18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

Measuring the accuracy of species distribution models: a review

Liu, C.[1], M. White[1] and G. Newell[1]

[1]Arthur Rylah Institute for Environmental Research, Department of Sustainability & Environment, 123 Brown Street, Heidelberg, Victoria 3084, Australia
Email: canran.liu@dse.vic.gov.au

Abstract: Species distribution models (SDMs) are empirical models relating species occurrence to environmental variables based on statistical or other response surfaces. Species distribution modeling can be used as a tool to solve many theoretical and applied ecological and environmental problems, which include testing biogeographical, ecological and evolutionary hypotheses, assessing species invasion and climate change impact, and supporting conservation planning and reserve selection. The utility of SDM in real world applications requires the knowledge of the model's accuracy. The accuracy of a model includes two aspects: discrimination capacity and reliability. The former is the power of the model to differentiate presences from absences; and the latter refers to the capability of the predicted probabilities to reflect the observed proportion of sites occupied by the subject species.

Similar methodology has been used for model accuracy assessment in different fields, including medical diagnostic test, weather forecasting and machine learning, etc. Some accuracy measures are used in all fields, e.g. the overall accuracy and the area under the receiver operating characteristic curve; while the use of other measures is largely restricted to specific fields, e.g. F-measure is mainly used in machine learning field, or is referred to by different names in different fields, e.g. "true skill statistic" is used in atmospheric science and it is called "Youden's J" in medical diagnostic field. In this paper we review those accuracy measures typically used in ecology. Generally, the measures can be divided into two groups: threshold-dependent and threshold-independent. Measures in the first group are used for binary predictions, and those in the second group are used for continuous predictions. Continuous predictions may be transformed to binary ones if a specific threshold is employed. In such cases, the threshold-dependent accuracy measures can also be used.

The threshold-dependent indices used in or introduced to SDM field include overall accuracy, sensitivity, specificity, positive predictive value, negative predictive value, odds ratio, true skill statistic, F-measure, Cohen's kappa, and normalized mutual information (NMI). However, since NMI only measures the agreement between two patterns, it cannot differentiate the worse-than-random models from the better-than-random models, which reduces its utility as an accuracy measure.

The threshold-independent indices used in or introduced to the SDM field include the area under the receiver operating characteristic curve (AUC), Gini index, and point biserial correlation coefficient. The proportion of explained deviance D^2 and its adjusted form have been also introduced into SDM field. But this adjusted metric has no theoretical foundation in the context of generalized linear modeling. Therefore, we provide another adjusted form, which was proposed by H. V. Houwelingen based on the asymptotic χ^2 distribution of the log-likelihood statistics. Its superiority over other related measures has been found through previous simulation studies. We also provide another analogous measure, the coefficient of determination R^2, which has had a long history in weather forecast verification and was also recommended for use in medical diagnosis. Though these measures D^2 and R^2 are routinely used to evaluate generalized linear models (GLMs), we argue that nothing prevents them from being applied to other GLM-like models.

In SDM accuracy assessment, discrimination capacity is often considered, but model reliability is frequently ignored. The primary reason for this is that no reliability measure has been introduced into the ecological literature. To meet this need we also suggest that root mean square error be used as a reliability measure. Its squared form, mean square error, has been used in meteorology for a long time, and is called Brier's score. We also discuss the effect of prevalence dependence of accuracy measures and the precision of accuracy estimates.

Keywords: *species distribution, prediction, accuracy measure, performance, prevalence*

1. INTRODUCTION

Species distribution models (SDMs) are used to predict the geographic range of a species from its occurrence and relevant environmental data. Species distribution modeling is essentially a binary classification problem with two training classes, presence and absence. 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 sites are given a 'probability' of being part of a species' distribution. It has become a useful tool for fundamental ecological and biogeographical research, and for biodiversity management and conservation (Guisan and Zimmermann, 2000). Model utility is dependant on an evaluation of performance. This is a critical element of model-building. Robust assessment of model performance identifies the relative strengths and weaknesses of models and delimits the range of uses to which they can be usefully applied.

