Global Ecology and Biogeography, (Global Ecol. Biogeogr.) (2012) **21**, 498–507

A Journal of Macroecology

Global Ecology and Biogeography

MACROECOLOGICAL METHODS

Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling

Alberto Jiménez-Valverde*

Department of Animal Biology, Faculty of Sciences, University of Málaga, 29071 Málaga, Spain and Azorean Biodiversity Group, University of Azores, Angra do Heroísmo, Portugal

*Correspondence: Alberto Jiménez-Valverde, Department of Animal Biology, Faculty of Sciences, University of Málaga, 29071 Málaga, Spain.
E-mail: alberto.jimenez@uma.es; alberto.jimenez.valverde@gmail.com

ABSTRACT

Aim The area under the receiver operating characteristic (ROC) curve (AUC) is a widely used statistic for assessing the discriminatory capacity of species distribution models. Here, I used simulated data to examine the interdependence of the AUC and classical discrimination measures (sensitivity and specificity) derived for the application of a threshold. I shall further exemplify with simulated data the implications of using the AUC to evaluate potential versus realized distribution models.

Innovation After applying the threshold that makes sensitivity and specificity equal, a strong relationship between the AUC and these two measures was found. This result is corroborated with real data. On the other hand, the AUC penalizes the models that estimate potential distributions (the regions where the species could survive and reproduce due to the existence of suitable environmental conditions), and favours those that estimate realized distributions (the regions where the species actually lives).

Main conclusions Firstly, the independence of the AUC from the threshold selection may be irrelevant in practice. This result also emphasizes the fact that the AUC assumes nothing about the relative costs of errors of omission and commission. However, in most real situations this premise may not be optimal. Measures derived from a contingency table for different cost ratio scenarios, together with the ROC curve, may be more informative than reporting just a single AUC value. Secondly, the AUC is only truly informative when there are true instances of absence available and the objective is the estimation of the realized distribution. When the potential distribution is the goal of the research, the AUC is not an appropriate performance measure because the weight of commission errors is much lower than that of omission errors.

Keywords
AUC, background data, commission/omission errors, misclassification cost, potential distribution, realized distribution, ROC curve, sensitivity, specificity, threshold.

INTRODUCTION

The AUC, i.e. the area under the receiver operating characteristic (ROC) plot, is a measurement of the discriminatory capacity of classification models. After being developed for radar signal detection, ROC graphs were adopted in medical research (Pepe, 2000), and in the last decade the summary AUC index has been accepted as the standard measure for assessing the accuracy of species distribution models (SDMs; Fielding & Bell, 1997; Lobo *et al.*, 2008). Taking into account sensitivity (*Se*), the proportion of instances of presence correctly predicted as presence, and specificity (*Sp*), the proportion of instances of absence correctly predicted as absence, the ROC curve plots *Se* versus (1 – *Sp*) (the commission error, i.e. the proportion of instances of absence wrongly predicted as presence) across all possible thresholds between 0 and 1. A model will be considered to discriminate better than chance if the curve lies above the diagonal of no discrimination, i.e. if the AUC is higher than 0.5 (Fig. 1; see

DOI: 10.1111/j.1466-8238.2011.00683.x

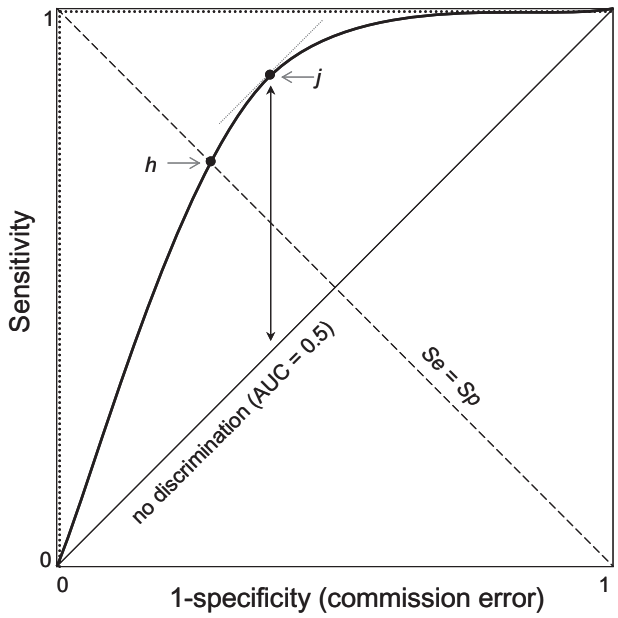

Figure 1 Idealized receiver operating characteristic (ROC) curves (dotted line, curve of a model showing perfect discrimination; black thick line, curve of a model showing imperfect discrimination) with the diagonal of no discrimination [area under the ROC curve (AUC) = 0.5] and the perpendicular along which sensitivity (*Se*) equals specificity (*Sp*) (dashed line). *h* is the point in which the ROC curve crosses the perpendicular to the line of no discrimination and *j* is the point where the slope of the tangent to the curve (thin grey line) equals 1. At *j* the Youden index and the vertical distance from the ROC curve to the diagonal of no discrimination reach their maximum values.

Krzanowski & Hand, 2009, for complete details of the ROC methodology). This statistical summary of the ROC graph has been widely adopted by SDM researchers as a standard measure of overall discriminatory capacity because it is independent of the threshold. In other words, it avoids the potential arbitrariness associated with the selection of the threshold needed to build 2 × 2 contingency matrices for calculating *Se*, *Sp* and the commission and omission error (i.e. the proportion of instances of presence wrongly predicted as absence) rates (Lobo *et al.*, 2008).

