Global Ecology and Biogeography, (Global Ecol. Biogeogr.) (2008) **17**, 145–151

ECOLOGICAL SOUNDING

AUC: a misleading measure of the performance of predictive distribution models

Jorge M. Lobo[1]*, Alberto Jiménez-Valverde[1] and Raimundo Real[2]

[1]*Departamento de Biodiversidad y Biología Evolutiva, Museo Nacional de Ciencias Naturales (CSIC), Madrid, Spain,* [2]*Laboratorio de Biogeografía, Diversidad y Conservación, Departamento de Biología Animal, Facultad de Ciencias, Universidad de Málaga, Spain*

*Correspondence: Jorge M. Lobo, Departamento de Biodiversidad y Biología Evolutiva, Museo Nacional de Ciencias Naturales (CSIC), Madrid, Spain.
E-mail: mcnj117@mncn.csic.es

ABSTRACT

The area under the receiver operating characteristic (ROC) curve, known as the AUC, is currently considered to be the standard method to assess the accuracy of predictive distribution models. It avoids the supposed subjectivity in the threshold selection process, when continuous probability derived scores are converted to a binary presence–absence variable, by summarizing overall model performance over all possible thresholds. In this manuscript we review some of the features of this measure and bring into question its reliability as a comparative measure of accuracy between model results. We do not recommend using AUC for five reasons: (1) it ignores the predicted probability values and the goodness-of-fit of the model; (2) it summarises the test performance over regions of the ROC space in which one would rarely operate; (3) it weights omission and commission errors equally; (4) it does not give information about the spatial distribution of model errors; and, most importantly, (5) the total extent to which models are carried out highly influences the rate of well-predicted absences and the AUC scores.

Keywords
AUC, distribution models, ecological statistics, goodness-of-fit, model accuracy, ROC curve.

INTRODUCTION

The area under the receiver operating characteristic (ROC) curve, known as the AUC, is widely used to estimate the predictive accuracy of distributional models derived from presence–absence species data. As the output of the different modelling techniques that use binary data as dependent variables produces continuous probabilities of presence (P), where P and $1 - P$ represent the degree to which each case is a member of one of the two events, a threshold is needed to predict class membership. Thus, the cases above this threshold would be predicted as presences, and the remaining cases would be absences. Comparing these binary transformed probabilities with the validation presence–absence data set enables the estimation of four different fractions in a two-by-two confusion matrix: the correctly predicted positive fraction or sensitivity; the correctly predicted negative fraction or specificity; the falsely predicted positive fraction (commission errors); and the falsely predicted negative fraction (omission errors). These four scores, and other measures of accuracy derived from the confusion matrix, such as the proportion of correct predictions (correct classification rate) and Cohen's kappa (Cohen, 1960), all depend on the discrimination threshold. In order to overcome the supposed subjectivity in the threshold selection process, the ROC curve plots sensitivity as a function of commission error (1 – specificity) as the threshold changes. The calculation of the area under this curve (the AUC score) provides a single-number discrimination measure across all possible ranges of thresholds. This discrimination measure is equivalent to the non-parametric Wilcoxon test (Hanley & McNeil, 1982), in which the rank of all possible pairs for presence and absence assigned probabilities is compared.

ROC curves were developed during World War II to assess the performance of radar receivers in signal detection (to estimate the trade-off between hit rates and false alarm rates), and were subsequently adopted in biomedical applications, mainly for comparing the performance of diagnostic tests (Pepe, 2000). In spite of its wide use and its generally good performance (Bradley, 1997), a lot of research effort has recently been devoted to the calculation of AUC score variations. This is being done to provide a measure of variance or to estimate the AUC's statistical significance (Provost & Fawcett, 2001; Fawcett, 2004; Schröder, 2004; Ferri *et al.*, 2005; Forman & Cohen, 2005). Since its first proposal as an appropriate method to estimate the accuracy of species distribution models (Fielding & Bell, 1997), many studies

DOI: 10.1111/j.1466-8238.2007.00358.x
www.blackwellpublishing.com/geb

have recommended its use in this field of research (Pearce & Ferrier, 2000; Manel *et al.*, 2001; McPherson *et al.*, 2004, among many others). However, some authors have begun to criticize the indiscriminate use of AUC as the standard measurement of accuracy in distribution models (Termansen *et al.*, 2006; Austin, 2007). In particular, Austin (2007) warns that 'reliance on AUC as a sufficient test of model success needs to be re-examined'. Agreeing with this concern, we examined some of the characteristics of this measure that question its reliability as a comparative measure of accuracy between model results. We also evaluated its general usefulness in distribution predictive modelling.

