution must be exercised to avoid confusion between the effects of these biological attributes and their associated data problems and those of prevalence.

When threshold-independent accuracy measures, such as area under the Receiver Operating Characteristic curve (Swets, 1988), are used to validate predictive models, confusing results have been obtained, as in some cases low prevalence values are related to high AUC scores, while the inverse has been found in other studies (see, e.g. Brotons *et al*., 2004; Luoto *et al*., 2005). McPherson *et al.* (2004) found best AUC scores with prevalence values around 0.5. But, if as pointed out before, there is no sound reason for models to perform poorly with unbalanced samples, what do these results mean? Effects of poor quality data can be misunderstood as false prevalence effects. For example, performance of species distribution models could depend on the sampling size of each event (independently of their relative size) and on the representativeness of the training data (i.e. presences and absences must be evenly distributed across the environmental and geographical gradient; a low sample size for an event implies poor representativeness), independently of prevalence. Additionally, the inclusion of false absences is surely a confounding factor present in many data sets whose effect will interact with prevalence. Poor quality data are usually associated with rare species, as presences are usually scarce and absences are prone to contain a high proportion of false data.

Thus, the true effect of prevalence is probably negligible when building predictive distribution models, and its 'ghostly' effect is due to other puzzling factors. To avoid this supposed unbalanced-sample problem, some authors recommend resampling the training data to balance presences and absences (McPherson *et al*., 2004; Liu *et al*., 2005). However, in the case of reliable training data, resampling would yield only a loss of information, mainly in rare species with scarce reliable data, and should be avoided.

RESCALING PROBABILITIES

Finally, fitted probabilities from probability maps published in research papers, if considered indicative of habitat suitability, could be misleading. While potential probability may range from 0 to 1, probabilities that do not surpass a minimum value due to low prevalence could erroneously be interpreted as low, even for well-established populations. Although it could seem paradoxical, a low value of fitted probability may be assigned to a known presence event (Pontius & Batchu, 2003), given that an under-represented event is less likely to occur in any sampling universe. To adjust the representativeness of the obtained probabilities adequately, favourability functions, such as the one proposed by Real *et al*. (2006), should be used, whose outputs are independent of prevalence due to the elimination of the random probability element. These favourability functions can be considered to be rescaling functions, as they convert logistic probabilities (P) into favourability values (F), assigning a value of $F = 0.5$ to the predictor conditions for which $P =$ prevalence (Real *et al*., 2006). Interestingly, whereas P-values for different species are not comparable site to site because of the prevalence bias, this is not the case for F-values, which are directly equivalent (Real *et al*., 2006).

COROLLARY

In conclusion, low prevalence is a property of low probability events, not a problem to be solved. Its effects on predictive tools are well known and, once accounted for, rare events should be accurately predicted if predictors are powerful and training data are reliable (especially absences) and neither spatially nor environmentally biased. These considerations are of special relevance in conservation biology, as low prevalence is usually a property of data from endangered species. Greater research effort aimed at improving both the quality of training data sets (Vaughan & Ormerod, 2003) and the processes of validating, and testing of models (Vaughan & Ormerod, 2005) should be made.

ACKNOWLEDGEMENTS

This paper was supported by a Fundación BBVA project (Diseño de una red de reservas para la protección de la Biodiversidad …) and a MEC Project (CGL2004-04309). A.J.-V. was supported by a Museo Nacional de Ciencias Naturales/C.S.I.C./Comunidad de Madrid PhD grant.

REFERENCES

Barbosa, A.M., Real, R., Olivero, J. & Vargas, J.M. (2003) Otter (*Lutra lutra*) distribution modeling at two resolution scales suited to conservation planning in the Iberian Peninsula. *Biological Conservation*, **114**, 377–387.

Berg, Å., Gärdenfors, U. & von Proschwitz, T. (2004) Logistic regression models for predicting occurrence of terrestrial molluscs in southern Sweden — importance of environmental data quality and model complexity. *Ecography*, **27**, 83–93.

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.

Chefaoui, R.M., Hortal, J. & Lobo, J.M. (2005) Potential distribution modelling, niche characterization and conservation status assessment using GIS tools: a case study of Iberian *Copris* species. *Biological Conservation*, **122**, 327–338.

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

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.

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

Guisan, A. & Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8**, 993–1009.

Guisan, A., Broennimann, O., Engler, R., Yoccoz, N.G., Vust, M., Zimmermann, N.E. & Lehmann, A. (2006) Using niche-based models to improve the sampling of rare species. *Conservation Biology*, **20**, 501–511.

Hosmer, D.W. & Lemeshow, S. (1989) *Applied logistic regression.* Wiley, New York.

Li, W., Wang, Z., Ma, Z. & Tang, H. (1997) A regression model for the spatial distribution of red-crown crane in Yancheng Biosphere Reserve, China. *Ecological Modelling*, **103**, 115–121.

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.

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

Manel, S., Williams, H.C. & Ormerod, S.J. (2001) Evaluating presence-absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, **38**, 921–931.

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.

Meggs, J.M., Munks, S.A., Corkrey, R. & Richards, K. (2004) Development and evaluation of predictive habitat models to assist the conservation planning of a threatened lucanid beetle, *Hoplogonus simsoni*, in north-east Tasmania. *Biological Conservation*, **118**, 501–511.

Peterson, A.T. (2003) Projected climate change effects on Rocky Mountain and Great Plains birds: generalities of biodiversity consequences. *Global Change Biology*, **9**, 647–655.

Pontius, R.G. & Batchu, K. (2003) Using the relative operating characteristic to quantify certainty in prediction of location of land cover change in India. *Transactions in GIS*, **7**, 467–484.

Raxworthy, C.J., Martínez-Meyer, E., Horning, N., Nussbaum, R.A., Schbeider, G.E., Ortega-Huerta, M.A. & Peterson, A.T. (2003) Predicting distributions of known 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*, in press.

Rouget, M., Richardson, D.M., Nel, J.I., Le Maitre, D.C., Egoh, B. & Mgidi, T. (2004) Mapping the potential ranges of major plant invaders in South Africa, Lesotho and Swaziland using climatic suitability. *Diversity and Distributions*, **10**, 475–484.

Russell, K.R., Mabee, T.J. & Cole, M.B. (2004) Distribution and habitat of Columbia Torrent Salamanders at multiple spatial scales in managed forests of northwestern Oregon. *Journal of Wildlife Management*, **68**, 403–415.

Schadt, S., Revilla, E., Wiegand, T., Knauer, F., Kaczensky, P., Breitenmoser, U., Bufka, L., Cervený, J., Koubek, P., Huber, T., Stanisa, C. & Trepl, L. (2002) Assessing the suitability of central European landscapes for the reintroduction of Eurasian lynx. *Journal of Applied Ecology*, **39**, 189–203.

Swets, K. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.

Thuiller, W., Lavorel, S., Araújo, M.B., Sykes, M.T. & Prentice, I.C.

(2005a) Climate change threats to plant diversity in Europe. *Proceedings of the National Academy of Sciences USA*, **102**, 8245–8250.

Thuiller, W., Richardson, D.M., Pyšek, P., Midgley, G.F., Hughes, G.O. & Rouget, M. (2005b) Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. *Global Change Biology*, **11**, 2234–2250.

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.