Most of the time, absence data (i.e. environmental data describing locations where absence of the species is almost certainly known) are not available and presence data (i.e. environmental data about a sample of locations that are known instances of presence) are the only information that modellers have about the species. In recent years, SDM researchers have devoted much of their effort to developing techniques to deal with this situation. For example, envelope and distance-based methods only require presence data (e.g. Busby, 1991; Farber & Kadmon, 2003). Other methods such as Maxent (Phillips *et al.*, 2006) or GARP (Stockwell & Peters, 1999) use presence and background data (environmental data about a random sample of locations with no information about the absence or presence of the species) to train the algorithms. In the same way, background data have also been used with classical presence–absence techniques such as generalized linear models (e.g. Elith *et al.*, 2006). To use the AUC without instances of absence, the ROC plot has to be modified so that instead of plotting *Se* against (1 − *Sp*), it is plotted against the proportion of the background locations predicted as presences (or the proportionate area predicted as presence) for all possible thresholds (Phillips *et al.*, 2006; Peterson *et al.*, 2008). Under this new scenario, the models are still ranked according to their AUC, i.e. the higher the better (Phillips *et al.*, 2006).

In this paper, by the general term SDMs I refer to both potential and realized distribution models *sensu* Jiménez-Valverde *et al.* (2008) and the corresponding ecological niche models (ENMs) and distribution models *sensu* Soberón (2010). Whereas potential distribution models (or ENMs) estimate the regions where the species could survive and reproduce due to the existence of suitable environmental conditions, realized distribution models estimate the regions where the species actually lives. Numerous applications of SDMs require estimations of the potential distribution, for instance the prediction of species invasions, the assessment of the impact of climate change on species distributions, the discovery of new species or the understanding of species evolutionary and biogeographic history (see Peterson, 2006). Estimates of realized distribution are of interest in many conservation planning studies (Peterson, 2006). Thus, the distinction is far from trivial and different strategies, in terms of data and modelling techniques, are required for approaching one concept or another (Jiménez-Valverde *et al.*, 2008; Soberón & Nakamura, 2009). In the same way, the strategies used to evaluate the models need to be different because the weight of commission errors is definitively lower in the case of the potential distribution than in the case of the realized distribution (Lobo *et al.*, 2008; Peterson *et al.*, 2008).

Recently, the AUC has been severely criticized in the SDM field (Lobo *et al.*, 2008; Peterson *et al.*, 2008). The need to evaluate the models and the current huge reliance of SDM researchers on the AUC statistic make it urgent that this matter is addressed. In this paper, I analyse the question of the supposed advantage of the AUC over threshold-dependent measures. I also address the implications of using the AUC depending on whether the goal of the research is to estimate the potential or the realized distribution, which is closely related to the situation in which no instances of absence are available and background data are used to evaluate the models.

THE ROC CURVE

Key details about the ROC plot

The ROC space is defined by the *Se* and (1 − *Sp*) axes; the upper-left corner of this space, i.e. the point (0, 1), satisfies the equation

$$Se = Sp = 1. \quad (1)$$

The ROC curve for a model with perfect discriminatory capacity would go from point (0,0) to point (0,1), and from this to point

(1,1) (Fig. 1). In theory, the AUC does not depend on a particular portion of the curve (Zweig & Campbell, 1993). However, in practice, the conjunction of three features of the ROC graph leads me to suggest that the AUC value strongly depends on certain points of the curve: (1) the high discriminatory capacity corresponds to curves that are closer to point (0,1); (2) the ROC curve is – by definition – anchored to points (0,0) and (1,1); (3) the curve monotonically increases and is by nature convex. Thus, the 'point of inflexion' of the smoothed ROC graph (*j*) may be a determinant point in the curve. This point corresponds with that in which the slope of the tangent to the ROC curve equals 1 (Fig. 1) or with that in which the Youden index is maximized (Hilden, 1991), i.e.

$$\max\{(Se + Sp - 1)\}. \quad (2)$$

From equation 2 it is straightforward that this point also maximizes (*Se* + *Sp*), i.e. it minimizes the error or misclassification rate (Kaivanto, 2008). It is also the point that maximizes the vertical distance between the ROC curve and the diagonal of no discrimination (Fig. 1; Perkins & Schisterman, 2005).

Some authors have suggested identifying the upper-left point of the ROC curve by applying the Pythagorean theorem (e.g. Freeman & Moisen, 2008), satisfying

$$\min\left\{(1 - Se)^2 + (1 - Sp)^2\right\}. \quad (3)$$

However, note that although the length of the segment from point (0,1) to the point of the ROC curve that satisfies equation 3 is geometrically intuitive, it has no probabilistic basis, and therefore no practical interpretation (J. Hilden, pers. comm.). Actually, the point identified as such is not the one with the minimum overall error rate and does not have to coincide with *j* (Perkins & Schisterman, 2006). The origin of this misconception comes from the usual statement in the ROC literature that the nearer the ROC curve is to the upper-left corner of the ROC space, the higher the discriminatory capacity of the model. Proximity to the point (0,1) is meant as an imprecise definition here; it has never been intended to be naively interpreted as Euclidean distance (J. Hilden pers. comm.). Zweig & Campbell (1993, p. 565) wrote 'Qualitatively, the closer the plot is to the upper left corner, the higher the overall accuracy of the test'. The adjective 'qualitative(ly)', which is repeated twice more in the same context throughout the paper (see p. 566), calls attention to the vagueness of the statement.

Another important point is the one in which the ROC curve crosses the perpendicular to the line of no discrimination (*h*; Fig. 1), satisfying

$$Se = Sp. \quad (4)$$

Its centred location in the ROC curve makes this point intuitively important in the determination of the AUC. Since the points (0,0) and (1,1) are fixed, and because the ROC curve is usually convex, variations in the location of *h* in the ROC space may have the most dramatic changes in the AUC.

Let *T* be the threshold for a certain point in the curve. Note that if the costs of omission and commission errors are the same, the two thresholds, T_j and T_h, are appealing because: (1) given the trade-off between *Se* and *Sp* (an increase in *Se* implies a decrease in *Sp*, and vice versa; Shapiro, 1999), T_h balances the rate of correctly predicted presences and correctly predicted absences, and (2) T_j maximizes the rate of correct classifications. Additionally, if the prevalence is 0.5, T_h balances the costs of making commission and omission errors and T_j minimizes the overall misclassification cost. These are desirable properties in a good performance classifier (Fielding & Bell, 1997; Kaivanto, 2008).