DRAWBACKS OF AUC AS A MEASURE OF OBSERVATION–PREDICTION FIT IN SPATIAL DISTRIBUTION MODELLING

There are several recognized features of the ROC curve that prevent its use as a measure of model accuracy. Firstly, AUC scores ignore the actual probability values, being insensitive to transformations of the predicted probabilities that preserve their ranks (Ferri *et al.*, 2005). This could be an advantage as it makes possible the comparison of tests that yield numerical results on different measurement scales (Pepe, 2000). However, in this way, transformations for species occurrence probabilities, such as those proposed by Real *et al.* (2006), may dramatically change the prediction output but do not have any effect on the AUC score. AUC is a discrimination index that represents the likelihood that a presence will have a higher predicted value than an absence (Hosmer & Lemeshow, 2000, p. 162), regardless of the goodness-of-fit of the predictions (Vaughan & Ormerod, 2005; Quiñonero-Candela *et al.*, 2006; Reineking & Schröder, 2006). Therefore, it is possible that a poorly fitted model (overestimating or underestimating all the predictions) has a good discrimination power (Hosmer & Lemeshow, 2000, p. 163). It is also possible that a well-fitted model has poor discrimination, if probabilities for presences are only moderately higher than those for absences, for example. Hosmer & Lemeshow (1980) and Lemeshow & Hosmer (1982) proposed testing the fitness of a model according to the predicted probabilities (see revisions of this approach in Graubard *et al.*, 1997, and in Hosmer *et al.*, 1997). They tested whether the proportion of presences in different ranges of the predictor values corresponded to the predicted probabilities. A probability value of 0.3, for instance, does not predict that all sites with this value will be absences (or presences if a lower threshold is selected), but that 30% of sites with this value will support the species. If, for example, the probability of presence increases from a mean value of 0.4 in half of the territory to a mean value of 0.6 in the other half, and a goodness-of-fit test shows that data fit the predictions, the AUC value will nevertheless be 0.6 (remarkably low), meaning low discrimination but not low accuracy.

A second weakness of ROC plots is that they summarize test performance over regions of the ROC space in which one would rarely operate (see Baker & Pinsky, 2001). The ROC curve has been recommended because it summarises model performance over all conditions a model could operate in (Swets, 1988), using all the information provided by the predictive model (Fielding & Bell, 1997). However, researchers will rarely be interested in all possible situations, but rather in one or a few of them. For example, extreme right and left sides of the ROC space are generally useless, as they correspond to high false-positive and high false-negative rates, respectively (Baker & Pinsky, 2001). On the contrary, if we were interested in maximizing correctly predicted positives and commission errors were unimportant, then the central and left areas of the curve would be valueless. Partial ROC curves have been proposed as an alternative to entire ROC curves (Thompson & Zucchini, 1989; Baker & Pinsky, 2001), but the partial AUC does not avoid any of the remaining drawbacks pointed out in this contribution.

Third, and related to the second point, AUC weights omission and commission errors equally, while in many applications of distribution modelling, omission and commission errors may not have the same importance (Fielding & Bell, 1997; Peterson, 2006). For example, from a reserve-design point of view, misclassifications of absences (commission errors) must be regarded as a more serious drawback than the opposite. On the other hand, low omission errors are desirable when searching for new species or populations (see Peterson, 2006). In fact, some modelling techniques based only on known presences and focused on simulating the potential, not realized, distribution of the species explicitly weight omission errors more strongly than commission errors (Anderson *et al.*, 2003). When misclassification costs are unequal, summarising over all possible threshold values is flawed (Adams & Hand, 1999). If two ROC curves belonging to different models cross, we could decide that the one with the highest AUC value is the best although, perhaps, the other can be the best for our cost ratio decision criteria (Adams & Hand, 1999). Visual inspection of the complete ROC curves instead of considering only the numerical AUC values could be a better strategy in this case. Anyway, cost assignments to false presences and absences are always subjective, and the simplest way of doing this is to change the probability threshold above which presence is accepted until the desired rate between commission and omission errors is achieved. The independent examination of the percentage of presence and absence errors helps to select the best model according to the researcher's goals, rather than the use of a synthetic measure such as the AUC.