There are two facets to measuring the accuracy of species distribution models: discrimination capacity and reliability (Pearce and Ferrier, 2000) though the former is generally viewed as more important than the latter (Ash and Shwartz, 1999). Discrimination capacity measures 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 describes the agreement between predicted probabilities of occurrence and the observed proportions of sites occupied by the subject 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 modeling result is continuous. When the modeling result is binary, only discrimination capacity can be assessed. A range of indices are used 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, called a threshold. These indices are called threshold-dependent indices. Indices that can be applied directly to continuous situations are called threshold-independent indices. If the threshold value is changed systematically, the optimal value of any threshold-dependent indices can be obtained. 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 ones. All threshold-dependent indices are based on some or all of the elements of the confusion table (Table 1).

Table 1. Confusion table with sample parameters, where n is the 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.

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

Fielding and Bell (1997) reviewed some accuracy measures that can potentially be used in SDM. Couto (2003) reviewed some accuracy measures in the context of general spatial simulation models. This paper aims to comprehensively review the accuracy measures used in SDM, and provide additional measures, especially with regard to calibration measures which are still largely ignored in SDM field. We also draw attention to some outstanding issues relevant to the measurement of accuracy that require further attention.

2. THRESHOLD-DEPENDENT INDICES

The threshold-dependent accuracy measures are shown in Table 2. Sensitivity (Se) and specificity (Sp) are widely used in many disciplines including SDM. Se and Sp are conditional probabilities. The former is the probability that the model correctly predicts an observation of a species at a site and the latter is the probability that a known absence site is correctly predicted. While Se and Sp are probabilities conditional on the observed pattern, 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 on the predicted pattern. PPV is the probability that a site predicted as present is actually present and NPV is the probability that a site predicted as absent is actually absent. Although these two indices are widely used in medical diagnostic tests, they are rarely applied to SDM. In the field of image classification, Se and Sp are referred to as producer's accuracy, and PPV and NPV are called user's accuracy (Liu et al., 2007). In the fields of machine learning and information retrieval, precision and recall (Fawcett, 2006) are used instead of PPV and Se. These measures have been used in SDM (e.g. Drake et al., 2006). The pair Se and Sp and the pair PPV and NPV are complementary to each other (Hand, 2001).

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

Cohen's (1960) kappa is another widely used measure in various disciplines including SDM. It has been adopted to overcome the problem of overestimating accuracy with OA. It measures the extent to which the agreement between observed and predicted is higher than that expected by chance alone. This measure was originally formulated by Doolittle (1888) and later was rediscovered and extended by Heidke in 1926. It is commonly used in meteorology where it is known as Heidke's skill score (Stephenson, 2000). The chance adjustment for sensitivity and specificity have also been devised (Coughlin and Pickle, 1992), however the definition of 'chance' is open to interpretation (Hand, 2001).

Odds ratio (OR) is a familiar measure in the epidemiologic field (Glas et al., 2003), which 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 (e.g. 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). This measure is closely related to Yule's Y, and Yule's Q, which 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 the machine learning field, especially when the parameter $\beta = 1$ (Fawcett, 2006). This measure has been used in SDM (e.g. Drake et al., 2006). Prescribed in this way the F-measure will be undefined when all sites are predicted as one category (either presence or absence), as Drake et al. (2006) encountered. This can be resolved by some simple algebraic manipulation. The resultant formula is presented in Table 2.

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

Index	Definition	Reference
Overall accuracy	$OA = (n_{11} + n_{00}) / n$	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)
True skill statistic	$TSS = Se + Sp - 1$	Peirce (1884)
F measure	$F = (\beta^2 + 1) / \left(\beta^2 / Se + 1 / PPV\right)$ $= (\beta^2 + 1) n_{11} / (\beta^2 n_{1+} + n_{+1})$	Daskalaki et al. (2006)
Odds ratio	$OR = n_{11} n_{00} / n_{10} n_{01}$	Glas et al. (2003)
Yule's Y	$Y = (\sqrt{OR} - 1) / (\sqrt{OR} + 1)$	Karemer (2006)
Yule's Q	$Q = (OR - 1) / (OR + 1)$	Karemer (2006)
Kappa	$Kp = (OA - EA) / (1 - EA)$ where $EA = (n_{1+} n_{+1} + n_{0+} n_{+0}) / n^2$	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)

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. It has also been "rediscovered" and reworked more recently (Stephenson 2000). The first index to be derived from Pierce's line of reasoning is "Hanssen-Kuipers discriminant" or "Kuipers' performance index" (Hanssen and Kuipers, 1965).

