Journal of Applied Ecology 2006
43, 1223–1232

METHODOLOGICAL INSIGHTS

Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS)

OMRI ALLOUCHE, ASAF TSOAR and RONEN KADMON

Department of Evolution, Systematics and Ecology, Institute of Life Sciences, The Hebrew University, Givat-Ram, Jerusalem 91904, Israel

Summary

1. In recent years the use of species distribution models by ecologists and conservation managers has increased considerably, along with an awareness of the need to provide accuracy assessment for predictions of such models. The kappa statistic is the most widely used measure for the performance of models generating presence–absence predictions, but several studies have criticized it for being inherently dependent on prevalence, and argued that this dependency introduces statistical artefacts to estimates of predictive accuracy. This criticism has been supported recently by computer simulations showing that kappa responds to the prevalence of the modelled species in a unimodal fashion.

2. In this paper we provide a theoretical explanation for the observed dependence of kappa on prevalence, and introduce into ecology an alternative measure of accuracy, the true skill statistic (TSS), which corrects for this dependence while still keeping all the advantages of kappa. We also compare the responses of kappa and TSS to prevalence using empirical data, by modelling distribution patterns of 128 species of woody plant in Israel.

3. The theoretical analysis shows that kappa responds in a unimodal fashion to variation in prevalence and that the level of prevalence that maximizes kappa depends on the ratio between sensitivity (the proportion of correctly predicted presences) and specificity (the proportion of correctly predicted absences). In contrast, TSS is independent of prevalence.

4. When the two measures of accuracy were compared using empirical data, kappa showed a unimodal response to prevalence, in agreement with the theoretical analysis. TSS showed a decreasing linear response to prevalence, a result we interpret as reflecting true ecological phenomena rather than a statistical artefact. This interpretation is supported by the fact that a similar pattern was found for the area under the ROC curve, a measure known to be independent of prevalence.

5. *Synthesis and applications*. Our results provide theoretical and empirical evidence that kappa, one of the most widely used measures of model performance in ecology, has serious limitations that make it unsuitable for such applications. The alternative we suggest, TSS, compensates for the shortcomings of kappa while keeping all of its advantages. We therefore recommend the TSS as a simple and intuitive measure for the performance of species distribution models when predictions are expressed as presence–absence maps.

Key-words: AUC, Mahalanobis distance, predictive maps, ROC curves, sensitivity, specificity, woody plants

Journal of Applied Ecology (2006) **43**, 1223–1232
doi: 10.1111/j.1365-2664.2006.01214.x

Correspondence: Omri Allouche, Department of Evolution, Systematics and Ecology, Institute of Life Sciences, The Hebrew University, Givat-Ram, Jerusalem 91904, Israel (fax 972 2 6584741; e-mail omri_alush@hotmail.com).

Introduction

Ecologists and conservation managers increasingly rely on predictive models as a means for estimating patterns of species distribution (Loiselle *et al*. 2003; Vaughan & Ormerod 2003; Rushton, Ormerod & Kerby 2004; Sanchez-Cordero *et al*. 2005). Distribution models are used to evaluate the spreading potential of invading species (Peterson & Robins 2003; Rouget *et al*. 2004; Thuiller *et al*. 2005b), identify and manage threatened species (Engler, Guisan & Rechsteiner 2004; Norris 2004), prioritize places for biodiversity conservation (Araujo *et al*. 2004; Ortega-Huerta & Peterson 2004; Sanchez-Cordero *et al*. 2005) and evaluate the potential impact of climate change on patterns of species distribution (Skov & Svenning 2004; Beaumont, Hughes & Poulsen 2005; Bomhard *et al*. 2005; Thuiller *et al*. 2005a; Thuiller, Lavorel & Araújo 2005).

One fundamental issue in the development of distribution models is the assessment of predictive accuracy (Guisan & Thuiller 2005; Barry & Elith 2006). A quantitative assessment of model performance assists in determining the suitability of the model for specific applications and may help to identify those aspects of the model that need improvement (Vaughan & Ormerod 2005; Barry & Elith 2006; Guisan *et al*. 2006). An assessment of model performance can also provide a basis for comparing alternative modelling techniques (Loiselle *et al*. 2003; Segurado & Araujo 2004; Pearson *et al*. 2006) and enables the user to investigate how different properties of the data and/or the species affect the accuracy of predictive maps generated by the model (Kadmon, Farber & Danin 2003; Segurado & Araujo 2004; Reese *et al*. 2005; Seoane *et al*. 2005).

Models generating presence–absence predictions (hereafter presence–absence models) are usually evaluated by comparing the predictions with a set of validation sites and constructing a confusion matrix that records the number of true positive (a), false positive (b), false negative (c) and true negative (d) cases predicted by the model (Table 1). Models generating non-dichotomous scores on an ordinal scale (hereafter ordinal score models) are often evaluated by applying a certain threshold to transform the scores into a dichotomous set of presence–absence predictions, and constructing a corresponding confusion matrix. One simple measure of accuracy that can be derived from the confusion matrix is the proportion of correctly predicted sites (overall accuracy; Table 2). However, this measure was criticized for ascribing high accuracies for rare species (Fielding & Bell 1997; Manel, Dias & Ormerod 1999). Two alternative measures that are often derived from the confusion matrix are sensitivity and specificity. Sensitivity is the proportion of observed presences that are predicted as such, and therefore quantifies omission errors. Specificity is the proportion of observed absences that are predicted as such, and therefore quantifies commission errors (Table 2). Sensitivity and specificity are independent of each other when compared across models, and are also independent of prevalence ($(a + c)/n$, the proportion of sites in which the species was recorded as present; Table 1).