Moreover, in contrast to biomedical applications where positive or negative events are clearer (a patient has cancer or not, with little doubt), when recording presence–absence data of a species, absences have a higher degree of uncertainty than presences. Apparent absences may be due, simply, to low detectability of the species, or may correspond to non-sampled areas. Because of this, false absences are more likely to occur than false presences and, consequently, commission errors should not weigh as much as omission errors. A compound discrimination measure such as AUC could then be misleading. This fact is of special concern given the extended use of background data as pseudo-absences, i.e., using randomly selected sites where the species has not been reported as absences in the training process (Elith *et al.*, 2006), a procedure that inflates the number of false absences.

A fourth weakness is that ROC plots do not provide information about the spatial distribution of model errors (Pontius & Schneider, 2001), since it is impossible to decide if the biases are homogeneously distributed across the modelled territory, or if the lack of discrimination is due to the incapacity to correctly predict a specific region. Ignorance of the spatial distribution of errors is a common drawback to all single-number measures of accuracy, such as sensitivity, specificity and other measures derived from contingency tables. Researchers usually ignore the spatial location of errors, but their spatial arrangement can give interesting clues about ecological and biogeographical questions, and can change the relative weight of those errors (i.e., a model with randomly distributed errors is not the same as a model with spatially aggregated errors, the latter probably indicating the role of unaccounted for spatially structured variables).

Fifth, species distribution data are referred to a concrete geographical extent, and increasing the geographical extent outside presence environmental domain entails obtaining higher AUC scores. This feature decisively prevents the use of AUC for testing accuracy in predictive distribution modelling. For example, in a clinical investigation, the purpose may be to estimate if a drug has a significant effect on a sample of patients compared with a control population. In this case, avoiding a bias in the control population (for example, people with a special resistance to the disease) is crucial to the experimental design. In species distribution research, it has been proposed that 'invented' absences from environmentally distant areas be used as pseudo-absences (Engler *et al.*, 2004; Lobo *et al.*, 2006). In this case, using pseudo-absences more environmentally distant from the presences increases the rate of well-predicted absences and the AUC scores. This would be analogous to selecting a control sample in a clinical study comprising people gradually more distant from the conditions that cause the incidence of the disease (for example, selecting the resistant population). The more environmentally distant the absences, the better they will be predicted even with a bad model. A model that overpredicts presences will have a low commission error if the number of absences is much higher than that of presences as a consequence of increasing the extent of the study area. If, for example, 10 actual presences are compared to 90 absences and a model predicts presences in 20 sites (an overprediction of 100%), the sensitivity will be 1, the specificity will be 0.89, and the AUC will be interpreted as outstanding. Thus, high sensitivity, specificity and AUC scores can be artificially obtained by simply increasing the extent of the territory. Consequently, the accuracy of different models for the same species should not be assessed using AUC if they differ in the total extent analysed.

On the other hand, different species usually have distinct ratios between the extent of occurrence and the whole extent of the territory under study, i.e. they differ in their relative occurrence area. The smaller this ratio, the higher the number of absences and the more likely it is that absence data are environmentally distant from the presence domain. This is probably the reason why rare species are usually 'better predicted' than widespread ones (e.g. Brotons *et al.*, 2004; Arntzen, 2006; Hernández *et al.*, 2006; McPherson & Jetz, 2007); it is a pure and inevitable methodological question. Consequently, AUC values cannot be used to compare model accuracy between species when they differ in their relative area of occurrence. In the same way, similar values of specificity and AUC may imply large differences in the degree of overprediction. A 5% commission error is not equivalent for a 'common' and a 'rare' species. For the latter, this commission error can imply a several fold increase in the distributional area of the species.

The dependence of the AUC on the choice of geographical extent has been outlined by other authors in papers not specifically devoted to this question (Wiley *et al.*, 2003; Termansen *et al.*, 2006), but the deep implications of this effect continue without being considered. Both AUC and specificity scores depend on the relative occurrence area, which implies that between-species model accuracy comparisons should not be approached using these measures.