The difference is that the adjustment for chance in Kuiper's index is based on historical climatological relative frequencies, whereas that in Pierce's measure is based on sample relative frequencies (Murphy, 1996). The second index is the "true skill statistic" (TSS) (Flueck, 1987). Some people also referred to it as "Pierce skill score" in reference to its original discovery (Stephenson, 2000). It has been introduced to SDM by Allouche et al. (2006).

TSS is equivalent to Youden's index J, which was developed by Youden (1951) and is widely used in medical diagnostic tests. It is defined as the average of the net prediction success rate for present sites and that for absent sites. It has gained considerable theoretical interest over many years (Böhning et al., 2008), and it is the best available summary measure of model performance in medical diagnostic tests (Biggerstaff, 2000). This index is closely related to the arithmetic mean of sensitivity and specificity (see Table 2).

The normalized mutual information (NMI) was introduced to ecology by Fielding and Bell (1997), and used in SDM by Manel et al. (2001). NMI is undefined whenever there is 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 the worse-than-random models from the better-than-random models, and as a result is not a useful accuracy measure.

3. THRESHOLD-INDEPENDENT INDICES

The threshold-independent accuracy measures are shown in Table 3. Area under the curve (AUC) of receiver operating characteristic is one of the most widely used accuracy measures in various disciplines including ecology though it has received some criticism (Lobo et al., 2008). In the context of SDM, the AUC of a model is equivalent to the probability that the model will rank a randomly chosen species 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). 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. It is a correlation coefficient rather than AUC (Hand and Till, 2001; Kraemer, 2006). The Gini coefficient was 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 is of no practical use (e.g. Briggs and Zaretzki, 2008). Therefore, partial AUC (i.e. PAUC) was proposed (McClish, 1989), which is the average sensitivity over a fixed range of the false positive rate. The choice of such "regions" has to be made on a case-by-case basis, and the PAUC does not possess a probabilistic interpretation (Lee and Hsiao, 1996).

The maximum overall accuracy and maximum kappa are frequently used in SDM in a threshold-independent way to indicate a model's predictive capacity (e.g. Guisan et al., 1998; Liu et al., 2005). Point biserial correlation coefficient (r_{pb}) is also used in SDM (e.g. Elith et al., 2006). It is the Pearson product moment correlation coefficient calculated under the condition that one variable (i.e. the observed species occurrence) is binary and the other (i.e. the predicted probability) is ordinal (Karemer, 2006). 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 was used in subsequent studies (e.g. Engler et al., 2004). The coefficient of determination R^2 was also suggested for generalized regression model assessment (Ash and Shwartz, 1999). Mean absolute prediction error (MAPE) (Schemper 2003) and mean cross entropy (MXE) (Caruana and Niculescu-Mizil 2004) have also been used as accuracy measures, but not been used in SDM.

4. DISCUSSION AND CONCLUSION

Through this review of the use of accuracy indices in SDM, the following issues have been identified:

