Research

Ecography 34: 232–243, 2011
doi: 10.1111/j.1600-0587.2010.06354.x

Subject Editor: Miguel Araújo. Accepted 14 March 2010

Measuring and comparing the accuracy of species distribution models with presence–absence data

Canran Liu, Matt White and Graeme Newell

C. Liu (canranliu@hotmail.com), M. White and G. Newell, Arthur Rylah Inst. for Environmental Research, Dept of Sustainability and Environment, 123 Brown Street, Heidelberg, Victoria 3084, Australia.

Species distribution models have been widely used to predict species distributions for various purposes, including conservation planning, and climate change impact assessment. The success of these applications relies heavily on the accuracy of the models. Various measures have been proposed to assess the accuracy of the models. Rigorous statistical analysis should be incorporated in model accuracy assessment. However, since relevant information about the statistical properties of accuracy measures is scattered across various disciplines, ecologists find it difficult to select the most appropriate ones for their research. In this paper, we review accuracy measures that are currently used in species distribution modelling (SDM), and introduce additional metrics that have potential applications in SDM. For the commonly used measures (which are also intensively studied by statisticians), including overall accuracy, sensitivity, specificity, kappa, and area and partial area under the ROC curves, promising methods to construct confidence intervals and statistically compare the accuracy between two models are given. For other accuracy measures, methods to estimate standard errors are given, which can be used to construct approximate confidence intervals. We also suggest that as general tools, computer-intensive methods, especially bootstrap and randomization methods can be used in constructing confidence intervals and statistical tests if suitable analytic methods cannot be found. Usually, these computer-intensive methods provide robust results.

Species distribution modelling (SDM) has become a useful tool for fundamental ecological and biogeographical research and it is an increasingly important tool for biodiversity management and conservation planning (Guisan and Zimmermann 2000, Araújo and Pearson 2005, Thuiller et al. 2005, Pearson et al. 2007). Species distribution models are used to predict the geographic range of a species from occurrence records and relevant environmental data. Two types of model output are common: binary results where sites are classified as either part of the distribution of the species or outside their distribution; and continuous results where each site is given a probability of being part of the species distribution or suitability for the species. This kind of modelling methodology has long been used in medical diagnostics (Yerushalmy 1947, Hand 1992), where patients (rather than sites) are predicted as either positive – suffering from or susceptible to a particular malady or otherwise, and meteorology (Finley 1884, Murphy and Winkler 1984, Glahn 2004), where a specific meteorological event is predicted to occur or not occur.

Assessing the utility of such models requires an evaluation of the performance or accuracy of the models. This is a critical element of model-building (Guisan and Zimmermann 2000). Robust assessment of performance will identify the relative strengths and weaknesses of models and delimit the range of uses to which they can be usefully applied.

Model performance assessment is based on accuracy measures calculated from a set of independent test data (Heikkinen et al. 2006). Researchers often calculate the values for the accuracy measures they choose without any indication of the precision for the calculated values. If the test data set is large, this is unlikely to be a significant problem since the variation will usually become smaller as sample size increases. However, ecological researchers often develop models with relatively small datasets due to lack of available survey data, budgetary constraints, or because of the restricted distribution or rarity of the species. In this situation, simply calculating a value for each of the accuracy measures used is inadequate, since variations associated with the estimated accuracy measures exist to some unknown extent. Ignorance of this aspect may ultimately provide misleading conclusions. Therefore, as good practice, assessment of the precision of the estimated accuracy measure, e.g. a standard error, or even better, a confidence interval (CI) for each of the estimated accuracy measures, should be provided (Jolliffe 2007). When different models are contrasted, a formal statistical test should be conducted, or the confidence interval for the difference of measure between two models be provided (see Fielding (2007) for

additional comments on this topic). Although some methods for statistical inference related to accuracy measures have been introduced to ecology (Fielding and Bell 1997, Couto 2003, Allouche et al. 2006), many promising methods exist outside the ecological literature, especially those for constructing confidence intervals and for comparing models using the same test data set, which are scattered in other fields, including statistics, biometry, medicine, psychology, meteorology and machine learning.

In this paper, we review the various accuracy measures that are currently applied to SDM, and suggest several additional, promising measures. For each measure, we attempt to provide methods to calculate the standard error and construct confidence intervals, as well as provide methods to compare alternate models in terms of their accuracy measures. When more than one method for a specific measure exists in the literature, we attempt to provide details for the most appropriate and parsimonious approach. More than one method may be provided if a consensus is not evident from the literature. Our aim is to provide ecological researchers with a useful approach to implementing the most appropriate statistical analysis of the accuracy measure they have selected to use. We also provide six appendices (Supplementary material Appendix S1–S6) as supplementary material to give more details for the recommended methods.

Accuracy measures

There are two facets to measuring the accuracy of species distribution models: discrimination capacity and reliability (Pearce and Ferrier 2000), although the former has been generally viewed as more important than the latter (Ash and Shwartz 1999). Discrimination capacity refers to a model's ability to distinguish between sites where the subject species has been detected (presence sites) and those sites where the species is known to be absent (absence sites). Reliability refers to the agreement between predicted probabilities of occurrence and the observed proportions of sites occupied by the species (Pearce and Ferrier 2000). Reliability is an essential attribute of the quality of probabilistic predictive models.

Both aspects of model performance (discrimination capacity and reliability) can be assessed when the modelling result is continuous. When the modelling result is binary, only discrimination capacity can be assessed. A range of indices are available to evaluate either discrimination capacity and/or reliability. A number of these can only be applied to binary results or to continuous results that have been transformed into a binary solution by using a specific cut-off value or threshold. These are called threshold-dependent measures. Those that can be applied directly to continuous predictions are called threshold-independent measures. If the threshold value is changed systematically, the optimal value of a threshold-dependent measure can be obtained (subject to an agreed definition of optimality). Since this process is not dependent on a specific threshold value, these types of optimal values can also be treated in the same manner as threshold-independent values.

Threshold-dependent measures

All threshold-dependent measures are based on some or all of the elements of the confusion matrix (Table 1). Some measures in this group are shown in Table 2. Sensitivity (Se) and specificity (Sp) are conditional probabilities widely used in many disciplines including SDM. Se is the probability that the model correctly predicts an observation of a species at a site, and Sp is the probability that a known absence site is correctly predicted. While both Se and Sp are probabilities conditional on the observations, positive predictive value (PPV, also called positive predictive power) and negative predictive value (NPV, also called negative predictive power) are their counterparts that are conditional upon the predictions. PPV is the probability that a site predicted as presence is actually a presence and NPV is the probability that a site predicted as absence is truly an absence. In the field of image classification, Se and Sp are referred to as producer's accuracy for the two classes – presence and absence, and PPV and NPV are called user's accuracy for the two classes (Liu et al. 2007). In the fields of machine learning and information retrieval, positive predictive value and sensitivity are called precision and recall (Fawcett 2006) respectively. These measures have been used in SDM (Drake et al. 2006). The pair Se and Sp and the pair PPV and NPV are complementary to each other (Hand 2001).

Positive and negative likelihood ratios (PLR and NLR) are two indices frequently used in medical diagnostic tests. PLR is the ratio of predicted presence sites among the real presence sites to that among the real absence sites, and NLR is the ratio of predicted absence sites among the real presence sites to that among the real absence sites (Glas et al. 2003). When false positives are zero, PLR is undefined; when true negatives are zero, NLR is undefined. Likelihood ratios are purported to be more efficient and more powerful performance measures than sensitivity, specificity and predictive values alone, by combining the attributes of both sensitivity and specificity into one index (Riddle and Stratford 1999).

Single overall measures of model performance are generally preferred by researchers. Overall accuracy (a_o), defined as the probability that a site (either presence or absence) is correctly predicted, is the most common measure used in various disciplines including ecology (Fielding and Bell 1997). Its application can be traced back to Finley (1884) who employed this measure for assessing the accuracy of tornado activity forecasts.

Table 1. Confusion table with sample parameters, where n is total number of sites, n_{+j} is the number of sites predicted as class j (j = 0, 1), n_{i+} is the number of sites observed as class i (i = 0, 1), n_{ij} is the number of sites observed as class i and predicted as class j, and class 0 is absence and class 1 is presence.

		Observed		
		Presence	Absence	Total
Predicted	Presence	n_{11}	n_{01}	n_{+1}
	Absence	n_{10}	n_{00}	n_{+0}
	Total	n_{1+}	n_{0+}	n

Table 2. Threshold-dependent accuracy measures. See Table 1 for the explanation of the basic parameters.