CONVERSION TO BINARY MAPS: IS THE THRESHOLD CRITERION SUBJECTIVE?

The main argument in favour of ROC curves is that they do not depend on a threshold, whose selection is thought to be subjective. However, this argument is questionable, since the criteria to select the best threshold to convert continuous predicted values to binary predictions have been recently improved. The logistic function, for instance, is symmetric with the inflection point located at the 0.5 probability value (Real *et al.*, 2006). So, when the training data contain the same number of presences and absences, 0.5 is the obvious and correct threshold if the same cost is assigned to commission and omission errors. However, when any of the two events is higher than the other, mean probabilities are biased towards the most common event (Hosmer & Lemeshow, 1980; Cramer, 1999). This effect is inevitable because logistic probabilities are computed based on the values of the predictors as well as on the relative proportion of presences and absences (Real *et al.*, 2006). When this happens, the 0.5 threshold is incorrect, and a prevalence-dependent threshold is needed (Jiménez-Valverde & Lobo, 2006; Jiménez-Valverde & Lobo, 2007). The supposed ambiguity in the threshold selection process is produced by this inevitable and well-known mathematical effect. Simply, the threshold must be adjusted to the prevalence of the training data.

Paradoxically, once the AUC score was developed as a threshold-independent measure, researchers proposed methods for selecting a threshold from this curve. It has been assumed that in ROC plots the optimal classifier point is the one that maximizes the sum of sensitivity and specificity (Zweig & Campbell, 1993). However, Jiménez-Valverde & Lobo (2007) have found that a threshold that minimizes the difference between sensitivity and specificity performs slightly better than one that maximizes the sum if commission and omission errors are equally costly. When the threshold changes from 0 to 1, the rate of well-predicted presences diminishes while the rate of well-predicted absences increases. The point where both curves cross can be considered the appropriate threshold if both types of errors are equally weighted (Fig. 1a). In a ROC plot, this point lies at the intersection of the ROC curve and the line perpendicular to the diagonal of

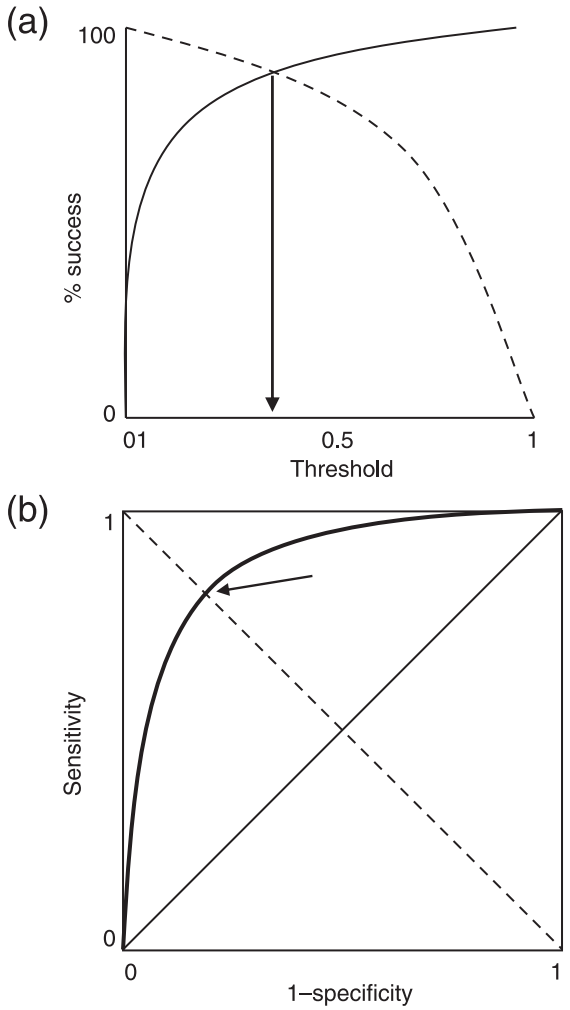

Figure 1 (a) Variation in the percentage of success in the prediction of presences (continuous line) and absences (broken line) with the change in the threshold used to discriminate both states from a continuous probability variable. The arrow represents the threshold that minimizes the difference between sensitivity and specificity. (b) ROC plot in which an arrow shows the 'most north-western' point.