Here, I used a virtual species to study the relationship between the AUC value and the *Se* and *Sp* obtained using T_j and T_h, to test the hypothesis that the AUC is closely related to certain points in the ROC curve. Furthermore, for comparison and just because it is widely used in SDM (Freeman & Moisen, 2008; but see the Discussion), I also explored the relationship with the threshold that maximizes the kappa statistic (κ; Cohen, 1960). To corroborate the consistency of the main result in a real-world situation, I used the results of the SDMs of 48 arthropod species on Terceira Island published in Jiménez-Valverde *et al.* (2009b).

ROC analysis with background data

A common situation in SDMs is a complete lack of absence data (Lobo *et al.*, 2010). As a result, in recent years, despite its weaknesses (see Warton & Shepherd, 2010), an approach taken from the resource selection literature (Manly *et al.*, 2002) has gained popularity with SDM researchers. Background data are randomly selected from the area of study, and the idea behind this approach is to find a function which discriminates between the instances of presence and background locations. When no instances of absence are available, and (1 – *Sp*) is replaced by the number of background data predicted as present in the ROC plot, Wiley *et al.* (2003) and Phillips *et al.* (2006) argued that the maximum AUC value (AUC_{max}) depends on the actual (unknown) area of distribution of the species. As a result, AUC_{max} no longer has to be equal to 1, and is inversely related to the area of distribution. In other words, the bigger the area, the lower AUC_{max} will be. In practice, since the area of distribution of the species is unknown (this is why the models are needed!), AUC_{max} is also unknown. This just recognizes that among the background data there are presences, and that the main interest is in identifying those background data that could be presences based on the predictors. This is a clear acknowledgement that background data are – in no way – a 'gold standard' (a group of cases, independent from the training set, in which the state of the dependent variable – presence and absence of the species – is known without error). Notice that, in the same way that the AUC_{max} no longer has to be 1, the AUC corresponding to a no-better-than-chance prediction no longer has to be 0.5. Every point along the line of no discrimination in the ROC plot (Fig. 1) satisfies

$$Se = 1 - Sp. \quad (5)$$

The explanation for equation 5 is that if a model predicts no better than chance, one expects the same proportion of correctly

and wrongly predicted presences and absences (Fawcett, 2006). However, when one *explicitly* acknowledges that an unknown percentage of background data are in fact instances of presence, this reasoning no longer makes any sense. For this same reason, the comparison of models between species is flawed as there is no point in considering species with the highest AUC values to be better predicted than those species with the lowest ones; this will entirely depend on the *unknown* area of distribution of each species. That the boundaries of the AUC are no longer fixed at 0.5 and 1, and that higher AUC values are no longer equivalent to better predictions, implies that AUC theory is clearly violated, hampering the evaluation and comparison of models.

Evaluating potential or realized distribution models

Different situations require different objectives; the distinction between potential and realized distribution is essential. Instances of presence and instances of absence, both coming from a standardized and unbiased sampling, are necessary if the goal is to estimate the realized distribution (Jiménez-Valverde *et al.*, 2008; Soberón & Nakamura, 2009; Ward *et al.*, 2009). Most importantly, unbiased instances of absence are a must if one wants to evaluate how accurately a model estimates a species' realized distribution (Jiménez-Valverde *et al.*, 2008). However, in the case of the potential distribution, the evaluation of the models is not so straightforward as there is no 'gold standard' available for validation. The situation is analogous to the use of background data to evaluate the models; since the potential distribution of the species is unknown, there is no reason to penalize the models for the number of absence or background data – or the extent of the area – predicted as presence. Here, I will use a virtual species to show the implications of using the AUC to evaluate estimations of potential distributions.

METHODS

Relationship between AUC, and *Se* and *Sp* yielded by different threshold criteria

I used a real environmental landscape to generate my simulated data. Annual temperature range (*temprng*) and precipitation of the wettest month (*precpw*) for the area in North America from 38.55 to 46.88 latitude and from −123.07 to −111.40 longitude were obtained from the WorldClim database (http://www.worldclim.org; Hijmans *et al.*, 2005). In this way, a spatial lattice with 1,400,000 cells (30 arcsec resolution) was considered and each cell was characterized by these two climatic variables. Variables were standardized (mean = 0, SD = 1) and the probability of presence of the species ($y = 1$) in each cell was conditioned on each independent variable (x) following the Gaussian (symmetric bell-shaped) function:

$$p(y=1|x) = c\exp\left[-(x-u)^2/2t^2\right] \quad (6)$$

where c is the maximum value of $p(y = 1|x)$, u is the optimum [the value of x where $p(y = 1|x)$ is maximum] and t is the tolerance of the response (ecological width). The final probability of presence of the simulated species $p(y = 1)$ is the product $p(y = 1|temprng) \times p(y = 1|precpw)$. To generate a species with a prevalence of 0.5 in the extent under consideration (1,400,000/2 presences), $c = 1$, $u = 0$ and $t = 0.75$ for the two variables. To finally determine the presence of the species in a cell, random numbers were picked from a uniform distribution; if the random number was lower than $p(y = 1)$, then the species was considered present in that cell (Appendix S1 in Supporting Information).