Index	Definition	Reference
Overall accuracy	$a_o = (n_{11} + n_{00})/n$	Finley (1884)
Sensitivity (recall)	$Se = n_{11}/n_{1+}$	Fielding and Bell (1997)
Specificity	$Sp = n_{00}/n_{0+}$	Fielding and Bell (1997)
Positive predictive value	$PPV = n_{11}/n_{+1}$	Fielding and Bell (1997)
Negative predictive value	$NPV = n_{00}/n_{+0}$	Fielding and Bell (1997)
Positive likelihood ratio	$PLR = Se/(1 - Sp)$	Glas et al. (2003)
Negative likelihood ratio	$NLR = (1 - Se)/Sp$	Glas et al. (2003)
True skill statistic	$TSS = Se + Sp - 1$	Peirce (1884)
F measure	$F = (\beta^2 + 1)/(\beta^2/Se + 1/PPV) = (\beta^2 + 1)n_{11}/(\beta^2 n_{1+} + n_{+1})$	Daskalaki et al. (2006)
Odds ratio	$OR = (n_{11}n_{00})/(n_{10}n_{01})$	Glas et al. (2003)
Yule's Y	$Y = (\sqrt{OR} - 1)/(\sqrt{OR} + 1) = (\sqrt{n_{11}n_{00}} - \sqrt{n_{10}n_{01}})/(\sqrt{n_{11}n_{00}} + \sqrt{n_{10}n_{01}})$	Kraemer (2006)
Yule's Q	$Q = (OR - 1)/(OR + 1) = (n_{11}n_{00} - n_{10}n_{01})/(n_{11}n_{00} + n_{10}n_{01})$	Kraemer (2006)
Phi coefficient	$\phi = \dfrac{n_{11}n_{00} - n_{01}n_{10}}{\sqrt{n_{1+}n_{0+}n_{+1}n_{+0}}}$	Kraemer (2006)
Kappa	$\kappa = (a_o - a_e)/(1 - a_e)$ where $a_e = (n_{1+}n_{+1} + n_{0+}n_{+0})/n^2$	Cohen (1960)
Normalized mutual information	$NMI = (H_o - H_{o\|p})/H_o$ where $H_o = (n\log n - n_{1+}\log n_{1+} - n_{0+}\log n_{0+})/n$ $H_{o\|p} = (n_{+1}\log n_{+1} + n_{+0}\log n_{+0} - \sum_{i=0}^{1}\sum_{j=0}^{1} n_{ij}\log n_{ij})/n$	Finn (1993)
Extreme dependency score	$EDS = 2\ln(n_{1+}/n)/\ln(n_{11}/n) - 1$	Stephenson et al. (2008)

Cohen's (1960) kappa is another widely used measure in various disciplines including SDM. It has been adopted to alleviate the problem of overestimating accuracy with a_o. It measures the extent to which the agreement between observed and predicted is higher than that expected by chance alone. This measure has been used in meteorology since the late of 1800s (Murphy 1996), and it is commonly called Heidke's skill score in that field (Stephenson 2000).

The odds ratio (OR) is a familiar measure in epidemiology (Glas et al. 2003), and is defined as the ratio of the odds of positivity in the presences relative to the odds of positivity in the absences, or the ratio of the odds of positivity in predicted presences relative to the odds of positivity in predicted absences. This index has also been introduced to SDM (Fielding and Bell 1997), and has been used in a few studies (Manel et al. 2001). OR is unbounded and is undefined when either false positives or false negatives are zero, which is not an unusual situation, especially for models with high accuracy. In this case, adding 0.5 to each of the four cells of Table 1 is a common practice to calculate an approximation of the OR (Glas et al. 2003). However, two closely related measures, Yule's Y and Yule's Q, have better properties than OR. They can be considered as correlation coefficients ranging from -1 to 1 (Kraemer 2006). Furthermore, simple manipulation solves the "no definition" problem (Table 2). Yule's Q has also been termed the Gamma coefficient (Kraemer 2006) and odds ratio skill score (Stephenson 2000).

F-measure, which is the weighted harmonic mean of precision and recall (Daskalaki et al. 2006), is widely used in machine learning field, especially when the parameter $\beta = 1$ (Fawcett 2006). This measure has been used in SDM (Drake et al. 2006). Prescribed in this common way, the F-measure will be undefined when all sites are predicted as one class (i.e. absence), as Drake et al. (2006) encountered. This can be resolved by some simple algebraic manipulation as presented in Table 2.

Stimulated by Gilbert's (1884) remarks on the accuracy of Finley's (1884) tornado forecasts, Peirce (1884) proposed a "measure of the science of the method", which is the difference between true positive rate and false negative rate. This metric has more recently been "rediscovered" and reworked (Stephenson 2000), and termed variously "Hanssen-Kuipers discriminant" or "Kuipers' performance index" (Hanssen and Kuipers 1965), "true skill statistic" (TSS) (Flueck 1987), or "Pierce skill score" (Stephenson 2000). It has recently been introduced to SDM by Allouche et al. (2006) using the TSS term.

TSS is also equivalent to Youden's index J, which was developed by Youden (1950) and is widely used in medical diagnostic tests. It is defined as the average of the net prediction success rate for presence sites and that for absence sites. It has gained considerable theoretical interest over many years (Böhning et al. 2008), and it is considered the best available summary measure of model performance in medical diagnostic tests by some researchers (Biggerstaff 2000). This index is closely related to the arithmetic mean of sensitivity and specificity (Table 2).

The Phi coefficient (ϕ) is a counterpart of the Pearson product correlation coefficient that measures the strength of the relationship between two dichotomous variables (Kraemer 2006). To our knowledge, it has not been used in SDM. But its intuitive meaning makes it attractive for this application.

The normalized mutual information (NMI) was introduced to ecology by Fielding and Bell (1997) as well as Couto (2003), and used in SDM by Manel et al. (2001). NMI is undefined whenever there is a zero in any cell of the confusion matrix. However, this problem can easily be solved if we take $\lim_{x \to 0} x \ln x$, which resolves to 0 (Finn 1993), instead of calculating $0 \ln 0$ directly, which is

undefined. However, as Liu et al. (2007) discussed, NMI has some weaknesses. It only measures the agreement between two patterns. It cannot differentiate "worse-than-random" models from "better-than-random" models, and as a result is not considered a useful accuracy measure.

Rare species are a major conservation concern to ecologists and management agencies. In addition to difficulties with predictive modelling of their distribution, assessment of accuracy for these models is challenging. It is well-recognized that accuracy measures without chance-adjustment are not suitable in this situation, and that chance-adjusted measures are also problematic in this situation since these measures have trivial, non-informative limits as species becomes rarer (Stephenson et al. 2008). In order to address this issue, the extreme dependency score (EDS) was introduced by Coles et al. (1999). It falls in the range $[-1, 1]$, where -1 corresponds to the worst predictions, 0 to random predictions, and 1 to perfect predictions. However, EDS does not properly use the information about false presences and true absences. Therefore, it should be provided together with the ratio of predicted presences to true presences in order to give a complete summary of model performance (Stephenson et al. 2008). EDS has been used in meteorology (Stephenson et al. 2008), and has the potential to be applied to SDM.

Threshold-independent measures

Threshold-independent accuracy measures are shown in Table 3. Area under the receiver operating characteristic curve (AUC) is one of the most widely used accuracy measures in various disciplines, including ecology (Raes and ter Steege 2007). In the context of SDM, the AUC of a model is equivalent to the probability that the model will rank a randomly chosen presence site higher than a randomly chosen absence site (Pearce and Ferrier 2000). This is equivalent to the Wilcoxon test of ranks (Hanley and McNeil 1982). AUC can also be interpreted as the average value of sensitivity over all possible values of specificity or alternatively as the average value of specificity over all possible values of sensitivity (Jiang et al. 1996). The AUC is also closely related to the Gini coefficient (Breiman et al. 1984), which is twice the area between the diagonal and the ROC curve. The Gini coefficient is a correlation

Table 3. Threshold-independent accuracy measures. See Table 1 for the explanation of the basic parameters.