no discrimination (Fig. 1b), i.e., the 'northwesternmost' point of the ROC curve. The two thresholds can be easily computed without using the ROC curve. Both thresholds are highly correlated and, more importantly, they also correlate with prevalence (Liu *et al.*, 2005; Jiménez-Valverde & Lobo, 2007). As a general rule, a good classifier needs to minimize the false positive and negative rates or, similarly, to maximize the true negative and positive rates. Thus, if we place equal weight on presences and absences there is only one correct threshold. This optimal threshold, the one that minimizes the difference between sensitivity and specificity, achieves this objective and provides a balanced trade-off between commission and omission errors. Nevertheless, as pointed out before, if different costs are assigned to false negatives and false positives, and the prevalence bias is always taken into account, the threshold should be selected according to the required criteria. It is also necessary to underline that the transformation of continuous probabilities into binary maps is frequently necessary for many practical applications that rely on making decisions (e.g., reserve selection).

CONTINUOUS PROBABILITY MAPS

AUC is commonly used in distribution modelling literature, even when discrimination capacity is not the main objective. We could be interested in the continuous probability map (Vaughan & Ormerod, 2003), as it reveals the whole gradient in habitat suitability. In this case, the selection of an appropriate threshold would be of concern only for measuring discrimination capacity. We need to consider that, because the mean probability bias is due to unbalanced prevalence, raw probability scores do not reflect habitat suitability (Jiménez-Valverde & Lobo, 2006; Real *et al.*, 2006). It is necessary to rescale these probabilities if we want to use continuous and not binary predictions (Jiménez-Valverde & Lobo, 2006; Real *et al.* 2006). Then, after the proper rescaling, not only a measure of discrimination capacity is necessary, but also a measure of habitat suitability accuracy. This is an interesting challenge in distribution modelling: do probability scores reflect real adequacy? The ROC curve says nothing about this, so a high AUC score does not imply suitability accuracy. Researchers must avoid the indiscriminate use of the AUC and reconsider the evaluation procedure of their models in accordance with their goal.

THE APPROPRIATE USE OF AUC IN DISTRIBUTION MODEL ASSESSMENT

Despite widespread use in several research fields, AUC has serious drawbacks when applied to species distribution modelling. It was rapidly adopted by presence–absence modellers as a measure of discrimination due to its good performance in other research areas and its ease in understanding and computation. However, apart from the flaws common to all disciplines, the uncertainty of absences and, mainly, the spatial dimensions of distributions are specific characteristics of species data that prevent the use of AUC in distribution modelling. The real value of AUC is that it provides a measure of the degree to which a species is restricted to a part of the variation range of the modelled predictors, so that presences can be told apart from absences. If a species is widespread and the probability of presence increases steadily with predictor values, an accurate model will have low AUC values, which will only denote the true generalist nature of the species distribution. In conclusion, AUC provides information about the generalist or restricted distribution of a species along the range of predictor conditions in the study area, but it does not provide information about the good performance of the model.

WHERE TO FROM HERE?

The overall agreement between the observed training data and the model output scores is guaranteed by the modelling methods. Modelling techniques are based on an inductive process, and inductive models should be assessed on the basis of their induction rules. The error derived from the incorrect application of the induction rules could be discarded if the software used is

reliable. Different methods could be compared, however, according to the assumptions implicit in the induction rules. Some profile methods, for instance, assume that the species is equally likely to occur within the environmental range of the presences, whereas GLMs tend to define a function of gradual probability variation according to the environmental conditions.

Nevertheless, even inductive models may fail to reflect the training data in specific parts of the territory analysed. The display of the deviance between observations and predictions (the subtraction of the output score from 1 for presences and from 0 for absences) throughout the territory may reflect the areas where commission error (negative deviance) and omission error (positive deviance) aggregate spatially.

Model distributional simulations can be confirmed using new information of the same type used to calibrate it (generally presence/absence data), but must always consider that the percentages of correct presences and absences depend on the relative occurrence area. The accuracy of distribution models can also be estimated by examining the consistency of model predictions with new observational data of a type different from that used to calibrate the models. The results of a field survey in a portion of the territory could be used to evaluate, for example, if the model scores correlate with the species abundance, density of breeding pairs, or productivity, which could also help in the interpretation of the biogeographical meaning of the results.