The lattice was divided into 70 sublattices of equal size (Appendix S1); each sublattice contained 20,000 spatially contiguous cells. Sublattices with a prevalence (number of presences/20,000) lower than 0.01 or higher than 0.99 were excluded from the model training step (Jiménez-Valverde *et al.*, 2009a), which left 68 sublattices for model parameterization. Logistic regressions (generalized linear models with binomial distribution and logit-link function; McCullagh & Nelder, 1989) were used to model the presence–absence simulated data in each of the 68 sublattices, including as independent variables the quadratic terms of *temprng* and *precpw* and their interaction. Probabilities of presence were converted into binary maps using three threshold criteria: (1) the threshold that satisfies equation 2 (T_j; Hilden, 1991); (2) the threshold that satisfies equation 4 (T_h; equivalent to the threshold that minimizes the difference between *Se* and *Sp* in empirical ROC curves; Jiménez-Valverde & Lobo, 2007); and (3) the threshold that maximizes κ (T_{kappa}; Freeman & Moisen, 2008). In the case of T_{kappa}, ties occurred (i.e. the maximum κ value was obtained with two or more thresholds); to break them, T_h was used as a second criterion. These three criteria were applied both to the probabilities of the training sublattice (herein 'internal' evaluation) and to the probabilities yielded by the models when they were used to predict the distribution of the species in the 69 (70 sublattices minus the sublattice used for model training) non-training sublattices (herein 'external' evaluation); internal and external AUC values were also calculated. Appendix S2 shows the structure of the simulation process. The basic idea of this simulation scheme was to generate models with varying degrees of predictive capacity, depending on the environmental characteristics of the sublattices with respect to those of the whole lattice. As a final result, a matrix of 68 rows (cases; the 68 sublattices available for models training) × 20 columns (variables; internal and external AUC values, internal and external threshold values for the three criteria and internal and external *Se* and *Sp* values for the three criteria) was available for further analyses. The relationship between the AUC values and *Se* and *Sp* for each threshold criterion was explored in both the internal and the external evaluation data. All the analyses were done in R (R Development Core Team, 2008) using the PresenceAbsence package (Freeman, 2009); the function glm was used to fit the logistic models.

Jiménez-Valverde *et al.* (2009b) modelled the realized distribution of 48 arthropod species (sample sizes ranging from 69 to 590) on Terceira Island using neural networks. Presences and absences were obtained from a standardized sampling, and several climatic, historical and anthropogenic variables were

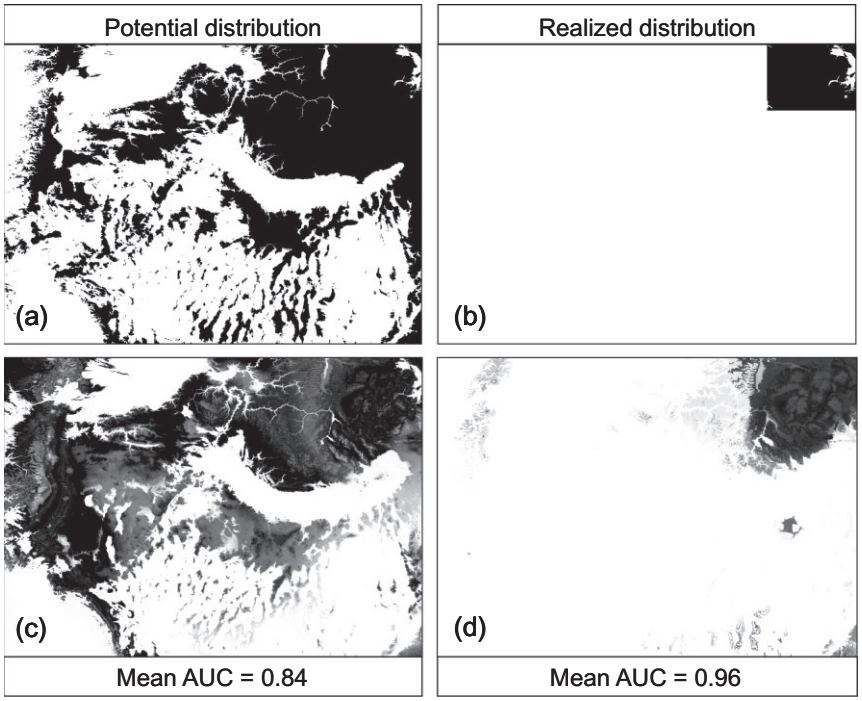

Figure 2 (a) Potential distribution of the simulated species (in black); the species can inhabit all locations (cells of the lattice) with *temp* values lower than an arbitrary score. (b) Realized distribution of the species (in black); unknown factors restrict the distribution of the species (c) Distribution model using Maxent and *tempw and precpcv* as predictors; the model approximated the potential distribution of the species and yielded a mean area under the receiver operating characteristic curve (AUC) of 0.84 (± 0.002). (d) Distribution model using Maxent and *precpwarm and tempd* as predictors; the model approximated the realized distribution of the species and yielded a mean AUC of 0.96 (± 0.001).

used as predictors. They evaluated their models using a jackknife procedure and calculated the AUC values and *Se* and *Sp* using the T_h criterion (see Table 2 in their paper). Their jackknife approach is an intermediate evaluation procedure ('data splitting') between my internal ('resubstitution'; using the same training data to evaluate the models) and external ('independent evaluation'; using independent data) evaluations. The relationship between the AUC values and *Se* (≈ *Sp*) in this empirical data was explored.

The AUC and the potential distribution

Considering the same spatial lattice, another species with a very simple response to the environment was simulated; it was able to live below a certain annual mean temperature (*temp*) value and was unable to exist above it. The potential distribution was restricted only by *temp* (Fig. 2a). However, because of unknown non-climatic factors, the species actually occupied just a small portion of the suitable area (the realized distribution; Fig. 2b). The actual presence cells (forming the realized distribution) were randomly divided into a training (*c.* 30,000) and a testing set (*c.* 30,000) 20 times, each time using a different seed, and modelled with Maxent (version 3.2.1; Phillips *et al.*, 2006) using as background data absence cells from the entire lattice. Then, the mean suitability of each cell and the mean AUC using the auto features option and the logistic output were calculated. This experiment was conducted twice, first using the mean temperature of the warmest quarter (*tempw*) and precipitation seasonality (*precpcv*) and then using the mean temperature of the driest quarter (*tempd*) and the precipitation of the warmest quarter (*precpwarm*) as predictors (variables were obtained from the WorldClim database). The first set of predictors showed a higher correlation with the determinant of the potential distribution (*temp*) than the second set. Neither the true climatic variable that determines the species' potential distribution nor the unknown variable(s) that determines the realized distribution were included in the modelling process, mimicking a real-world modelling exercise in which these variables are completely ignored.