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

Index	Definition	Reference
Maximum overall accuracy	$MXOA = \max(OA)$	Liu et al. (2005)
Maximum kappa	$MXKp = \max(Kp)$	Guisan et al (1998)
Area under ROC curve	$AUC = \frac{1}{n_1 n_0}\sum_{i=1}^{n_1}\sum_{j=1}^{n_0} I(p_{1i}, p_{0j})$ where, $I(p_{1i}, p_{0j}) = \begin{cases} 0 & if \quad p_{1i} < p_{0j} \\ 0.5 & if \quad p_{1i} = p_{0j} \\ 1 & if \quad p_{1i} > p_{0j} \end{cases}$ p_{0i} and p_{1j} are the predicted value for the absence site i and presence site j. n_1 and n_0 are the number of present and absent sites respectively.	Mason and Graham (2002)
Gini index	$Gini = 2AUC - 1$	Hand and Till (2001)
Point biserial correlation coefficient	$r_{pb} = \frac{\sum_{o_i=1} p_i - \frac{n_1}{n}\sum_{i=1}^{n} p_i}{\sqrt{n_1(1-\frac{n_1}{n})(\sum_{i=1}^{n} p_i^2 - \frac{1}{n}\sum_{i=1}^{n} p_i)}}$ where, o_i is the observed value for site i (1 for presence, 0 for absence), p_i is the predicted value for site i.	Karemer (2006)
Proportion of Explained deviance	$D^2 = 1 - \log L(\hat{\beta}) / \log L(\hat{\beta}_0)$ where, $L(\hat{\beta}) = \sum_{i=1}^{n}[o_i \log p_i + (1-o_i)\log(1-p_i)]$ $L(\hat{\beta}_0) = p\log p + (1-p)\log(1-p)$, $p = \frac{1}{n}\sum_{i=1}^{n} o_i$	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 (1951)
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	$R^2 = 1 - \sum_{i=1}^{n}(p_i - o_i)^2 \Big/ \sum_{i=1}^{n}(p - o_i)^2$	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)

(1) Almost all SDM studies only consider the discrimination capacity, reliability is rarely evaluated. There are some useful indices that measure model reliability, e.g. Brier's score (i.e. mean square error) or its square root (i.e. root mean square error), which have been used in other fields including meteorology (e.g. Brier 1950) and machine learning (e.g. Caruana and Niculescu-Mizil, 2004). We recommend these measures also be included in SDM accuracy assessment.

(2) The precision of the estimated accuracy is also important information for model accuracy assessment. Though statistical properties for many measures are known (see Couto 2003, Obuchowski 2005, Allouche et al. 2006, and the other references cited) and can be used to calculate the variance or standard deviation or confidence interval for the calculated accuracy, such reporting is unusual in modeling species distributions. We recommend that this information be given for each accuracy measure used. If the theoretical statistical characteristics are not known for the measures used, resampling methods (e.g. bootstrap) can be used to calculate the variance and confidence interval for the estimated accuracy.

(3) The sample size for the test data needed to obtain a reliable estimate of model performance also needs to be considered, as this is closely related to the statistical properties of accuracy measures. Small-sized test datasets may lead to unstable accuracy measurements, which may result in misleading conclusions on model accuracy.

(4) The prevalence dependence of accuracy measures also requires further attention. Existing studies have lead to inconsistent conclusions. The main reason for this is that the design of these studies does not differentiate between the effect of model-building data prevalence and the effect of test data prevalence *per se*.

REFERENCES

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

Ash, A., and M. Shwartz, (1999), R^2: a useful measure of model performance when predicting a dichotomous outcome. *Statistics in Medicine*, 18, 375-384.

Biggerstaff, B.J. (2000), Comparing diagnostic tests: a simple graphic using likelihood ratios. *Statistics in Medicine*, 19, 649-663.

Böhning, B., W. Böhning, and H. Holling, (2008), Revisiting youden's index as a useful measure of the misclassification error in meta-analysis of diagnostic studies. *Statistical Methods in Medical research*, doi:10.1177/0962280207081867.

Breiman, L., J. Friedman, R. Olshen, and C. Stone, (1984), *Classification and regression trees*. Belmont, CA: Wadsworth International Group.

Brier, G.W. (1950), Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78**,** 1-3.

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

Caruana, R., and A. Niculescu-Mizil, (2004), Data mining in metric space: an empirical analysis of supervised learning performance criteria. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, August 22–25, 2004, Seattle, Washington, USA. pp. 69-78.

Cohen, J. (1960), A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-40.

Coughlin, S.S., and L.W. Pickle, (1992), Sensitivity and specificity-like measures of the validity of a diagnostic test that are corrected for chance agreement. *Epidemiology* 3, 178-181.