Index	Definition	Reference
Maximum overall accuracy	$a_{max} = \max(a_o)$	Stockwell and Peterson (2002)
Maximum kappa	$\kappa_{max} = \max(\kappa)$	Guisan et al. (1998)
Maximum vertical distance	$MVDr = \max(TSS)$	Lee (1999)
Area under ROC curve (AUC)	$\hat{\theta} = \frac{1}{n_{1+}n_{0+}} \sum_{i=1}^{n_{0+}} \sum_{j=1}^{n_{1+}} \phi(X_i, Y_j)$ where $\phi(X,Y)$ equals 1 if $Y > X$, 1/2 if $Y = X$, and 0 otherwise; X_i and Y_j are the predicted value for the absence site i and presence site j.	Mason and Graham (2002)
Partial area under ROC curve (PAUC)	See the text	He and Escobar (2008)
Gini index	$Gini = 2AUC - 1$	Hand and Hill (2001)
Point biserial correlation coefficient	$r_{pb} = \frac{\bar{P}_1 - \bar{P}_0}{\sigma} \sqrt{\frac{n_{1+}n_{0+}}{n^2}}$ where, $\bar{P}_1 = \sum_{o_i=1} p_i/n_{1+}$, $\bar{P}_0 = \sum_{o_i=0} p_i/n_{0+}$, $\sigma = \left[\sum_{i=1}^{n} (P_i - \bar{P})^2/n\right]^{1/2}$, $\bar{P} = \sum_{i=1}^{n} p_i/n$, p_i and o_i are the predicted value and observed value (1 for presence and 0 for absence) for site i.	Tate (1954)
Rank biserial correlation coefficient	$r_{rb} = 2(\bar{R}_1 - \bar{R}_2)/n$ where $\bar{R}_1$ and $\bar{R}_0$ are the mean ranks for the predicted values of the presence and absence sites respectively.	Glass (1966)
Proportion of explained deviance	$D^2 = 1 - V/V_0$ where, $V = -2 \sum_{i=1}^{n} [o_i \log p_i + (1 - o_i)\log(1 - p_i)]$ $V_0 = -2n[p \log p + (1-p) log (1-p)]$, $p = n_{1+}/n$	Mittlböck and Schemper (1996)
Adjusted proportion of explained deviance	$D^2_{adj} = 1 - \frac{n-1}{n-m}(1 - D^2)$	Guisan and Zimmermann (2000)
Mean square error	$MSE = \frac{1}{n} \sum_{i=1}^{n} (p_i - o_i)^2$	Brier (1950)
Root mean square error	$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (p_i - o_i)^2}$	Caruana and Niculescu-Mizil (2004)
Coefficient of determination	$R^2 = 1 - \frac{1}{n} \sum_{i=1}^{n} (p_i - o_i)^2 \Big/ [p(1-p)]$	Ash and Shwartz (1999)
Mean absolute prediction error	$MAPE = \frac{1}{n} \sum_{i=1}^{n} \lvert p_i - o_i \rvert$	Schemper (2003)
Mean cross entropy	$MXE = -\frac{1}{n} [\sum_{o_i=1} \ln p_i + \sum_{o_i=0} \ln(1 - p_i)]$	Caruana and Niculescu-Mizil (2004)

coefficient, unlike AUC (Hand and Hill 2001, Kraemer 2006), and has been used in SDM by Engler et al. (2004).

AUC has been criticized by some researchers as it can give a misleading picture of model performance since it covers parts of the prediction range that are of no practical use (Briggs and Zaretzki 2008, Lobo et al. 2008). Therefore, partial AUC (i.e. PAUC) was proposed (McClish 1989), which is the part of the area under the ROC curve with false positives falling in the range [FPR_l, FPR_h]. Dividing the PAUC by ($FPR_h - FPR_l$) gives the normalized partial area (nPAUC), which can be interpreted as the average sensitivity over a fixed range of false positive rates [FPR_l, FPR_h] (He and Escobar 2008).

Jiang et al. (1996) proposed a further measure, called partial area index, which is the area under the ROC curve, but above a pre-selected sensitivity (Se_0) divided by the constant ($1 - Se_0$). This metric can be interpreted as the average value of specificity over all values of sensitivity between Se_0 and 1.

However, the choice of false positive range for McClish's method and Se_0 for Jiang et al.'s method needs to be made on a case-by-case basis, and neither of these methods allow for probabilistic interpretations (Lee and Hsiao 1996). Since the two methods are conceptually the same and McClish's method is more flexible, it will be discussed in detail elsewhere in this paper.

Lee and Hsiao (1996) and Lee (1999) introduced other accuracy indices, among which *MVDr* (i.e. the maximum vertical distance between the ROC curve and the diagonal line) is a promising one. It can be shown that $MVDr = \max(TSS)$.

The maximum overall accuracy and maximum kappa are frequently used in SDM in a threshold-independent way to indicate a model's predictive capacity (Guisan et al. 1998, Liu et al. 2005). The point biserial correlation coefficient (r_{pb}) has also been used in SDM (Elith et al. 2006), which is the Pearson product moment correlation coefficient, calculated under the condition that one variable (i.e. the observed species occurrence) is dichotomous while the remaining variable (i.e. the predicted probability) is continuous (Kraemer 2006). A closely related measure, rank biserial correlation coefficient (r_{rb}), has also been applied in SDM (Phillips et al. 2009), which is the Spearman rank correlation coefficient, calculated under the condition that one variable is dichotomous and the other is ordinal (Kraemer 2006). Glass (1966) modified it to improve its statistical behaviour.

Guisan and Zimmermann (2000) introduced the proportion of explained deviance (D^2) and its adjusted form into ecology to assess the performance of generalized linear models, and the latter has been used in subsequent studies (Engler et al. 2004). While adjustment for the number of explanatory variables is useful in the model-building stage in order to avoid over-fitting, and to obtain a parsimonious model, it seems unnecessary when the model is assessed using independent test data.

Other threshold-independent measures include Brier's (1950) score, which is actually the mean square error (MSE), and Brier's skill score, which is equivalent to the coefficient of determination (R^2). Both metrics are widely used in meteorology for forecast verification (Bradley et al. 2008). R^2 has also been recommended for generalized regression model performance assessment (Ash and Shwartz 1999). The square root of R^2 – a correlation measure was proposed by Zheng and Agresti (2000) to summarize the predictive power of a generalized linear model. Mean absolute prediction error (MAPE) (Schemper 2003), the root mean square error (RMSE) and mean cross entropy (MXE) (Caruana and Niculescu-Mizil 2004) have also been used as accuracy measures in other fields and have potential application in SDM.

Methods of estimation and statistical inference for accuracy measures

Methods for simple proportion measures

Methods for assessment of one model

Some simple measures are just binomial proportions, including sensitivity, specificity, and overall accuracy, as well as false positive and false negative error rates. Suppose sample size is m and the "event" (true/false positive or negative) of interest occurs k times, then the measure of interest θ is estimated as $\hat{\theta} = k/m$, and the variance can be estimated as $\hat{\sigma}^2 = \hat{\theta}(1-\hat{\theta})/m$ (Kraemer 1992). For example, if the measure is sensitivity ($\theta = Se$), $k = n_{11}$, $m = n_{1+}$, then $\hat{\theta} = n_{11}/n_{1+}$, and $\hat{\sigma}^2 = n_{11}n_{10}/n_{1+}^3$.

Under the above conditions, k has a binomial distribution with parameters m and θ, and the Clopper-Pearson confidence interval (CI_{CP}) for θ, (θ_L, θ_U), with a coverage probability – the proportion of the time that the CI contains the true value of interest – of at least $1-\alpha$ can be obtained from this distribution (Seaman et al. 1996, Brown et al. 2001, Pires and Amado 2008).

If m is not too small, and $\hat{\theta}$ not too close to 0 or 1 so that as a rule of thumb $\min(m\hat{\theta}, m(1-\hat{\theta})) > 5$, then the CI for the measure θ can be calculated from the asymptotic normal approximation to the binomial distribution with mean θ estimated as $\hat{\theta}$ and variance σ^2 estimated as $\hat{\sigma}^2 = \hat{\theta}(1-\hat{\theta})/m$ (Seaman et al. 1996), i.e. $(\hat{\theta}-\theta)/\hat{\sigma}$ is asymptotically normally distributed, which gives the $100(1-\alpha)$ % CI of the measure θ as $\hat{\theta} \mp z_{1-\alpha/2}\hat{\sigma}$, where $z_{1-\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the standard normal distribution (Brown et al. 2001), e.g. if $\alpha = 0.05$, $z_{1-\alpha/2} = 1.96$; the upper operator corresponds to the lower limit of the CI, and the lower operator corresponds to the upper limit of the CI. This denotation will be used throughout the paper. When the lower limit goes below 0, it will be replaced with 0; and when the upper limit goes above 1, it will be replaced with 1 (Pires and Amado 2008). The interval formed in this way is called the standard Wald confidence interval (CI_{sW}). In order for the estimate of the measure to be reasonably unbiased and the estimated variance reasonably accurate, Kraemer (1992) suggested that each marginal sum (i.e. n_{i+} and $n_{+i}, i = 0,1$) must be at least 10.