RESULTS

Relationship between AUC and *Se* and *Sp* yielded by different threshold criteria

Only with T_h was there an almost identity relationship between the AUC and the *Se* and *Sp* in the internal as well as in the external evaluation, with significant positive correlations as high as 0.99 and 0.97, respectively (Figs 3 & 4). Note, however, that the slope of the regression line was always significantly lower than 1 (internal evaluation, *Se*: $t = -6.83$; internal evaluation, *Sp*: $t = -7.26$; external evaluation, *Se*: $t = -11.66$; external evaluation, *Sp*: $t = -12.10$, $P < 0.01$ in all cases). With T_j and T_{kappa} the relationship was far from identity and was even negative in the external evaluation (Figs 3 & 4).

Empirical AUC and *Se* (and *Sp*) values from Jiménez-Valverde *et al.* (2009b) showed a relationship close to the identity line (again, the slope of the regression line was significantly lower than 1; *Se*: $t = -5.30$, *Sp*: $t = -5.22$, $P < 0.01$ in both cases) and a significant positive correlation of 0.98 (Fig. 5).

The AUC and the potential distribution

Using *tempw* and *precpcv* as predictor, the potential distribution was approximated and the mean AUC was 0.84 (± 0.002 SD)

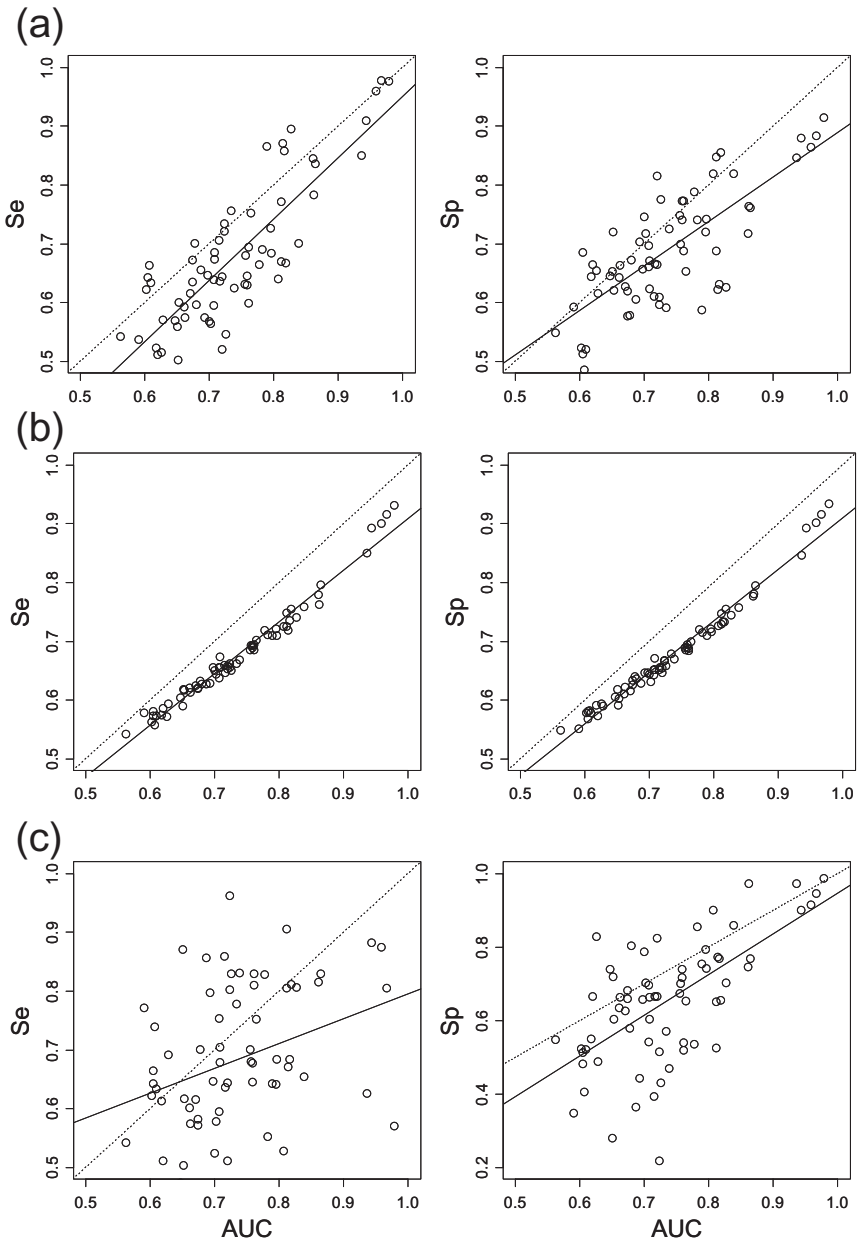

Figure 3 Relationship between the area under the receiver operating characteristic curve (AUC) values and sensitivity (*Se*) and specificity (*Sp*) using the three threshold criteria in the internal evaluation: (a) T_p, (b) T_h, (c) T_{kappa}. The dotted line indicates the identity line.

(Fig. 2c); when including *tempd* and *precpwarm*, the model approximated the realized distribution (Fig. 2d) and the mean AUC increased to 0.96 (± 0.001 SD). Despite the fact that the first model was the best if our goal was to estimate the potential distribution, the second model was the best according to the AUC.

DISCUSSION

The AUC is closely related to *Se* and *Sp* yielded by the threshold criterion that makes *Se* = *Sp*

There is a strong relationship between AUC values and *Se* and *Sp* values yielded by the T_h criterion. These results show that there is not much difference between evaluating and comparing the models using the AUC or *Se* (and *Sp*) after applying the T_h criterion, as these measures closely covary. The fact that this relationship was held in the real data suggests: (1) that the pattern is robust and not an artefact of the simulation exercise and (2) that, as should be expected, it is insensitive to the modelling technique. Hilden (1991, p. 98) already noted that '. . . typically A [AUC] varies in parallel with any other measure of diagnosticity that one might propose'. Recall that using the T_h cutoff assumes equal costs of omission and commission errors. Although the AUC assumes nothing about the relative costs of both errors (actually, it is an average of all possible cost ratio situations; Adams & Hand, 2000), in practice, the equal-cost criterion has a determinant weight on the AUC value. However, because the relative magnitude of omission and commission errors changes with the problem under consideration, the equal-