Strictly speaking, only closed models can be validated (Oreskes *et al.*, 1994). Species distribution models are incomplete approximations because the distribution of a species is always influenced by an unknown number of non-independent factors that interact spatially in an unknown manner. In this case, the agreement between simulated and observational data implies the calculation of the accuracy of the model within a confirmatory iterative procedure in which the result of each model should be viewed as a unique 'distributional hypothesis' limited to the used predictors, and the extent and location of the considered region.

The relevance of omission and commission errors depends on the modelling purpose. If we want to generate a distributional simulation able to reflect all the environmentally suitable places in which a species can occur according to a group of environmental variables (the potential distribution), we need to use profile techniques that only use presence data, or discrimination techniques that use 'invented' absences outside the environmental domain of presences. In this way we will not incorporate absence data from climatically suitable localities in which the species does not occur due to historical factors, biotic interactions or dispersal limitation processes (Ricklefs & Schluter, 1993; Hanski, 1998; Pulliam, 1988; Pulliam, 2000). Including absences from *a priori* favourable environmental localities inevitably diminishes the predicted range size approaching the simulation to the realized distribution. We claim that the current distribution of species should be modelled incorporating such distribution restriction forces and using reliable presence and absence data, and that these geographical representations need to be confirmed using reliable distributional information, always considering that model predictions are reliant on the conditions in which the model has been carried out (the predictors used, the quality and number of data points of the dependent variable, as well as the extent and resolution considered).

Accuracy measures proposed in the literature (Fielding & Bell, 1997) can be used to compare techniques for the same species at the same extent. In this case, instead of using only the AUC, we propose that sensitivity and specificity should be also reported, so that the relative importance of commission and omission errors can be considered to assess the method performance. Unfortunately, we cannot recommend any useful method to compare model performance among species. To the same extent, species unavoidably differ in range sizes and, therefore, in their relative areas of occurrence. Thus, the extent-dependence of model accuracy measures hinders their use as a means to compare model performance between species.

ACKNOWLEDGEMENTS

Boris Schröder clarified some points about the ROC curve and different thresholds. A. T. Peterson and two anonymous referees gave valuable comments on the manuscript. 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).

REFERENCES

Adams, N.M. & Hand, D.J. (1999) Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition*, **32**, 1139–1147.

Anderson, R.P., Lew, D. & Peterson, A.T. (2003) Evaluating predictive models of species' distributions: criteria for selecting optimal models. *Ecological Modelling*, **162**, 211–232.

Arntzen, J.W. (2006) From descriptive to predictive distribution models: a working example with Iberian amphibians and reptiles. *Frontiers in Zoology*, **3**, 8.

Austin, M. (2007) Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological Modelling*, **200**, 1–19.

Baker, S. & Pinsky, P. (2001) A proposed design and analysis for comparing digital and analog mammography: special receiver operating characteristic methods for cancer screening. *Journal of the American Statistical Association*, **96**, 421–428.

Bradley, A.P. (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, **30**, 1145–1159.

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.

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **41**, 687–699.

Cramer, J.S. (1999) Predictive performance of the binary logit model in unbalanced samples. *The Statistician*, **48**, 85–94.

Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G.,

Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S., Wisz, M.S. & Zimmermann, N.E. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

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

Fawcett, T. (2004) *ROC graphs: notes and practical considerations for data mining researchers*. Kluwer Academic Publishers, Netherlands. Available at http://home.comcast.net/~tom.fawcett/public_html/papers/ROC101.pdf

Ferri, C., Flach, P., Hernández-Orallo, J. & Senad, A. (2005) Modifying ROC curves to incorporate predicted probabilities. *Proceedings of the Second Workshop on ROC Analysis in Machine Learning* (ed. by N. Lachiche, C. Ferri, S. Macskassy and A. Rakotomamonjy), pp. 33–40. International Conference on Machine Learning, Bonn, Germany.

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.

Forman, G. & Cohen, I. (2005) Beware the null hypothesis: critical value tables for evaluating classifiers. *Lecture Notes in Computer Science*, **3720**, 133–145.

Graubard, B.I., Korn, E.L. & Midthune, D. (1997) Testing goodness-of-fit for logistic regression with survey data. *Proceedings of the American Statistical Association*, available at http://www.amstat.org/sections/srms/proceedings/y1997.html.