However, it has been criticized that the coverage probabilities of the CI_{sW} can be erratically poor even when the proportion θ is not close to 0 or 1 (Brown et al. 2001). Therefore, various alternative CIs have been proposed (Vollset 1993, Newcombe 1998a, Brown et al. 2001, Pires and Amado 2008). After comparing eleven methods, Brown et al. (2001) recommended the Wilson interval (CI_W) for small samples and the Agresti-Coull interval

(CI_{AC}) for large samples. They also showed that CI_W also has better performance than the exact Clopper-Pearson interval. CI_W has been used in weather forecasting (Stephenson 2000). Vollset (1993) recommended the CI_W with continuity correction (CI_{Wcc}) from comparison of seventeen methods. From comparison of seven methods, Newcombe (1998a) recommended CI_W and CI_{Wcc}, while a comparison of twenty methods led Pires and Amado (2008) to recommend CI_{AC}.

Methods for comparison of two models

For those indices that are in the form of proportions, the difference of each index between two models can be compared statistically. Two situations will be considered: 1) the two samples used to assess the two models are independent, i.e. the test sites used for the two models are different; and 2) the two samples are dependent, and we only consider the most common situation that the same set of sites are used for the two models. The corresponding data are called independent data and paired (or here, interchangeably, dependent) data in the two situations respectively.

For the situation with independent data, suppose sample sizes are m_1 and m_2, and the "event" of interest occurs k_1 and k_2 times for the two samples. The measure of interest θ for the two samples is estimated as $\hat{\theta}^{(1)} = k_1/m_1$ and $\hat{\theta}^{(2)} = k_2/m_2$, and the difference $d = \theta^{(1)} - \theta^{(2)}$ can be estimated as $\hat{d} = \hat{\theta}^{(1)} - \hat{\theta}^{(2)}$ with the variance estimated as $\hat{\sigma}^2 = k_1(m_1 - k_1)/m_1^3 + k_2(m_2 - k_2)/m_2^3$ (Newcombe 1998b). With this variance estimate, a standard Wald confidence interval for the difference can be constructed, but it behaves very poorly (Newcombe 1998b). Among the eleven methods compared by Newcombe (1998b), he found that the two methods based on the tail area profile likelihood achieve the best coverage properties, but the calculation is complex; fortunately, two methods based on the Wilson score (CI_{Ws} for without continuity correction and CI_{Wscc} for with continuity correction) perform well and are easy to calculate.

For the situation with paired data, a contingency table (Table 4) can be constructed. In this table, we use m, instead of n, to denote the sample size, i.e. the number of sites used in the calculation of the target index. Only when the overall accuracy (OA) is the target index, $m = n$; for other indices, $m \neq n$. For example, if the target index is sensitivity (Se), $m = n_{1+}$; and if the target index is specificity (Sp), $m = n_{0+}$.

Table 4. Contingency table with sample parameters, where m is the number of "sites" (NS) involved in the calculation of the target index. The "sites" mean the presence sites for sensitivity, absence sites for specificity, and general sites for the overall accuracy. e is the NS predicted correctly by both models, h is the NS predicted incorrectly by both models, f is the NS predicted correctly by model 1 and incorrectly by model 2, and g is the NS predicted incorrectly by model 1 and correctly by model 2.

		NS predicted by model 1		
		Correct	Incorrect	Total
NS Predicted by model 2	Correct	e	g	e+g
	Incorrect	f	h	f+h
	Total	e+f	g+h	m

For the data in Table 4, the index for model 1 and 2 can be estimated as $\hat{\theta}^{(1)} = (e+f)/m$ and $\hat{\theta}^{(2)} = (e+g)/m$ respectively. The difference of the index between the two models, $d = \theta^{(1)} - \theta^{(2)}$, can be estimated as $\hat{d} = \hat{\theta}^{(1)} - \hat{\theta}^{(2)} = (f-g)/m$. The variance of this difference d can be estimated as $\hat{\sigma}^2 = [(e+h)(f+g) + 4fg]/m^3$ (Newcombe 1998c) or equivalently $\hat{\sigma}^2 = (1/m^2)[f+g-(f-g)^2/m]$ (Hawass 1997). With this variance estimation, the Wald confidence interval for the difference d can be constructed as $\hat{d} \mp z_{1-\alpha/2}\hat{\sigma}$, where $z_{1-\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the standard normal distribution (Newcombe 1998c). This method can be applied only if the sample size is over 40; below 40, this should only be applied if the sample size is above 20, and that $m\hat{\theta}^{(1)}$, $m(1-\hat{\theta}^{(1)})$, $m\hat{\theta}^{(2)}$, $m(1-\hat{\theta}^{(2)})$ are all greater than 5; otherwise, the critical value from the Student's t distribution should replace the normal critical value $z_{1-\alpha/2}$ (Hawass 1997). However, as stated by Newcombe (1998c), this method can produce confidence limits outside the range of validity. After comparing ten methods, he found that the methods based on the profile likelihood are superior to the others, and a computationally simpler method based on the Wilson score interval (CI_{WsP}) also performed well.

For paired data, the difference in accuracy between two models $d = \theta^{(1)} - \theta^{(2)}$ can also be tested with the McNemar test, which tests the null hypothesis H_0: $\theta^{(1)} = \theta^{(2)}$ with the test statistic $\chi_1^2 = (f-g)^2/(f+g)$, which has an asymptotic χ^2 distribution with one degree of freedom (Hawass 1997). The statistic with a continuity correction $\chi_1^2 = (|f-g|-1)^2/(f+g)$ may be preferred (Hawass 1997). This method can be applied if the sample size is greater than 40; otherwise, it should be applied only when the sample size is between 20 and 40, and $f+g \geq 10$ (Hawass 1997).

It is better to simultaneously test the significance of difference of both sensitivities and specificities between two models than to separately test for sensitivity and specificity, since the type I error of the test can be better controlled in a simultaneous test (Hawass 1997). The null hypothesis that $Se^{(1)} = Se^{(2)}$ and $Sp^{(1)} = Sp^{(2)}$ can be tested by the extended McNemar test $\chi_2^2 = (c-a)^2/(c+a) + (d-b)^2/(d+b)$, where a, b, c, and d are from another contingency table (Table 5) (Hawass 1997).

Methods for the Kappa statistic

The standard Wald confidence interval for the kappa coefficient (κ) can be calculated as $\kappa \mp z_{1-\alpha/2}\sigma$, where σ is the standard deviation of κ (Blackman and Koval 2000),

Table 5. Contingency table used in the extended McNemar test for the simultaneous sensitivities and specificities significance test with paired data. T, F, P and A stand for true, false, presence and absence repectively. "-" means that there is no value for the cell.

Model 2	Model 1			
	TP	FP	FA	TA
TP	x	-	c	-
FP	-	y	-	d
FA	a	-	u	-
TA	-	b	-	v

which can be estimated as $\hat{\sigma} = \sqrt{a_o(1-a_o)/n}/(1-a_e)$ (Cohen 1960, Kundel and Polansky 2003) (see Table 2 for a_o and a_e). The asymptotic variance (σ^2) of κ can also be estimated in several other ways: $\hat{\sigma}^2_{FCE}$ (Fleiss et al. 1969, see also Hanley 1987), $\hat{\sigma}^2_{BK}$ (Bloch and Kraemer 1989), and $\hat{\sigma}^2_G$ (Garner 1991). Jackknife method has also been used to calculate the variance ($\hat{\sigma}^2_J$) of kappa (Blackman and Koval 2000).

After comparing the lower bound of the confidence intervals based on the above four variance estimates, Blackman and Koval (2000) found that there is no uniformly best method, and the accuracy of estimation depends on the species' prevalence, sample size and true degree of agreement. When degree of agreement is moderate ($0.4 \le \kappa < 0.6$), prevalence is extreme ($P \le 0.2$ or $P \ge 0.8$), and sample size is large ($n \ge 40$), the jackknife method is recommended; when the degree of agreement is slight to fair ($0 \le \kappa < 0.4$), prevalence is not too extreme ($0.1 < P < 0.9$), and sample size is not too small ($n \ge 20$), BK and FCE methods are recommended; when the degree of agreement is at least moderate ($\kappa \ge 0.4$), prevalence is not too extreme ($0.1 < P < 0.9$), and sample size is not too small ($n \ge 20$), the Garner method is recommended (Blackman and Koval 2000).