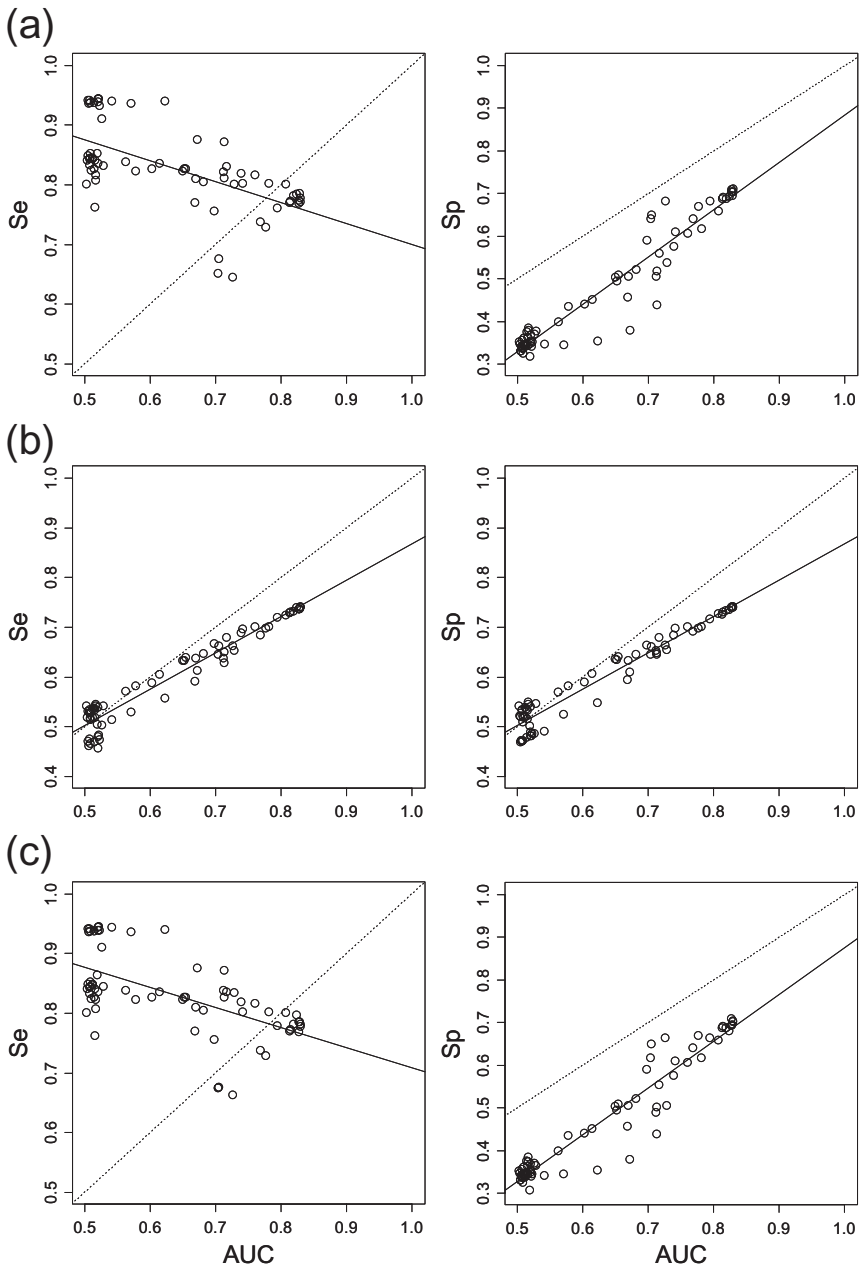

Figure 4 Relationship between the area under the receiver operating characteristic curve (AUC) values and sensitivity (*Se*) and specificity (*Sp*) using the three threshold criteria in the external evaluation: (a) T_j, (b) T_h, (c) T_{kappa}. The dotted line indicates the identity line.

cost criterion may not be optimal (see, for example, the discussion about the AUC and the potential distribution below). I argue that reporting different fractions of the 2×2 contingency matrix, together with the prevalence of the instances of presence and the threshold value (and the reasons for choosing such a threshold), is much more informative than reporting just the summary AUC value. This is in accordance with the argument raised by Lobo and collaborators (2008) who suggested that once the 'optimal' threshold is calculated the arbitrariness associated with its selection disappears, thus questioning the advantage of the AUC.

The fact that the slopes of the regression lines were consistently lower than 1 implies that low AUC values will lead to very similar *Se* and *Sp* values (notice that, ideally, an AUC value of 0.5 unavoidably entails an *Se* and *Sp* value of 0.5) whereas high AUC values will lead to lower *Se* and *Sp* values. This can be a somewhat devilish effect: it is always more appealing to report the highest performance index – in this case, the AUC. For instance, if one has to choose between an AUC of 0.85 and an *Se* (and *Sp*) of 0.75 (real case represented by a solid circle in Fig. 5), it is likely that one would prefer to support one's results with the AUC. However, to report the AUC without the associated ROC plot is less informative than to report *Se* and *Sp* (and other statistics derived from the contingency table) at a certain 'optimal' threshold (which must be reported as well as the threshold criterion applied).

At this point, I want to stress that my criticism is not about the ROC curve itself but the summary AUC index. When both presences and true absences are available (see next section), the ROC

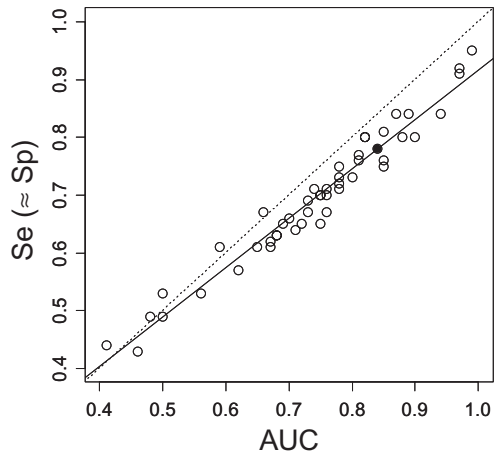

Figure 5 Relationship between the area under the receiver operating characteristic curve (AUC) and sensitivity (*Se*) (≈ specificity, *Sp*) using the T_h criterion of the realized distribution models of 48 arthropod species on Terceira Island (Jiménez-Valverde *et al.*, 2009b). The solid dot corresponds to a species model with an AUC = 0.85 and *Se* = *Sp* = 0.75. The dotted line indicates the identity line.

plot can constitute a very informative and powerful tool for assessing and comparing predictive models (see Fawcett, 2006; Krzanowski & Hand, 2009). But summarizing the ROC graph into a single measure such as the AUC unavoidably implies a loss of information; ignoring its properties and flaws can distort and bias our perception of model performances.