Table 1. An error matrix used to evaluate the predictive accuracy of presence–absence models. *a*, number of cells for which presence was correctly predicted by the model; *b*, number of cells for which the species was not found but the model predicted presence; *c*, number of cells for which the species was found but the model predicted absence; *d*, number of cells for which absence was correctly predicted by the model

		Validation data set	
		Presence	Absence
Model	Presence	*a*	*b*
	Absence	*c*	*d*

Table 2. Measures of predictive accuracy calculated from a 2×2 error matrix (Table 1). Overall accuracy is the rate of correctly classified cells. Sensitivity is the probability that the model will correctly classify a presence. Specificity is the probability that the model will correctly classify an absence. The kappa statistic and TSS normalize the overall accuracy by the accuracy that might have occurred by chance alone. In all formulae $n = a + b + c + d$

Measure	Formula
Overall accuracy	$\frac{a+d}{n}$
Sensitivity	$\frac{a}{a+c}$
Specificity	$\frac{d}{b+d}$
Kappa statistic	$\frac{\left(\frac{a+d}{n}\right) - \frac{(a+b)(a+c)+(c+d)(d+b)}{n^2}}{1 - \frac{(a+b)(a+c)+(c+d)(d+b)}{n^2}}$
TSS	sensitivity + specificity – 1

An alternative method for assessing the accuracy of ordinal score models is the receiver operating characteristic (ROC) curve (Fielding & Bell 1997). ROC curves are constructed by using all possible thresholds to classify the scores into confusion matrices, obtaining sensitivity and specificity for each matrix, and then plotting sensitivity against the corresponding proportion of false positives (equal to 1 – specificity). The use of all possible thresholds avoids the need for a selection of a single threshold, which is often arbitrary (Manel, Dias & Ormerod 1999; Manel, Williams & Ormerod 2001; Liu *et al*. 2005), and allows appreciation of the trade-off between sensitivity and specificity (Pearce & Ferrier 2000). The area under the ROC curve (AUC) is often used as a single threshold-independent measure for model performance (Manel, Williams & Ormerod 2001; Thuiller 2003; Brotons *et al*. 2004; McPherson, Jetz & Rogers 2004; Thuiller, Lavorel & Araújo 2005).

AUC was shown to be independent of prevalence (Manel, Williams & Ormerod 2001; McPherson, Jetz & Rogers 2004) and is considered a highly effective

measure for the performance of ordinal score models. However, practical applications of species distribution models in conservation planning, such as the identification of biodiversity hotspots and the selection of representative conservation sites, often require presence–absence maps of species distribution, and thus a selection of a threshold for transforming ordinal scores into presence–absence predictions (Cumming 2000b; Loiselle *et al*. 2003; Berg, Gardenfors & von Proschwitz 2004). In such cases, predictive accuracy should be evaluated based on the selected threshold rather than on threshold-independent ROC curves. It should also be noted that some of the most frequently used models of species distribution (e.g. BioCLIM, Nix 1986; GARP, Stockwell & Peters 1999) generate dichotomous presence–absence predictions of species distribution, for which ROC curves cannot be applied.

The most popular measure for the accuracy of presence–absence predictions is Cohen's kappa (Shao & Halpin 1995; Manel, Williams & Ormerod 2001; Loiselle *et al*. 2003; Petit *et al*. 2003; Berg, Gardenfors & von Proschwitz 2004; Parra, Graham & Freile 2004; Pearson, Dawson & Liu 2004; Rouget *et al*. 2004; Segurado & Araujo 2004). This measure corrects the overall accuracy of model predictions by the accuracy expected to occur by chance (Table 2). The kappa statistic ranges from –1 to +1, where +1 indicates perfect agreement and values of zero or less indicate a performance no better than random (Cohen 1960; Table 2). Other advantages of kappa are its simplicity, the fact that both commission and omission errors are accounted for in one parameter, and its relative tolerance to zero values in the confusion matrix (Manel, Williams & Ormerod 2001).

In spite of its wide use, several studies have criticized the kappa statistic for being inherently dependent on prevalence and claimed that this dependency introduces bias and statistical artefacts to estimates of accuracy (Cicchetti & Feinstein 1990; Byrt, Bishop & Carlin 1993; Lantz & Nebenzahl 1996). In a recent study focusing on the evaluation of species distribution models, McPherson, Jetz & Rogers (2004) used numerical simulations to analyse the dependency of kappa on prevalence of the modelled species and found that kappa responds to variation in prevalence in a unimodal fashion. Based on this finding they concluded that 'kappa's sensitivity to prevalence overall, however, renders it inappropriate for comparisons of model accuracy between species or regions unless certain precautions are taken' McPherson, Jetz & Rogers (2004).

In this paper we explain the observed unimodal dependency of kappa on prevalence, and introduce into ecology a new measure for the performance of presence–absence distribution models, the true skill statistic (TSS), which corrects for this dependency while still keeping all of the advantages of kappa.