As Blackman and Koval (2000) stated, none of the above four methods perform well over all parameter values; and under extreme values of the true degree of agreement and species prevalence as well as in small samples, non-symmetric intervals may be used.

Flack (1987) compared the performance of four confidence interval estimation methods including the above FCE method, the Edgeworth skewness correction method (which is a jackknife-based method), the logarithmic transformation method, and the square-root transformation method. He found that the square-root transformation method provides confidence intervals with a better coverage rate, which is $1 - [\sqrt{1-\kappa} \pm z_{1-\alpha/2}\hat{\sigma}_{FCE}/(2\sqrt{1-\kappa})]^2$.

Hale and Fleiss (1993) compared three methods (including the above FCE method, the Edgeworth skewness correction method, and the Cornfield's test-based method), and found that the Cornfield's test-based method is better than the others. However, Lee and Tu (1994) compared the square-root transformation method, the FCE method and other two methods based on profile variance, and recommended a profile-variance-based method.

The variance and confidence intervals for kappa were also estimated with (the approximate) bootstrap method (Fung and Lee 1991). For small samples, the exact bootstrap confidence intervals can be constructed for κ (Klar et al. 2002).

The null hypothesis that $\kappa = 0$ can be tested using the statistic $Z = \hat{\kappa}/\sigma_0$, which has an asymptotic standard normal distribution under the null hypothesis, and σ_0 can be estimated as $\hat{\sigma}_0 = \sqrt{a_e/[n(1-a_e)]}$ (Cohen 1960, Sheskin 2007).

For two kappas $\hat{\kappa}^{(1)}$ and $\hat{\kappa}^{(2)}$ calculated from two independent samples, their equivalence can be tested with the statistic $z = (\hat{\kappa}^{(1)} - \hat{\kappa}^{(2)})/\sqrt{\hat{\sigma}^2_1 + \hat{\sigma}^2_2}$ (Cohen 1960, Sheskin 2007), which has an asymptotic standard normal distribution under the null hypothesis that the two kappas are equal, where $\hat{\sigma}^2_1$ and $\hat{\sigma}^2_2$ are the variance of the two kappas calculated with the FCE method.

For two kappas $\hat{\kappa}^{(1)}$ and $\hat{\kappa}^{(2)}$ calculated from the same set of sites, their equivalence can be tested with the statistic (Barnhart and Williamson 2002) $z = (\hat{\kappa}^{(1)} - \hat{\kappa}^{(2)})/\sqrt{\hat{\sigma}^2_1 + \hat{\sigma}^2_2 - 2\hat{\sigma}_{12}}$, which has an asymptotic standard normal distribution under the null hypothesis that the two kappas are equal, where $\hat{\sigma}^2_1$ and $\hat{\sigma}^2_2$ are the variance of the two kappas, which can be estimated with the FCE method as well as weighted least-squares (WLS) method introduced by Barnhart and Williamson (2002), and $\hat{\sigma}_{12}$ is the covariance between the two kappas, which needs to be calculated with WLS method.

Williamson et al. (2000) also used generalized estimating equation (GEE) method to compare two correlated kappas. McKenzie et al. (1996) used bootstrap and randomization methods to compare two correlated kappas. Vanbell and Albert (2008) used bootstrap method to estimate the mean and variance-covariance matrix of several correlated kappas, and hence provided a way to test their equivalence using Hotelling's T^2 statistic. They also compared their method with Barnhart and Williamson's (2002) WLS method and Williamson et al.'s (2000) GEE method, and found that the estimates of kappa obtained with the bootstrap method were slightly biased, and the bootstrap method also yielded slightly higher standard errors than the WLS and GEE methods.

Methods for other threshold-dependent measures

It has been proved that $\hat{\sigma}^2[\ln(\theta_1/\theta_2)] = (1-\hat{\theta}_1)/(m_1\hat{\theta}_1) + (1-\hat{\theta}_2)/(m_2\hat{\theta}_2)$ if θ_1 and θ_2 are two proportion measures from two independent samples (Agresti 2002). Taking $\theta_1 = Se$ and $\theta_2 = 1 - Sp$, we get $\hat{\sigma}^2(\ln PLR) = 1/n_{11} - 1/n_{1+} + 1/n_{01} - 1/n_{0+}$; and taking $\theta_1 = 1 - Se$ and $\theta_2 = Sp$, we get $\hat{\sigma}^2(\ln NLR) = 1/n_{10} - 1/n_{1+} + 1/n_{00} - 1/n_{0+}$. Therefore, we can get the standard Wald confidence intervals for ln *PLR* and ln *NLR*.

With an exponential transformation, we can calculate the confidence intervals for *PLR* and *NLR*. That is, if we denote (L,U) as the confidence intervals for ln *PLR* and ln *NLR*, the confidence intervals for *PLR* and *NLR* can be expressed as (exp(*L*),exp(*U*)).

Through simple transformation, we know that $PPV = [1 + (1-P)P^{-1}PLR^{-1}]^{-1}$ and $NPV = [1 + P(1-P)^{-1}NLR]^{-1}$, where P is the species prevalence, which can be estimated as $\hat{P} = n_{1+}/n$. This means that *PPV* is a monotonically increasing function of *PLR*, and *NPV* is a monotonically decreasing function of *NLR*. Therefore, the confidence intervals for *PPV* and *NPV* can be easily obtained by transforming those for *PLR* and *NLR*. That is, the lower and upper limits of the confidence interval for *PPV* correspond to the lower and upper limits of the confidence interval for *PLR*, and the lower and upper limits of the confidence interval for *NPV* correspond to the upper and lower limits of the confidence interval for *NLR*.

Confidence intervals for the predictive values (PPV and NPV), the negative likelihood ratio (NLR) and the reciprocal of the positive likelihood ratio (PLR^{-1}) have also been derived by Li et al. (2007) using two methods: Fieller's approach and the delta method, which are asymptotically equivalent. Using the lower and upper limits, we can also derive the confidence intervals for PPV and NPV.

The odds ratio (OR) can be easily tested for significance by considering the natural logarithm of OR (i.e. ln *OR*), which is asymptotically normally distributed with a standard error estimated as (Stephenson 2000) $\hat{\sigma} = (n_{11}^{-1} + n_{10}^{-1} + n_{01}^{-1} + n_{00}^{-1})^{1/2}$. The asymptotic $100(1-\alpha)$ % confidence interval for the OR can be estimated as $(L,U) = OR \times \exp(\mp z_{1-\alpha/2}\hat{\sigma})$. From here we can easily construct confidence intervals for Yule's Q (i.e. Gamma statistic) and Yule's Y, because they are monotonically increasing functions of OR, which are $((L-1)/(L+1),(U-1)/(U+1))$ and $((\sqrt{L}-1)/(\sqrt{L}+1),(\sqrt{U}-1)/(\sqrt{U}+1))$, respectively.

When sample size is small, the exact confidence limits (L, U) for OR can be calculated by using an algorithm developed by Thomas (1971).

The asymptotic variance of the true skill statistic (TSS) has been derived in two different ways: $\hat{\sigma}_1^2(TSS)$ (Seaman et al. 1996, Stephenson 2000, Allouche et al. 2006) and $\hat{\sigma}_2^2(TSS)$ (Hanssen and Kuiper 1965). The two expressions are asymptotically equivalent (Seaman et al. 1996). Using these variance estimates, the standard Wald confidence intervals can be constructed. However, if TSS is close to 1, or if sample size is small, the upper limit of the confidence interval may exceed 1, in which case 1 is taken as the upper limit instead (Woodcock 1976).

As for the Phi coefficient ϕ, it has been shown that $n\phi^2 \sim \chi_1^2$, therefore, the χ^2 statistic with one degree of freedom can be used to test the hypothesis that $\phi = 0$ (Myers and Well 2003).

The asymptotic variance of NMI (Forbes 1995) and the standard deviation of EDS (Stephenson et al. 2008) have also been derived. Therefore, the standard Wald confidence intervals for the two measures can be calculated. More accurate confidence intervals for EDS have been derived by Ferro (2007) by fitting a bivariate extreme-value model.

Methods for AUC

Both parametric and nonparametric approaches have been used for estimation and statistical inference regarding AUC. While the parametric approaches have distributional assumptions (usually binormal), nonparametric methods make no such assumptions.