Couto, P. (2003), Assessing the accuracy of spatial simulation models. *Ecological Modelling*, 167, 181-198.

Daskalaki, S., I. Kopanas, and N. Avouris, (2006), Evaluation of classifiers for an uneven class distribution problem. *Applied Artificial Intelligence*, 20, 381-417.

Doolittle, M.H. (1888), Association ratios. *Bull. Philos. Soc. Washington*, 7, 122-127.

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

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

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

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

Fielding, A.H., and J.F. Bell, (1997), A review of methods for the measurement of prediction errors in conservation presence/absence models. *Environmental Conservation* 24, 38-49.

Finley, J.P. (1884), Tornado predictions. *American Meteorological Journal*, 1, 85-88.

Finn, J.T. (1993), Use of the average mutual information index in evaluating classification error and consistency. *International Journal of Information Systems*, 7, 349-366.

Flueck, J.A. (1987), A study of some measures of forecast verification. Reprints, *10th Conference on Probability and Statistics in Atomaspheric Sciences*. Edmonton, AB, Canada, American Meteorological Society, pp. 69-73.

Gilbert, G.K. (1884), Finley's tornado predictions. *American Meteorological Journal*, 1, 166-172.

Glas, A.S., J.G. Lijmer, M.H. Prins, G.J. Bonsel, and P.M. Bossuyt, (2003), The diagnostic odds ratio: A single indicator of test performance. *Journal of Clinical Epidemiology*, 56, 1129–1135.

Guisan, A., J.P. Theurillat, F. Kienast, (1998), Predicting the potential distribution of plant species in an Alpine environment. *Journal of Vegetation Science*, 9, 65–74.

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

Hand, D.J. (2001), Measuring diagnostic accuracy of statistical prediction rules. *Statistica Neerlandica*, 55, 3-16.

Hand, D.J., and R.J. Hill, (2001), A simple generalization of the area under the ROC curve to multiple class classification problems. *Machine Learning*, 45, 171-186.

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

Hanssen A.J., and W.J. Kuipers, (1965), On the relationship between the frequency of rain and various meteorological parameters. *Meded Verhand*, 81, 2-15.

Kraemer, H.C. (2006), Correlation coefficients in medical research: from product moment correlation to the odds ratio. *Statistical Methods in Medical Research*, 15, 525-545.

Lee, W.C., and C.K. Hsiao, (1996), Alternative Summary Indices for the Receiver Operating Characteristic (ROC) Curve. *Epidemiology*, 7, 605-611.

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

Liu, C., P. Frazier, and K. Kumar, (2007), Comparative assessment of the measures of thematic classification accuracy. *Remote Sensing of Environment*, 107, 606-616.

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

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

McClish, D.K. (1989), Analyzing a portion of the ROC curve. *Medical Decision Making*, 9, 190-195.

McPherson, J.M., W. Jetz, and D.J. Rogers, (2004), The effect of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology*, 41, 811-823.

Mittlböck, M. and M. Schemper, (1996), Explained variation for logistic regression. *Statistics in Medicine*, 15, 1987-1997.

Murphy, A.H. (1996), The Finley affair: a signal event in the history of forecast verification. *Weather and Forecasting*, 11, 3-20.

Obuchowski, N.A. (2005), Fundamentals of clinical research for radiologists: ROC analysis. *AJR*, 184, 364-372.

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

Peirce, C.S. (1884), The numerical measure of the success of predictions. *Science*, 4, 453-454.

Schemper, M. (2003), Predictive accuracy and explained variation. *Statistics in Medicine*, 22, 2299–2308.

Sokolova, M., N. Japkowicz, and S. Szpakowicz, (2006), Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. *Lecture Notes in Computer Science*, 4304, 1015-1021.

Stephenson, D.B. (2000), Use of the "Odds Ratio" for diagnosing forecast skill. *Weather and Forecasting*, 15, 221-232.

Weisberg, S. (1980), *Applied linear regression*. Wiley, New York, pp. 283.

Youden, W.J. (1950), Index for rating diagnostic tests. *Cancer*, 3, 32-35.