We begin with a theoretical explanation for the unimodal dependence of kappa on prevalence. To do so we reformulate kappa in terms of prevalence, sensitivity and specificity. We then show analytically that TSS is largely immune to prevalence. We also compare the effect of prevalence on kappa and TSS using real data by modelling distribution patterns of 128 species of woody plants in Israel. Finally we discuss some methodological issues of kappa, TSS and other measures of accuracy, and their relevance for the performance of species distribution models.

Theoretical analysis

The mechanism underlying the unimodal dependency of the kappa statistic on prevalence can be understood by reformulating kappa in terms of three parameters: prevalence, sensitivity and specificity. Such derivation leads to the following form:

$$\text{Kappa} = \frac{P_o - P_e}{1 - P_e},\ P_o = P \cdot Sn + (1 - P) \cdot Sp,\quad P_e = -2(Sn + Sp - 1)P(1 - P) + P_o \qquad \text{eqn 1}$$

where P, Sn and Sp are prevalence, sensitivity and specificity, respectively, P_o is the observed accuracy and P_e is the accuracy expected to occur by chance. Kappa has an extremum at P that satisfies both $(Sn - Sp)P^2 - 2(1 - Sp)P + (1 - Sp) = 0$ and $0 \le P \le 1$. The extremum is a maximum when $Sn + Sp - 1 > 0$ and a minimum when $Sn + Sp - 1 < 0$. We will focus on the former case, which characterizes models with performance better than random. The prevalence that maximizes the kappa score of a given model is thus a function of the sensitivity and specificity of the model. If sensitivity and specificity are equal, a maximal kappa score is obtained for equal proportions of presences and absences. If sensitivity is larger than specificity, kappa is maximized by higher prevalence rates. If specificity is larger than sensitivity, kappa is maximized by lower prevalence rates. In any case, kappa inherently depends on prevalence. An alternative measure is thus required for assessing the performance of presence–absence models, which is largely insensitive to prevalence.

It is rewarding first to define theoretically when two modelling methods are of equal performance. It seems reasonable to assume that two methods are equal in their overall performance if they are equal in both sensitivity and specificity, and hence are equal in their ability to detect presences and absences. It also seems reasonable to expect that properties of the specific data set for which alternative methods are applied should not affect their rating. Taking into account the fact that the confusion matrix can be fully described by sensitivity, specificity, prevalence and the size of the validation set, an ideal measure of model performance should not be affected by prevalence or the size of the specific data set used for model validation (both being properties of the specific data set) and it should combine sensitivity and specificity so that both omission and commission errors are accounted for. We propose the true skill statistic (TSS), also known as the Hanssen–Kuipers discriminant, as a measure that satisfies these requirements.

This statistic, traditionally used for assessing the accuracy of weather forecasts (McBride & Ebert 2000; Saseendran *et al*. 2002; Elmore, Weiss & Banacos 2003; Accadia *et al*. 2005), compares the number of correct forecasts, minus those attributable to random guessing, to that of a hypothetical set of perfect forecasts (see Appendix S1 in the supplementary material). For a 2×2 confusion matrix TSS is defined as:

$$\mathrm{TSS} = \frac{ad - bc}{(a + c)(b + d)} = \text{Sensitivity} + \text{Specificity} - 1 \qquad \text{eqn 2}$$

Like kappa, TSS takes into account both omission and commission errors, and success as a result of random guessing, and ranges from –1 to +1, where +1 indicates perfect agreement and values of zero or less indicate a performance no better than random. However, in contrast to kappa, TSS is not affected by prevalence. It can also be seen that TSS is not affected by the size of the validation set, and that two methods of equal performance have equal TSS scores. In Appendix S1 in the supplementary material we describe in more detail the relation of TSS to kappa. TSS is a special case of kappa, given that the proportions of presences and absences in the validation set are equal.

Computer simulations were conducted to allow more thorough comparison of kappa and TSS and their responses to prevalence. Confusion matrices consisting of 100 cases each were created by tagging presences and absences to be correctly classified at probabilities equal to predetermined values of sensitivity and specificity, respectively. Three possible scenarios were simulated: (i) equal sensitivity and specificity (both set to 0·8); (ii) higher sensitivity (sensitivity = 0·9, specificity = 0·7); and (iii) higher specificity (sensitivity = 0·7, specificity = 0·9). The number of presence cases was varied systematically from 1 to 99 in increments of 1. For each level of prevalence we randomly simulated 100 000 confusion matrices. The kappa and TSS scores were determined for each of the 9 900 000 matrices and their mean values were calculated for each level of prevalence under the three scenarios. The corresponding theoretical expectations were also calculated for each value of prevalence based on equations 1 and 2. The results (Fig. 1) showed that TSS scores were largely unaffected by prevalence while kappa scores exhibited a unimodal response to prevalence, as found by McPherson, Jetz & Rogers (2004). We conclude that, in contrast to kappa, documented effects of prevalence on TSS can be interpreted as evidence for real ecological phenomenon rather than statistical artefacts.