Let $\{X_i\}$ and $\{Y_j\}$ be the sets of model predicted values that correspond to the n_{0+} absence sites and n_{1+} presence sites, respectively ($i = 1,2,\ldots,n_{0+}$; $j = 1,2,\ldots,n_{1+}$). The AUC can be estimated nonparametrically as $\hat{\theta}$ (Table 3).

The commonly used formula for the standard error of the nonparametric estimate ($\hat{\theta}$) of AUC was introduced by Hanley and McNeil (1982), which is $\hat{\sigma} = \{[\hat{\theta}(1-\hat{\theta}) + (n_{1+}-1)(Q_1-\hat{\theta}^2) + (n_{0+}-1)(Q_2-\hat{\theta}^2)]/(n_{1+}n_{0+})\}^{1/2}$, where Q_1 and Q_2 can be calculated approximately as $Q_1 = \hat{\theta}/(2-\hat{\theta})$ and $Q_2 = 2\hat{\theta}^2/(1+\hat{\theta})$. Hanley and McNeil (1982) also used another method to calculate the standard error of the nonparametric estimate ($\hat{\theta}$) of AUC in a more complex way. When the standard error is estimated, the confidence interval for $\hat{\theta}$ can be constructed, and the variance and the confidence interval for the difference of two independent AUCs can also be calculated.

In order to compare two dependent AUCs estimated from the same set of sites, $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$, the covariance or correlation coefficient (r) between them must be estimated, since $\mathrm{cov}(\hat{\theta}^{(1)}, \hat{\theta}^{(2)}) = r\hat{\sigma}_1^2\hat{\sigma}_2^2$, where $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are the variances of the two estimated AUCs. Hanley and McNeil (1983) introduced a method to calculate the correlation coefficient r. Two intermediate rank correlation coefficients r_P and r_A between the two methods are calculated for the presence sites and absence sites respectively. The predicted values for the presence sites $\{Y_j^{(k)} | j = 1,2,\ldots,n_{1+}\}$ (k = 1,2) are used to calculate r_P, and the predicted values for the absence sites $\{X_i^{(k)} | i = 1,2,\ldots,n_{0+}\}$ (k = 1,2) are used to calculate r_A. From the average of the two correlation coefficients (r_P and r_A) and the average of the two AUCs ($\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$), the correlation coefficient r between the two dependent AUCs can be obtained from Hanley and McNeil's (1983) Table I.

DeLong et al. (1988) proposed a method to nonparametrically estimate the variance-covariance matrix of a vector of AUCs from several models, and therefore, the confidence intervals for the AUC, and for the difference of any two dependent AUCs can be constructed. Hanley and Hajian-Tilaki (1997) and Lasko et al. (2005) gave useful descriptions of this method.

Hanley and Hajian-Tilaki (1997) introduced the jackknife method for estimating the variance and covariance of two dependent AUCs using pseudo values. These values have the same average as the calculated AUC values, but they behave like independent identically distributed samples, so that the standard methods can be used to calculate the variances and confidence intervals (Lasko et al. 2005). A pseudo value U_i for a particular data point i is calculated by taking the weighted difference of the $\hat{\theta}$ using all the data points and the $\hat{\theta}_{-i}$ generated with all but the point i, that is $U_i = n\hat{\theta} - (n-1)\hat{\theta}_{-i}$. The variance of the $\hat{\theta}$ is estimated as $\mathrm{var}(\hat{\theta}) = \mathrm{var}(U)/n$, and the covariance between the two dependent AUCs is estimated as $\mathrm{cov}(\hat{\theta}^{(1)}, \hat{\theta}^{(2)}) = \mathrm{cov}(U^{(1)}, U^{(2)})/n$. The variance of the difference between the two dependent AUCs can be estimated similarly as in the above.

Bandos et al. (2005) developed an exact permutation test for the comparison of two dependent AUCs. Their method is based on the assumption of the exchangeability of within subject (i.e. site here) rank-ratings. Therefore, instead of the raw predicted values for each site, the ranks of these values are used in the calculation. The hypothesis that the two dependent AUCs (i.e. $\theta^{(1)}$ and $\theta^{(2)}$) are equal can be tested by the permutation method through simulation, or using the test statistic $(\hat{\theta}^{(1)} - \hat{\theta}^{(2)})/\sqrt{V_\Omega(\hat{\theta}^{(1)} - \hat{\theta}^{(2)})}$, which has an asymptotically normal distribution, with mean 0 and variance (under the null hypothesis) estimated as $V_\Omega(\hat{\theta}^{(1)} - \hat{\theta}^{(2)})$.

Qin and Hotilovac (2008) compared nine non-parametric methods for constructing confidence intervals of the AUC, including Mann-Whitney and logit-transformation-based confidence intervals, DeLong's non-parametric interval, empirical likelihood-based interval and three bootstrap intervals. They found that the empirical likelihood-based interval has nice asymptotic properties and good coverage accuracy; however, it is complex and will not be described here. In contrast, the bootstrap percentile-t interval is

slightly conservative, but has good coverage accuracy when $AUC \geq 0.95$.

Molodianovitch et al. (2006) compared DeLong et al.'s (1988) nonparametric method, Wieand et al.'s (1989) binormal-based parametric method, and their own transformed normal method for the comparison of two correlated AUCs. They found that their method performs best, and the nonparametric method is robust for all kinds of data they studied.

Methods for PAUC

The partial AUC (PAUC) can be estimated both parametrically and nonparametrically. The original methods proposed by McClish (1989) and Jiang et al. (1996) are parametric in nature. Wieand (1989) proposed a generalized nonparametric method for both AUC and PAUC, and their asymptotic variances. However, these calculations are mathematically complicated and thus have not been widely used (He and Escobar 2008). Zhang et al. (2002) proposed a much simpler method to estimate the variance-covariance matrix for PAUCs, based on the nonparametric approach of DeLong et al. (1988). However, their variance formula was found to be incorrect by He and Escobar (2008), by whom another nonparametric method was introduced.

He and Escobar (2008) compared this method with McClish's (1989) method and Zhang et al.'s (2002) method through simulation. They found that their method and McClish's method produced very similar variances that are very close to the simulated variances for data generated from both Gaussian distribution and non-Gaussian distribution. However, for small sample sizes the behaviour of these estimates is still uncertain as the comparisons were made using large samples (at least 200 presences and 200 absences).

Li et al. (2008) compared several methods to test two PAUCs, including McClish's (1989) parametric method based on maximum likelihood estimation, their own method based on generalized p-value, and a nonparametric method with the point estimate calculated using the method described in this section and with the variance calculated using a bootstrap method. They found their method is more powerful than the others, however, the method is not only parametric, but also very complicated, which hinders its wider use.

Methods for other threshold-independent measures

The estimated point biserial correlation coefficient $\hat{r}_{pb}$ was shown by Tate (1954) to be asymptotically normally distributed, with mean r_{pb} and variance $\sigma^2(\hat{r}_{pb}) = [4PQ - r_{pb}^2(6PQ - 1)](1 - r_{pb}^2)^2/(4nPQ)$, where species prevalence P can be estimated as $\hat{P} = n_{1+}/n$ and $Q = 1 - P$, and r_{pb} can be substituted by the estimated value $\hat{r}_{pb}$.

The hypothesis that $r_{pb} = 0$ can be tested with the statistic $t = \hat{r}_{pb}\sqrt{(n-2)/(1-\hat{r}_{pb}^2)}$, which has a t-distribution with $n-2$ degrees of freedom (Myers and Well 2003).

Bradley et al. (2008) derived the sampling variance for the mean square error, i.e. Brier's score, and the R^2 based on sum-of-squares, also known as Brier's skill score.

Conclusion

In this paper we have reviewed various accuracy measures currently used in species distribution modelling; and also introduced some promising measures that have not previously been applied in this field. Among them, the partial AUC (i.e. PAUC) is worthy of recommendation. It is partly free of those criticisms raised to the use of AUC. However, the subjectivity in choosing the range of false positive rate for its calculation hinders its comparability among different studies. Perhaps a fixed range applicable for most problems could be introduced to standardize its application, as has taken place within genomic studies. As reported by Gribskov and Robinson (1996), many researchers in genomics have adopted AUC_{50}, the area under the lower portion of the ROC curve up to the first 50 false positives (He and Escobar 2008). However, using the absolute value of false positives may not be suitable in SDM, and a relative value, i.e. a false positive rate may be better. Therefore, we tentatively suggest $[0, 0.5]$ as the range of false positive rates to be used for the calculation of PAUC. Additionally, the true positive rate in some range, e.g. [0.5,1] can be restricted (i.e. the calculated PAUC is the area under the portion of the ROC curve that is within the upper-left quarter of the plot square). However, for this statistic there is no appropriate analytic method for statistical inference, and computer-intensive methods need to be used for variance estimation and hypothesis testing.