I want to highlight that the T_{kappa} criterion was included in this study just for comparative purposes. It is known that κ depends on the prevalence of the data (Thompson & Walter, 1988; Allouche *et al.*, 2006), and for this reason some authors have advised against its use in model evaluation (Jiménez-Valverde *et al.*, 2009a). The rationality behind these arguments is that κ was developed to test agreement between models, not the predictive accuracy of a model against the 'gold standard', and so the use of κ in model validation is highly questionable (Tooth & Ottenbacher, 2004).

Evaluating potential distribution estimations with the AUC does not make sense

Using the AUC for the evaluation of potential distribution modes, or for the evaluation of SDMs using background data instead of true absences, violates AUC theory. First, the evaluation of models using background data no longer fixes the limits of the AUC (0.5 and 1, respectively). Essentially, background data are not a 'gold standard', either for the potential distribution or for the realized distribution. Second, the results expounded in this study show that the AUC can be misleading as competing models will be ranked according to their capacity to approximate to the realized distribution, not the potential one. It can easily be understood that the more absence (or background) locations that are predicted as absences (i.e. the lower the proportionate area predicted as presence), the higher the AUC. In any case, this would only make sense when the goal is to estimate the realized distribution (and assuming that the presence data come from an unbiased sampling); however, when trying to estimate the potential distribution, there is no reason to penalize for 'overprediction'. As we have seen above, the AUC is a measure of the overall discrimination capacity of the models, i.e. it assumes nothing about the relative costs of omission and commission errors. When the objective of the research is the estimation of the potential distribution, the cost of both type of errors is not the same; commission errors are of much less concern than omission errors (Anderson *et al.*, 2003; Lobo *et al.*, 2008; Peterson *et al.*, 2008; Jiménez-Valverde *et al.*, 2011). When some idea about the relative importance of the errors is known, using an overall measure such as the AUC can be misleading (Adams & Hand, 1999). The goal of the research must be unequivocal with respect to the potential versus realized distribution framework; clearly, if the objective is to predict and evaluate estimations of potential distributions (i.e. in an ENM context *sensu* Soberón, 2010) the AUC is useless. On the contrary, if the objective is to produce and evaluate estimations of realized distributions, unbiased true instances of absence instead of background data are necessary in order to avoid violation of AUC theory. I want to point out that any other statistic that combines *Se* and *Sp* (and, so, omission and commission errors) in a single measure (e.g. the Youden index, maximum Kappa) may suffer from the same conceptual problems in this context.

Concluding remarks

The proper evaluation of models is a key challenge in SDM (Vaughan & Ormerod, 2005) and despite the wide acceptance of some standard methods/statistics, there is currently a strong ongoing debate as to their use (e.g. Allouche *et al.*, 2006; Jiménez-Valverde *et al.*, 2008; e.g. Lobo *et al.*, 2008; Peterson *et al.*, 2008; Santika, 2011). Lobo *et al.* (2008) seriously questioned the reliance on the AUC as a *single* measure of model performance. They highlighted that the AUC ignores the goodness-of-fit of the models, that it assumes equal costs for commission and omission errors, and that it is spatially independent (we have no information about the spatial distribution of the errors). But probably the most important point raised by Lobo and collaborators was that the AUC value is sensitive and positively related to the spatial extent under consideration (see also Jiménez-Valverde *et al.*, 2008, and Barve *et al.*, 2011), which seriously places into doubt the robustness of SDM studies. All these characteristics are not exclusive to the AUC but are common to any discrimination measure. Although they are not shortcomings per se, they can become serious drawbacks if researchers are not completely aware of them and if the conclusions of the studies rely entirely on just a single omnipotent number.

What has been shown in this paper is that the most written argument in favour of the AUC, i.e. that it avoids the reluctance to adopt a threshold to convert the probabilities into a presence–absence variable that others may regard as arbitrary or inappropriate, seems to be irrelevant in practice. On the other hand, the

use of background data to evaluate the models violates AUC theory. Finally, the AUC is not an appropriate performance measure for potential distribution models. Because the potential distribution is actually non-existent (it is the distribution that the species *could* have in the absence of historical, biotic and other restrictive factors not considered in the models), there is a priori no reason to penalize the models for the overall extension of the region that is predicted to be populated. Numerous applications of SDMs require estimations of the potential distributions, so this scheme of thought must be carefully considered by the researchers.

Critical thinking and debating is fundamental to the practise of science, otherwise we no longer engage in true inquiry. In a field such as SDM, in which results are applied to important areas of research like climate change, invasive species or disease spread, the debate about model evaluation should be encouraged. The development of a 'comprehensive toolbox of evaluation measures' (Elith & Leathwick, 2009, p. 691) is strengthened by challenging contributions. The AUC is just one of the many metrics that can be used to evaluate the discrimination capacity of predictive models. It may be handy in particular situations, and *together with the ROC plot* it can provide valuable information. However, I want to highlight that the AUC has sometimes been adopted in the SDM field in a rather uncritical manner. The choice to use any given statistic should be driven by what the situation conceptually demands; just reporting the AUC because it is easy (popular analysis packages report it by default) does not make it right. The widespread use of this single measure of overall performance without careful consideration of its properties and limitations can be unsuitable and harmful in many situations.