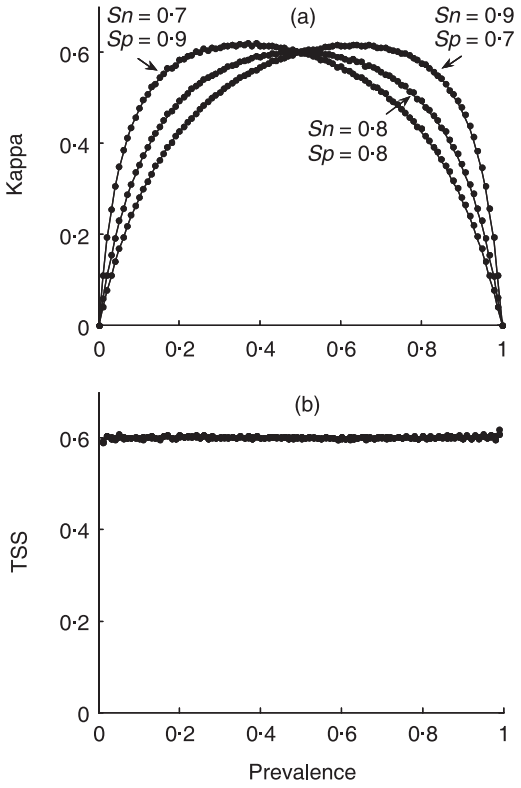

Fig. 1. Effect of prevalence on kappa and TSS based on theoretical calculations (continuous lines) and corresponding numerical simulations (dots indicating mean values of 100 000 randomly simulated confusion matrices for each level of prevalence). Three sets of parameters were used: (i) $Sn = 0{\cdot}8$, $Sp = 0{\cdot}8$; (ii) $Sn = 0{\cdot}9$, $Sp = 0{\cdot}7$; (iii) $Sn = 0{\cdot}7$, $Sp = 0{\cdot}9$. Differences between the three sets in TSS values are indistinguishable. The total number of cases (n) was set to 100 in all simulations.

Empirical analysis

STUDY SYSTEM

An empirical comparison of the responses of kappa and TSS to variation in prevalence was carried out by re-analysing the data used by Farber & Kadmon (2003) for introducing the Mahalanobis distance as an approach for species distribution modelling. This data set comprises 32 414 geo-referenced observations on the distribution of 128 woody species in Israel (median number of observations per species 159). The models developed by Farber & Kadmon (2003) were validated using an independent database consisting of lists of species recorded in 96 validation sites of 5×5 km covering the main climatic gradients of Israel. The same calibration and validation data sets were used here to compare the responses of kappa and TSS to prevalence.

METHODS

As in the theoretical analysis, prevalence was defined as the proportion of validation sites in which the relevant species was recorded. Predictive presence–absence maps were produced using the Mahalanobis distance. Three climatic factors were used as predictors in the models: mean annual rainfall, mean daily temperature of the hottest month (August) and mean minimum temperature of the coldest month (January). Further

details of the modelling approach and the data can be found in Farber & Kadmon (2003).

We quantified the accuracy of the predictive map produced for each of the 128 species using four measures of accuracy: kappa, TSS, sensitivity and specificity. We also calculated the AUC statistic for each species non-parametrically using the Wilcoxon statistic (Hanley & McNeil 1982). Each of these five measures was regressed against prevalence using two types of models: a linear model and a quadratic model.

RESULTS

When kappa was regressed against prevalence with a linear model, prevalence had a positive but very weak effect on kappa ($P = 0{\cdot}047$). The portion of variance explained by this model was extremely low (0·02). When the same data were analysed using a quadratic model the portion of variance explained increased to 0·12 and the coefficient of the quadratic term was negative and highly significant (< 0·001), as expected from the theoretical analysis.

The linear models constructed for AUC and TSS showed that both measures were negatively and significantly correlated with prevalence (Table 3). This response suggested that distribution ranges of rare species were more predictable than those of more common species. When AUC and TSS values were analysed by quadratic models, the coefficients of the quadratic term were not statistically significant (Table 3).

The effect of prevalence on sensitivity was not statistically significant for both the linear and quadratic models but the corresponding effect on specificity was negative and highly significant (Table 3). These results indicated that the decrease of TSS with increasing prevalence was caused by an increase in the magnitude of commission errors. As can be expected from these results, when the five measures of accuracy were plotted against prevalence, kappa showed a unimodal response, TSS, AUC and specificity showed a negative response, and sensitivity showed no response (Fig. 2).

Spearman correlation analysis indicated that all pair-wise correlations between AUC, TSS and kappa were statistically significant ($P < 0{\cdot}01$). However, the correlation between AUC and TSS was higher than the correlation of AUC with kappa or the correlation between TSS and kappa (0·85 vs. 0·65 and 0·66, respectively).