Hanley, J.A. & McNeil, B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.

Hanski, I. (1998) Metapopulation dynamics. *Nature*, **396**, 41–49.

Hernández, P.A., Graham, C.H., Master, L.L. & Albert, D.L. (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, **29**, 773–785.

Hosmer, D.W. & Lemeshow, S. (1980) A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics*, **10**, 1043–1069.

Hosmer, D.W. & Lemeshow, S. (2000) *Applied logistic regression*, 2nd edn. John Wiley & Sons, New York.

Hosmer, D.W., Hosmer, T., le Cessie, S. & Lemeshow, S. (1997) A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, **16**, 965–980.

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

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

Lemeshow, S. & Hosmer, D.W. (1982) A review of goodness of fit statistics for use in the development of logistic regression models. *America Journal of Epidemiology*, **115**, 92–106.

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

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

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. (2007) Effects of species' ecology on the accuracy of distribution models. *Ecography*, **30**, 135–151.

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.

Oreskes, N., Shrader-Frechette, K. & Belitz, K. (1994) Verification, validation and confirmation of numerical models in the earth sciences. *Science*, **263**, 641–646.

Pearce, J. & Ferrier, S. (2000) Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, **133**, 225–245.

Pepe, M.S. (2000) Receiver operating characteristic methodology. *Journal of the American Statistical Association*, **95**, 308–311.

Peterson, A.T. (2006) Uses and requirements of ecological niche models and related distributional models. *Biodiversity Informatics*, **3**, 59–72.

Pontius R.G., Jr & Schneider, L.C. (2001) Land-cover change model validation by an ROC method for the Ipswich watershed, Massachusetts, USA. *Agriculture, Ecosystems and Environment*, **85**, 239–248

Provost, F. & Fawcett, T. (2001) Robust classification for imprecise environments. *Machine Learning*, **42**, 203–231.

Pulliam, H.R. (1988) Sources, sinks and population regulation. *The American Naturalist*, **132**, 652–661.

Pulliam, H.R. (2000) On the relationship between niche and distribution. *Ecology Letters*, **3**, 349–361.

Quiñonero-Candela, J., Rasmussen, C.E., Sinz, F., Bousquet, O. & Schölkopf, B. (2006) Evaluating predictive uncertainty challenge. *Machine learning challenges: Evaluating predictive uncertainty, visual object classification, and recognising textual entailment* (ed. by J. Quiñonero-Candela, I. Dagan, B. Magnini and F. d'Alché-Buc). pp. 1–27. Springer, Heidelberg, Germany.

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

Reineking, B. & Schröder, B. (2006) Constrain to perform: regularization of habitat models. *Ecological Modelling*, **193**, 675–690.

Ricklefs, R. E. & Schluter, D. (1993) *Species diversity in ecological communities. Historical and geographical perspectives.* University Chicago Press, Chicago.

Schröder, B. (2004) *ROC plotting and AUC calculation transferability test software (version 1.3-7)*. Available from http://brandenburg.geoecology.uni-potsdam.de/users/schroeder/download.html.

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

Termansen, M., McClean, C.J. & Preston, C.D. (2006) The use of genetic algorithms and Bayesian classification to model species distributions. *Ecological Modelling*, **192**, 410–424.

Thompson, M.L. & Zucchini, W. (1989) On the statistical analysis of ROC curves. *Statistics in Medicine*, **8**, 1277–1290.

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.

Wiley, E.O., McNyset, K.M., Peterson, A.T., Robins, C.R. & Stewart, A.M. (2003) Niche modeling and geographic range predictions in the marine environment using a machine-learning algorithm. *Oceanography*, **16**, 120–127.

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

BIOSKETCHES

Jorge M. Lobo is a specialist in the biogeography and ecology of dung beetles. He is interested in the patterns and processes of species distribution from a macroecological perspective, the management of biodiversity information, and conservation biology.

Alberto Jiménez-Valverde is interested in broad-scale patterns of biodiversity. He is particularly interested in methods for modelling potential distributions of species in order to understand the relative importance of environmental, biotic and historical factor limits on geographical ranges.

Raimundo Real is a professor at the University of Málaga with research interests in biogeography and conservation, mainly of vertebrates, but also of invertebrates and plants.

Editor: José Alexandre F. Diniz-Filho