Special attention was paid to reviewing the methods for estimating standard errors of accuracy measures, as well as for constructing confidence intervals for both single measures and the difference of a measure between two models. For most accuracy measures reviewed in this paper we provided methods to estimate their standard errors. Using these estimates we can conduct significance tests by calculating Z test statistics; and we can also construct the standard Ward confidence interval $\hat{\theta} \mp z_{1-\alpha/2}\hat{\sigma}$ for the accuracy measure θ. However, this is a very approximate approach as it relies on the asymptotic distribution, which means large samples are needed to guarantee accuracy. Methods that can be used for small samples are generally more attractive. They should be used if sample is not large enough, as is usually the case.

We suggest that confidence intervals should be provided for each accuracy measure used in accuracy assessment. Even when comparing two models, the confidence interval for the difference of the measure between the two models should be provided. Confidence intervals contain more information than statistical tests. For some measures including those not reviewed in this paper, if no suitable analytical methods exist, computer-intensive methods (such as bootstrap or randomization methods) can be used for both statistical tests and confidence interval construction.

However, it is not our intention that all the accuracy measures described in this paper should be used, since some measures contain very similar information as discussed by Liu et al. (2007). Further work is still needed to investigate the behaviours and their relations of the measures discussed in this paper. Before that work is done, we can not give a list of preferred measures. However, we provide the following general suggestion. For continuous modelling results, it is better to report both the discriminatory power of the model

(e.g. AUC, PAUC, κ_{max}, *MVDr*, r_{pb}) and the reliability of the model (e.g. D^2, R^2, *RMSE*); for binary modelling results, it is better to report both the producer's accuracy (i.e. Se and Sp) and the user's accuracy (i.e. NPV and PPV) in addition to other overall measurements (e.g. a_0, κ, TSS); and whatever accuracy measures are used, it is preferable to provide their confidence intervals.

Acknowledgements – We thank Alan H. Fielding, Miguel Araújo, Lars B. Pettersson, Michael Scroggie, and an anonymous reviewer for their thoughtful comments, which allowed us to improve the manuscript greatly. The authors are supported by the funding from Biodiversity and Ecosystem Services, Dept of Sustainability and Environment, Victoria, Australia.

References

Agresti, A. 2002. Categorical data analysis, 2nd ed. – Wiley.

Allouche, O. et al. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). – J. Appl. Ecol. 43: 1223–1232.

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

Ash, A. and Shwartz, M. 1999. R^2: a useful measure of model performance when predicting a dichotomous outcome. – Stat. Med. 18: 375–384.

Bandos, A. I. et al. 2005. A permutation test sensitive to differences in areas for comparing ROC curves from a paired design. – Stat. Med. 24 : 2873–2893.

Barnhart, H. X. and Williamson, J. M. 2002. Weighted least-squares approach for comparing correlated kappa. – Biometrics 58: 1012–1019.

Biggerstaff, B. J. 2000. Comparing diagnostic tests: a simple graphic using likelihood ratios. – Stat. Med. 19: 649–663.

Blackman, N. J. and Koval, J. J. 2000. Interval estimation for Cohen's kappa as a measure of agreement. – Stat. Med. 19: 723–741.

Bloch, D. A and Kraemer, H. C. 1989. 2×2 kappa coefficients: measures of agreement or association. – Biometrics 45: 269–287.

Böhning, B. et al. 2008. Revisiting Youden's index as a useful measure of the misclassification error in meta-analysis of diagnostic studies. – Stat. Methods Med. Res. 17: 543–54.

Bradley, A. A. et al. 2008. Sampling uncertainty and confidence intervals for the Brier score and Brier skill score. – Weather Forecast. 23: 992–1006.

Breiman, L. et al. 1984. Classification and regression trees. – Wadsworth International Group, Belmont, CA.

Brier, G. W. 1950. Verification of forecasts expressed in terms of probability. – Mon. Weather Rev. 78: 1–3.

Briggs, W. M. and Zaretzki, R. 2008. The skill plot: a graphical technique for evaluating continuous diagnostic tests. – Biometrics 63: 250–261.

Brown, L. D. et al. 2001. Interval estimation for a binomial proportion. – Stat. Sci. 16: 101–117.

Caruana, R. and Niculescu-Mizil, A. 2004. Data mining in metric space: an empirical analysis of supervised learning performance criteria. – In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 22–25, 2004, Seattle, WA, USA, pp. 69–78.

Cohen, J. 1960. A coefficient of agreement for nominal scales. – Educ. Psychol. Meas. 20: 37–40.

Coles, S. et al. 1999. Dependence measures for extreme value analyses. – Extremes 2: 339–365.

Couto, P. 2003. Assessing the accuracy of spatial simulation models. – Ecol. Model. 167: 181–198.

Daskalaki, S. et al. 2006. Evaluation of classifiers for an uneven class distribution problem. – Appl. Artif. Intell. 20: 381–417.

DeLong, E. R. et al. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a non-parametric approach. – Biometrics 44: 837–845.

Drake, J. M. et al. 2006. Modelling ecological niches with support vector machines. – J. Appl. Ecol. 43: 424–432.

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

Engler, R. et al. 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. – J. Appl. Ecol. 41: 263–274.

Fawcett, T. 2006. An introduction to ROC analysis. – Pattern Recogn. Lett. 27: 861–874.

Ferro, C. A. T. 2007. A probability model for verifying deterministic forecasts of extreme events. – Weather Forecast. 22: 1089–1100.

Fielding, A. H. 2007. Cluster and classification techniques for the biosciences. – Cambridge Univ. Press.

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

Finley, J. P. 1884. Tornado predictions. – Am. Meteorol. J. 1: 85–88.

Finn, J. T. 1993. Use of the average mutual information index in evaluating classification error and consistency. – Int. J. Inform. Syst. 7: 349–366.

Flack, V. 1987. Confidence intervals for the interrater agreement measure kappa. – Commun. Stat. Theory 16: 953–968.

Fleiss, J. L. et al. 1969. Large-sample standard errors of kappa and weighted kappa. – Psychol. Bull. 72: 323–327.

Flueck, J. A. 1987. A study of some measures of forecast verification. – In: 10th Conference on Probability and Statistics in Atomaspheric Sciences. Edmonton, AB, Canada, American Meteorological Society, pp. 69–73.

Forbes, A. D. 1995. Classification-algorithm evaluation: five performance measures based on confusion matrices. – J. Clin. Monit. 11: 189–206.

Fung, K. P. and Lee, J. 1991. Bootstrap estimate of the variance and confidence interval of kappa. – Br. J. Ind. Med. 48: 503–504.

Garner, J. B. 1991. The standard error of Cohen's kappa. – Stat. Med. 10: 767–775.

Gilbert, G. K. 1884. Finley's tornado predictions. – Am. Meterol. J. 1: 166–172.

Glahn, B. 2004. Discussion of verification concepts in forecast verification: a practitioner's guide in atmospheric science. – Weather Forecast. 19: 769–775.

Glas, A. S. et al. 2003. The diagnostic odds ratio: a single indicator of test performance. – J. Clin. Epidemiol. 56: 1129–1135.

Glass, G. V. 1966. Note concerning rank biserial correlation. – Educ. Psychol. Meas. 26: 623–631.

Gribskov, M. and Robinson, N. L. 1996. The use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. – Comput. Chem. 20: 25–33.

Guisan, A. and Zimmermann, N. E. 2000. Predictive habitat distribution models in ecology. – Ecol. Model. 135: 147–186.

Guisan, A. et al. 1998. Predicting the potential distribution of plant species in an Alpine environment. – J. Veg. Sci. 9: 65–74.

Hale, C. A. and Fleiss, J. L. 1993. Interval estimation under two study designs for kappa with binary classifications. – Biometrics 49: 523–534.