Table 3. Results of linear regression models ($y = b_0 + b_1x$) and quadratic regression models ($y = b_0 + b_1x + b_2x^2$) for the effect of prevalence of woody plant species on five measures of accuracy (kappa, TSS, AUC, sensitivity and specificity). Asterisks indicate significance levels of regression coefficients; $*P < 0{\cdot}05$, $***P < 0{\cdot}001$

Measure	Regression model	b_1	b_2	Adjusted R^2
kappa	Linear	0·25*		0·02
	Quadratic	1·54***	–2·33***	0·12
TSS	Linear	–0·52***		0·13
	Quadratic	–0·31***	–0·38	0·13
AUC	Linear	–0·25***		0·14
	Quadratic	–0·192	–0·10	0·13
Sensitivity	Linear	–0·11		< 0·01
	Quadratic	–0·25	0·26	< 0·01
Specificity	Linear	–0·41***		0·14
	Quadratic	–0·06	–0·64	0·15

Discussion

McPherson, Jetz & Rogers (2004) demonstrated with numerical simulations that kappa, one of the most common measures of predictive accuracy in ecology, is inherently sensitive to prevalence, showing a unimodal dependency with a maximum at intermediate levels of prevalence. This bias has long been recognized in other research fields, such as clinical epidemiology (Cicchetti & Feinstein 1990; Byrt, Bishop & Carlin 1993; Lantz & Nebenzahl 1996). In this paper we provide an analytical explanation for the results obtained by McPherson, Jetz & Rogers (2004), and propose a new measure of accuracy, TSS, that is insensitive to prevalence while still keeping all the advantages of the kappa statistic. We also provide empirical data supporting the hypothesis that the two measures of accuracy respond differentially to variation in prevalence, and demonstrate that the relationship between kappa and prevalence is unimodal, as expected from the theoretical analysis. Several previous studies have documented unimodal responses of kappa to species' prevalence (Manel, Williams & Ormerod 2001; Petit *et al*. 2003; Liu *et al*. 2005) but none of these studies attributed this response to statistical artefact.

In our empirical analysis, TSS showed a negative response to prevalence, a result we interpret as indicative of a true effect of prevalence (or ecological characteristics associated with prevalence) on predictive accuracy. The fact that AUC, which is known to be independent of prevalence, showed a similar response to prevalence supports this interpretation. We explain this result by the fact that prevalent species often occupy wide niches. The area of predicted presence for such species is therefore much larger than that of scarce species. The increased area allows the Mahalanobis distance method to keep high levels of sensitivity (correctly predicting a presence as such) but results in a decrease in specificity, as many locations where the species is absent are erroneously predicted as presence locations.

Evidence for negative effects of prevalence on the accuracy of species distribution models was found in several previous studies. For example, Guisan & Hofer (2003) analysed the distribution of reptiles in Switzerland using generalized linear models (GLM) and found that highly common species showed exceptionally low values of predictive accuracy. Segurado & Araujo (2004) compared the performance of seven modelling techniques in predicting the distribution of amphibians

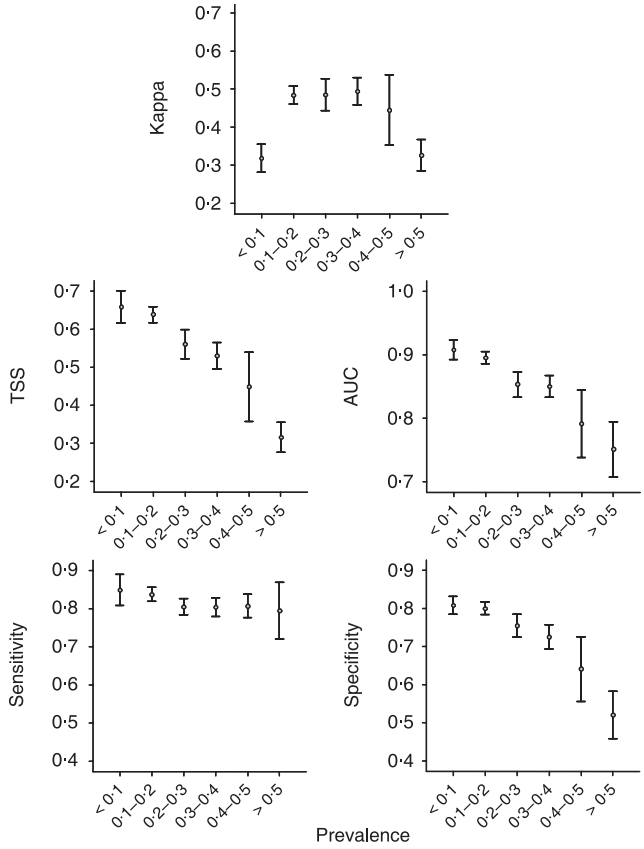

Fig. 2. Effect of prevalence on kappa, TSS, AUC, sensitivity and specificity of predictive maps produced for 128 species of woody plants. Species were categorized into six prevalence categories and the mean (± 1 SE) score of each measure was calculated for each prevalence category.

and reptiles in Portugal and found that, in general, widespread species had greater overall errors. Stockwell & Peterson (2002) analysed patterns of bird distribution in Mexico with GARP and found that range size had a negative effect on the accuracy of model predictions. They suggested that widespread species show local adaptations in ecological characteristics and that ignoring such ecological differentiation overestimates the species' distribution range and reduces model accuracy.

As can be verified by assigning $P = 0{\cdot}5$ in equation 1, when the proportions of presences and absences are equal, kappa is equal to TSS. The bias caused to kappa by unequal proportions of presences and absences in the validation set has led some to suggest that efforts should be made to collect validation sets such that prevalence would be around 50% (Lantz & Nebenzahl 1996; Hoehler 2000; McPherson, Jetz & Rogers 2004). Unfortunately this recommendation is of questionable practicability in ecological applications, particularly for rare species for which a small number of presence records is available. TSS satisfies this recommendation without requiring special adjustments or sampling efforts.