ACKNOWLEDGEMENTS

I would like to give special thanks to Jørgen Hilden for his helpful, generous and intelligent comments which greatly improved previous versions of this paper. My views have been influenced considerably by the discussions and collaborative work with Jorge M. Lobo. The editor J. A. F. Diniz-Filho, A. T. Peterson and two anonymous referees raised important points that enhanced the final version of the article. Elizabeth A. Freeman helped me with the PresenceAbsence package. Lucía Maltez kindly reviewed the English. I was supported by the MEC Juan de la Cierva Program.

REFERENCES

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

Adams, N.M. & Hand, D.J. (2000) An improved measure for comparing diagnostic tests. *Computers in Biology and Medicine*, **30**, 89–96.

Allouche, O., Tsoar, A. & Kadmon, R. (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, **43**, 1223–1232.

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.

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S.P., Peterson, A.T., Soberón, J. & Villalobos, F. (2011) The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling* (in press).

Busby, J.R. (1991) BIOCLIM. A bioclimate analysis and prediction system. *Nature conservation: cost effective biological surveys and data analysis* (ed. by C.R. Margules and M.P. Austin), pp. 64–68. CSIRO, Melbourne.

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37–46.

Elith, J. & Leathwick, J. (2009) Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution and Systematics*, **40**, 677–697.

Elith, J., Graham, C.H., Anderson, R.P. *et al*. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Farber, O. & Kadmon, R. (2003) Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. *Ecological Modelling*, **160**, 115–130.

Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recognition Letters*, **27**, 861–874.

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.

Freeman, E. (2009) *PresenceAbsence: presence–absence model evaluation. R package version 1.1.3*. Available at: http://www.R-project.org (accessed March 2009)

Freeman, E.A. & Moisen, G.G. (2008) A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling*, **217**, 48–58.

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

Hilden, J. (1991) The area under the ROC curve and its competitors. *Medical Decision Making*, **11**, 95–101.

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

Jiménez-Valverde, A., Lobo, J.M. & Hortal, J. (2008) Not as good as they seem: the importance of concepts in species distribution modelling. *Diversity and Distributions*, **14**, 885–890.

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

Jiménez-Valverde, A., Diniz, F., Azevedo, E.B. & Borges, P.A.V. (2009b) Species distribution models do not account for abundance: the case of arthropods on Terceira Island. *Annales de Zoologici Fennici*, **46**, 451–464.

Jiménez-Valverde, A., Decae, A.E. & Arnedo, M.A. (2011) Environmental suitability of new reported localities of the funnelweb spider *Macrothele calpeiana*: an assessment using potential distribution modelling with presence-only techniques. *Journal of Biogeography*, doi: 10.1111/j.1365-2699.2010.02465.x

Kaivanto, K. (2008) Maximization of the sum of sensitivity and specificity as a diagnostic cutpoint criterion. *Journal of Clinical Epidemiology*, **61**, 517–518.

Krzanowski, W.J. & Hand, D.J. (2009) *ROC curves for continuous data*. Chapman and Hall, Boca Raton, FL.

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

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

McCullagh, P. & Nelder, J.A. (1989) *Generalized linear models*. Chapman and Hall, London.

Manly, B.F.J., McDonald, L., Thomas, D.L., McDonald, T.L. & Erickson, W.P. (2002) *Resource selection by animals: statistical design and analysis for field studies*. Kluwer Press, New York.

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

Perkins, N.J. & Schisterman, E.F. (2005) The Youden index and the optimal cut-point corrected for measurement error. *Biometrical Journal*, **47**, 428–441.

Perkins, N.J. & Schisterman, E.F. (2006) The inconsistency of 'optimal' cutpoints obtained using two criteria based on the receiver operating characteristic curve. *American Journal of Epidemiology*, **163**, 670–675.

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

Peterson, A.T., Papeş, M. & Soberón, J. (2008) Rethinking receiver operating characteristic analysis applications in ecological niche modelling. *Ecological Modelling*, **213**, 63–72.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

R Development Core Team (2008) *R: a language and environment for statistical computing*, version 2.7.2.R Foundation for Statistical Computing, Vienna. Available at: http://www.R-project.org (accessed August 2008).

Santika, T. (2011) Assessing the effect of prevalence on the predictive performance of species distribution models using simulated data. *Global Ecology and Biogeography*, **20**, 181–192.

Shapiro, D.E. (1999) The interpretation of diagnostic tests. *Statistical Methods in Medical Research*, **8**, 113–134.

Soberón, J. (2010) Niche and area of distribution modeling: a population ecology perspective. *Ecography*, **33**, 159–167.

Soberón, J. & Nakamura, M. (2009) Niches and distributional areas: concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences USA*, **106**, 19644–19650.

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

Thompson, W.D. & Walter, S.D. (1988) A reappraisal of the kappa coefficient. *Journal of Clinical Epidemiology*, **41**, 949–958.

Tooth, L.R. & Ottenbacher, K.J. (2004) The *k* statistic in rehabilitation research: an examination. *Archives of Physical Medicine and Rehabilitation*, **85**, 1371–1376.

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

Ward, G., Hastie, T., Barry, S., Elith, J. & Leathwick, J. (2009) Presence-only data and the EM algorithm. *Biometrics*, **65**, 554–563.

Warton, D.I. & Shepherd, L.C. (2010) Poisson point process models solve the 'pseudo-absence problem' for presence-only data in ecology. *Annals of Applied Statistics*, **4**, 1383–1402.

Wiley, E.O., McNyset, K.M., Peterson, A.T., Robins, C.R. & Stewart, A.M. (2003) Niche modelling 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 characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, **39**, 561–577.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1 The spatial lattice and the virtual species.
Appendix S2 Structure (pseudocode) of the simulation process.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

BIOSKETCHES

Alberto Jiménez-Valverde is currently a Juan de la Cierva researcher at the University of Málaga. He is interested in understanding the relative importance of environmental, biotic and historical factors in limiting species geographical ranges. He is also very interested in methodological and conceptual issues related to species distribution models, and in the ecology and biogeography of spiders.

Editor: José Alexandre F. Diniz-Filho