Hand, D. J. 1992. Statistical methods in diagnosis. – Stat. Methods Med. Res. 1: 49–67.
Hand, D. J. 2001. Measuring diagnostic accuracy of statistical prediction rules. – Stat. Neerl. 55: 3–16.
Hand, D. J. and Hill, R. J. 2001. A simple generalization of the area under the ROC curve to multiple class classification problems. – Mach. Learn. 45: 171–186.
Hanley, J. A. 1987. Standard error of the kappa statistic. – Psychol. Bull. 102: 315–321.
Hanley, J. A. and McNeil, B. J. 1982. The meaning eand use of the area under a receiver operating characteristic (ROC) curve. – Radiology 143: 29–36.
Hanley, J. A. and McNeil, B. J. 1983. A method of comparing the areas under receiver operating characterisitc curves derived from the same cases. – Radiology 148: 839–848.
Hanley, J. A. and Hajian-Tilaki, K. O. 1997. Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: an update. – Acad. Radiol. 4: 49–58.
Hanssen, A. J. and Kuipers, W. J. 1965. On the relationship between the frequency of rain and various meteorological parameters. – Meded Verhand 81: 2–15.
Hawass, N. E. D. 1997. Comparing the sensitivities and specificities of two diagnostic procedures performed on the same group of patients. – Br. J. Radiol. 70: 360–366.
He, Y. and Escobar, M. 2008. Nonparametric statistical inference method for partial areas under receiver operating characteristic curves, with application to genomic studies. – Stat. Med. 27: 5291–5308.
Heikkinen, R. K. et al. 2006. Methods and uncertainties in bioclimatic envelope modelling under climate change. – Pro. Phys. Geogr. 30: 1–27.
Jiang, Y. et al. 1996. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. – Radiology 143: 29–36.
Jolliffe, I. T. 2007. Uncertainty and inference for verification measures. – Weather Forecast. 22: 637–650.
Klar, N. et al. 2002. An exact bootstrap confidence interval for κ in small samples. – Statistician 51: 467–478.
Kraemer, H. C. 1992. Evaluating medical tests: objective and quantitative guidelines. – Sage Publ.
Kraemer, H. C. 2006. Correlation coefficients in medical research: from product moment correlation to the odds ratio. – Stat. Methods Med. Res. 15: 525–545.
Kundel, H. L. and Polansky, M. 2003. Measurement of observer agreement. – Radiology 228: 303–308.
Lasko, T. A. et al. 2005. The use of receiver operating characteristic curves in biomedical informatics. – J. Biomed. Inform. 38: 404–415.
Lee, J. J. and Tu, Z. N. 1994. A better confidence interval for kappa (κ) on measuring agreement between two raters with binary outcomes. – J. Comput. Graph. Stat. 3: 301–321.
Lee, W. C. 1999. Probabilistic analysis of global performances of diagnostic tests: interpreting the Lorenz curve-based summary measures. – Stat. Med. 18: 455–471.
Lee, W. C. and Hsiao, C. K. 1996. Alternative summary indices for the receiver operating characteristic curve. – Epidemiology 7: 605–611.
Li, C. R. et al. 2008. On the exact interval estimation for the difference in paired areas under the ROC curves. – Stat. Med. 27: 224–242.
Li, J. et al. 2007. Prevalence-dependent diagnostic accuracy measures. – Stat. Med. 26: 3258–3273.
Liu, C. et al. 2005. Selecting thresholds of occurrence in the prediction of species distributions. – Ecography 28: 385–393.
Liu, C. et al. 2007. Comparative assessment of the measures of thematic classification accuracy. – Remote Sens. Environ. 107: 606–616.
Lobo, J. M. et al. 2008. AUC: a misleading measure of the performance of predictive distribution models. – Global Ecol. Biogeogr. 17: 145–151.
Manel, S. et al. 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. – J. Appl. Ecol. 38: 921–931.
Mason, S. J. and Graham, N. E. 2002. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves statistical significance and interpretation. – Q. J. R. Meteorol. Soc. 128: 2145–2166.
McClish, D. K. 1989. Analyzing a portion of the ROC curve. – Med. Decis. Making 9: 190–195.
McKenzie, D. P. et al. 1996. Comparing correlated kappas by resampling: is one level of agreement significantly different from another? – J. Psychiatr. Res. 30: 483–92.
Mittlböck, M. and Schemper, M. 1996. Explained variation for logistic regression. – Stat. Med. 15: 1987–1997.
Molodianovitch, K. et al. 2006. Comparing the areas under two correlated ROC curves: parametric and non-parametric approaches. – Biom. J. 48: 745–757.
Murphy, A. H. 1996. The Finley affair: a signal event in the history of forecast verification. – Weather Forecast. 11: 3–20.
Murphy, A. H. and Winkler, R. L. 1984. Probability forecasting in meteorology. – J. Am. Stat. Assoc. 79: 489–500.
Myers, J. L. and Well, A. 2003. Research design and statistical analysis, 2nd ed. – Lawrence Erlbaum Associates, Hillsdale, NJ.
Newcombe, R. G. 1998a. Two-sided confidence intervals for the single proportion: comparison of seven methods. – Stat. Med. 17: 857–872.
Newcombe, R. G. 1998b. Interval estimation for the difference between independent proportions: comparison of eleven methods. – Stat. Med. 17: 873–890.
Newcombe, R. G. 1998c. Improved confidence intervals for the difference between binomial proportions based on paired data. – Stat. Med. 17: 2635–2650.
Pearce, J. and Ferrier, S. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. – Ecol. Model. 133: 225–245.
Pearson, R. G. et al. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. – J. Biogeogr. 34: 102–117.
Peirce, C. S. 1884. The numerical measure of the success of predictions. – Science 4: 453–454.
Phillips, S. J. et al. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. – Ecol. Appl. 19: 181–197.
Pires, A. M. and Amado, C. 2008. Interval estimators for a binomial proportion: comparison of twenty methods. – REVSTAT 6: 165–197.
Qin, G. and Hotilovac, L. 2008. Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test. – Stat. Methods Med. Res. 17: 207–221.
Raes, N. and ter Steege, H. 2007. A null-model for significance testing of presence-only species distribution models. – Ecography 30: 727–736.
Riddle, D. L. and Stratford, P. W. 1999. Interpreting validity indexes for diagnostic tests: an illustration using the Berg balance test. – Phys. Ther. 79: 939–948.
Schemper, M. 2003. Predictive accuracy and explained variation. – Stat. Med. 22: 2299–2308.
Seaman, R. et al. 1996. Confidence intervals for some performance measures of yes-no forecasts. – Aust. Meteorol. Mag. 45: 49–53.
Sheskin, D. J. 2007. Handbook of parametric and nonparametric statistical procedures, 4 ed. – Chapman and Hall/CRC.

Stephenson, D. B. 2000. Use of the ''odds ratio'' for diagnosing forecast skill. – Weather Forecast. 15: 221–232.
Stephenson, D. B. et al. 2008. The extreme dependency score: a non-vanishing measure for forecasts of rare events. – Meterol. Appl. 15: 41–50.
Stockwell, D. R. B. and Peterson, A. T. 2002. Effects of sample size on accuracy of species distribution models. – Ecol. Model. 148: 1–13.
Tate, R. F. 1954. Correlation between a discrete and a continuous variable: point-biserial correlation. – Ann. Math. Stat. 25: 603–607.
Thomas, D. G. 1971. Algorithm AS-36: exact confidence limits for the odds ratio in a 2×2 table. – Appl. Stat. 20: 105–110.
Thuiller, W. et al. 2005. Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. – Global Change Biol. 11: 2234–2250.
Vanbelle, S. and Albert, A. 2008. A bootstrap method for comparing correlated kappa coefficients. – J. Stat. Comput. Simul. 78: 1009–1015.
Vollset, S. E. 1993. Confidence intervals for a binomial proportion. – Stat. Med. 12: 809–824.

Download the Supplementary material as file E6354 from <www.oikos.ekol.lu.se/appendix>.

Wieand, S. 1989. A family of non-parametric statistic for comparing diagnostic markers with paired or unpaired data. – Biometrika 76: 585–592.
Williamson, J. M. et al. 2000. Modeling kappa for measuring dependent categorical agreement data. – Biostatistics 1: 191–202.
Woodcock, F. 1976. The evaluation of yes-no forecasts for scientific and administrative purposes. – Mon. Weather Rev. 104: 1209–1214.
Yerushalmy, J. 1947. Statistical problems in assessing methods of medical diagnosis. – Public Health Rep. 62: 1432–1449.
Youden, W. J. 1950. Index for rating diagnostic tests. – Cancer 3: 32–35.
Zhang, D. D. et al. 2002. A non-parametric method for the comparison of partial areas under ROC curves and its application to large health care data sets. – Stat. Med. 21: 701–715.
Zheng, B. and Agresti, A. 2000. Summarizing the predictive power of a generalized linear model. – Stat. Med. 19: 1771–1781.