An alternative approach to obtaining a validation set with effective prevalence of 50% is by random resampling from the data available (Stockwell & Peterson 2002). This method suffers from stochasticity in the selection of points, but if the validation set is large enough, or if performance is averaged over several randomly drawn validation sets, the results would converge to the TSS statistic.

The dependence of kappa on prevalence has led to the development of several modifications of kappa that attempt to 'adjust' for this dependency (e.g. the prevalence and bias adjusted kappa, PABAK, proposed by Byrt, Bishop & Carlin 1993; the adjusted kappa coefficient, κ' described by Hoehler 2000). However, PABAK does not consider agreement as a result of chance and therefore assigns high scores to algorithms of poor predictive ability (Henderson 1993; Fielding & Bell 1997; Manel, Williams & Ormerod 2001; Olden, Jackson & Peres-Neto 2002). The adjusted kappa coefficient was also criticized for being sensitive to variation in prevalence (Hoehler 2000).

AUC is commonly used as a measure of model performance (Manel, Williams & Ormerod 2001; Thuiller 2003; Brotons *et al*. 2004; McPherson, Jetz & Rogers 2004; Thuiller, Lavorel & Araújo 2005) and has been shown to be independent of prevalence, both theoretically (Hanley & McNeil 1982; Zweig & Campbell 1993) and empirically (Manel, Williams & Ormerod 2001; McPherson, Jetz & Rogers 2004). AUC is a threshold-independent measure of model performance and is therefore particularly suitable for evaluating the performance of ordinal score models such as logistic

regression. Yet, for practical applications, a dichotomous prediction of presence–absence is often required, and hence a threshold must be applied to transform the probability/suitability scores to presence–absence data. For example, most reserve selection algorithms require presence–absence data on species composition in the relevant area (Church, Stoms & Davis 1996; Howard *et al*. 1998; Margules & Pressey 2000; Tsuji & Tsubaki 2004). As available data are often partial, species distribution models are often used to predict the presence or absence of species in candidate areas (Loiselle *et al*. 2003; Ortega-Huerta & Peterson 2004; Sanchez-Cordero *et al*. 2005). Estimates of biodiversity hotspots are also often based on presence–absence predictions (Cumming 2000b; Schmidt *et al*. 2005). ROC plots cannot be constructed for presence–absence predictions and, therefore, AUC is not applicable for evaluating the accuracy of predictive maps used in such applications.

TSS provides a threshold-dependent measure of accuracy that is readily applied for presence–absence predictions. Our theoretical analysis demonstrates that it is not affected by prevalence, and our empirical analysis indicates that its values are highly correlated with those of the threshold-independent AUC statistic. These findings suggest that TSS can serve as an appropriate alternative to AUC in cases where model predictions are formulated as presence–absence maps. Several recent studies have jointly used AUC as a threshold-independent and kappa as a threshold-dependent measure of predictive accuracy (Thuiller 2003; Huntley *et al*. 2004; Pearson, Dawson & Liu 2004; Araujo *et al*. 2005; Pearson *et al*. 2006). Our results suggest that TSS should be preferred over kappa as a threshold-dependent measure in such studies.

As is evident from equation 2, TSS assigns equal weights to sensitivity and specificity. Practical applications might require different weights. In conservation planning, for example, when predicting distribution of endangered species, one may wish to weight sensitivity more than specificity. Unlike the kappa score, weights can easily be introduced to the TSS in a straightforward manner.

While our theoretical and empirical results support the superiority of TSS over the kappa statistics, a thorough comparison of the two measures should also consider their variance and its dependency on prevalence. It can be shown that the variance of TSS is given by:

$$V(\mathrm{TSS}) = \frac{Sn(1-Sn)}{N \cdot P} + \frac{Sp(1-Sp)}{N(1-P)} \qquad \text{eqn 3}$$

where Sn, Sp, P and N are the sensitivity, specificity, prevalence and size of the validation set, respectively. As evident from Fig. 3a, TSS is highly variable for extremely low and high levels of prevalence. The variability is caused by the large variability in sensitivity for small data sets with very low prevalence range (as randomness can easily change the parameter from 1 to 0 or vice versa) and in specificity for small data sets with very high prevalence range. The variance of kappa can be calculated non-parametrically for a finite N, by determining the probability of obtaining each possible confusion matrix and the corresponding kappa score. Given a prevalence P, sensitivity Sn, specificity Sp and N cases, the probability of obtaining a confusion matrix with values a, b, c and d (as defined in Table 1) is given by:

$$f(a, b, c, d \mid N, P, Sn, Sp) = \binom{NP}{a} Sn^a (1-Sn)^b \binom{N(1-P)}{d} Sp^d (1-Sp)^c. \qquad \text{eqn 4}$$

A plot of the variance of kappa against prevalence (Fig. 3a) indicates that kappa does not suffer from the border effects obtained for TSS, and that its variability decreases for extremely low or high prevalence. However, the absolute value of kappa also decreases for low or high prevalence (Fig. 1a). A more informative comparison of the two measures should therefore be based on their coefficient of variation (CV; the standard deviation divided by the mean). Such a comparison reveals similar curves for the two measures (Fig. 3b). Similar patterns have been shown to characterize AUC as well (Cumming 2000a; McPherson, Jetz & Rogers 2004). Instability at extremely low or high levels of prevalence seems to be inherent in any model with low number of cases in one of the cells of the confusion matrix (Nelson & Cicchetti 1995).

Finally, when discussing the suitability of TSS vs. kappa as measures for model performance, one should

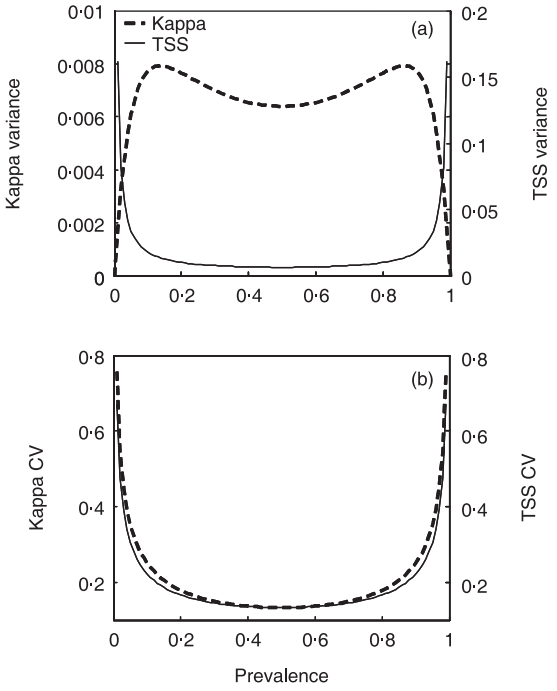

Fig. 3. Effect of prevalence on the variance (a) and coefficient of variation (b) of kappa (dashed lines) and TSS (continuous lines). Results are shown for n = 100 and equal scores of sensitivity and specificity (both 0·8).

distinguish between tests of agreement and validation. Kappa was originally designed to measure reliability of predictions by assessing agreement between two or more observers (Tooth & Ottenbacher 2004). In such applications none of the observers is treated as a 'gold standard', i.e. is known to be accurate. For such a purpose the prevalence effect on kappa is much desired and kappa should not be adjusted for it (Hoehler 2000). However, in validation tests such as those performed for evaluating the performance of distribution models, a gold standard obviously exists. Under such circumstances the prevalence effect of kappa turns against it, and sensitivity and specificity, which are not applicable for the purpose of assessing agreement between two observers, become very informative. TSS accounts for both sensitivity and specificity and is therefore better suited than kappa for measuring performance of a method in the presence of a gold standard.

CONCLUSIONS

In a recent review of the challenge of testing models of species distribution, Vaughan & Ormerod (2005) concluded that adequate testing of such models is still scarce and that their true value cannot yet be appraised. The results of this study support their conclusions and provide theoretical and empirical support that kappa, one of the most widely used measures of model performance in ecology, has serious limitations that make it unsuitable for such applications. The alternative we suggest, the TSS statistic, compensates for the shortcomings of kappa while keeping all of its advantages, and provides results that are highly correlated with those of the threshold-independent AUC statistic. We therefore recommend the TSS as a simple and intuitive measure for the performance of predictive maps generated by presence–absence models.

Acknowledgements

We thank A. Ben-Nun for assistance with GIS issues, U. Motro for fruitful discussions of statistical issues, and W. Thuiller and an anonymous referee for valuable comments on an earlier version of the manuscript. The research was supported by The Israel Science Foundation (grant no. 545/03) and the Ministry of Environment.

References

Accadia, C., Mariani, S., Casaioli, M., Lavaqnini, A. & Speranza, A. (2005) Verification of precipitation forecasts from two limited-area models over Italy and comparison with ECMWF forecasts using a resampling technique. *Weather and Forecasting*, **20**, 276–300.

Araujo, M.B., Cabeza, M., Thuiller, W., Hannah, L. & Williams, P.H. (2004) Would climate change drive species out of reserves? An assessment of existing reserve-selection methods. *Global Change Biology*, **10**, 1618–1626.

Araujo, M.B., Pearson, R.G., Thuiller, W. & Erhard, M. (2005) Validation of species–climate impact models under climate change. *Global Change Biology*, **11**, 1504–1513.

Barry, S. & Elith, J. (2006) Error and uncertainty in habitat models. *Journal of Applied Ecology*, **43**, 413–423.

Beaumont, L.J., Hughes, L. & Poulsen, M. (2005) Predicting species' distributions: use of climatic parameters in BIOCLIM and its impact on predictions of species' current and future distributions. *Ecological Modelling*, **186**, 250–269.

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

Bomhard, B., Richardson, D.M., Donaldson, J.S., Hughes, G.O., Midgley, G.F., Raimondo, D.C., Rebelo, A.G., Rouget, M. & Thuiller, W. (2005) Potential impacts of future land use and climate change on the Red List status of the Proteaceae in the Cape Floristic Region, South Africa. *Global Change Biology*, **11**, 1452–1468.

Brotons, L., Thuiller, W., Araújo, M.B. & Hirzel, A.H. (2004) Presence–absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, **27**, 437–448.

Byrt, T., Bishop, J. & Carlin, J.B. (1993) Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, **46**, 423–429.

Church, R.L., Stoms, D.M. & Davis, F.W. (1996) Reserve selection as a maximal covering location problem. *Biological Conservation*, **76**, 105–112.

Cicchetti, D.V. & Feinstein, A.R. (1990) High agreement but low kappa. II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, **43**, 551–558.

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

Cumming, G.S. (2000a) Using between-model comparisons to fine-tune linear models of species ranges. *Journal of Biogeography*, **27**, 441–455.

Cumming, G.S. (2000b) Using habitat models to map diversity: pan-African species richness of ticks (Acari: Ixodida). *Journal of Biogeography*, **27**, 425–440.

Doswell, C.A., Daviesjones, R. & Keller, D.L. (1990) On summary measures of skill in rare event forecasting based on contingency tables. *Weather and Forecasting*, **5**, 576–585.

Elmore, K.L., Weiss, S.J. & Banacos, P.C. (2003) Operational ensemble cloud model forecasts. Some preliminary results. *Weather and Forecasting*, **18**, 953–964.

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

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

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

Guisan, A. & Hofer, U. (2003) Predicting reptile distributions at the mesoscale: relation to climate and topography. *Journal of Biogeography*, **30**, 1233–1243.

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

Guisan, A., Lehmann, A., Ferrier, S., Austin, M., Overton, J.M.C.C., Aspinall, R. & Hastie, T. (2006) Making better biogeographical predictions of species' distributions. *Journal of Applied Ecology*, **43**, 386–392.

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

Henderson, A.R. (1993) Assessing test accuracy and its clinical consequences: a primer for receiver operating characteristic curve analysis. *Annals of Clinical Biochemistry*, **30**, 521–539.

Hoehler, F.K. (2000) Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *Journal of Clinical Epidemiology*, **53**, 499–503.

Howard, P.C., Viskanic, P., Davenport, T.R.B., Kigenyi, F.W., Baltzer, M., Dickinson, C.J., Lwanga, J.S., Matthews, R.A. & Balmford, A. (1998) Complementarity and the use of indicator groups for reserve selection in Uganda. *Nature*, **394**, 472–475.

Huntley, B., Green, R.E., Collingham, Y.C., Hill, J.K., Willis, S.G., Bartlein, P.J., Cramer, W., Hagemeijer, W.J.M. & Thomas, C.J. (2004) The performance of models relating species' geographical distributions to climate is independent of trophic level. *Ecology Letters*, **7**, 417–426.

Kadmon, R., Farber, O. & Danin, A. (2003) A systematic analysis of factors affecting the performance of climatic envelope models. *Ecological Applications*, **13**, 853–867.

Lantz, C.A. & Nebenzahl, E. (1996) Behavior and interpretation of the *k* statistic: resolution of two paradoxes. *Journal of Clinical Epidemiology*, **49**, 431–434.

Liu, C., Berry, P.M., Dawson, T.P. & Pearson, R.G. (2005) Selecting thresholds of occurrence in the prediction of species' distributions. *Ecography*, **28**, 385–393.

Loiselle, B.A., Howell, C.A., Graham, C.H., Goerck, J.M., Brooks, T., Smith, K.G. & Williams, P.H. (2003) Avoiding pitfalls of using species distribution models in conservation planning. *Conservation Biology*, **17**, 1591–1600.

McBride, J.L. & Ebert, E.E. (2000) Verification of quantitative precipitation forecasts from operational numerical weather prediction models over Australia. *Weather and Forecasting*, **15**, 103–121.

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

Manel, S., Dias, J.M. & Ormerod, S.J. (1999) Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecological Modelling*, **120**, 337–347.

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

Margules, C.R. & Pressey, R.L. (2000) Systematic conservation planning. *Nature*, **405**, 243–253.

Nelson, L.D. & Cicchetti, D.V. (1995) Assessment of emotional functioning in brain-impaired individuals. *Psychological Assessment*, **7**, 404–413.

Nix, H.A. (1986) A biogeographic analysis of Australian elapid snakes. *Atlas of Elapid Snakes of Australia* (Ed. R. Longmore), pp. 4–15. Australian Government Publications Service, Canberra, Australia.

Norris, K. (2004) Managing threatened species: the ecological toolbox, evolutionary theory and declining-population paradigm. *Journal of Applied Ecology*, **41**, 413–426.

Olden, J.D., Jackson, D.A. & Peres-Neto, P.R. (2002) Predictive models of fish species distributions: a note on proper validation and chance predictions. *Transactions of the American Fisheries Society*, **131**, 329–336.

Ortega-Huerta, M.A. & Peterson, A.T. (2004) Modelling spatial patterns of biodiversity for conservation prioritization in north-eastern Mexico. *Diversity and Distributions*, **10**, 39–54.

Parra, J.L., Graham, C.C. & Freile, J.F. (2004) Evaluating alternative data sets for ecological niche models of birds in the Andes. *Ecography*, **27**, 350–360.

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

Pearson, R.G., Dawson, T.P. & Liu, C. (2004) Modelling species distributions in Britain: a hierarchical integration of climate and land-cover data. *Ecography*, **27**, 285–298.

Pearson, R.G., Thuiller, W., Araujo, M.B., Martinez-Meyer, E., Brotons, L., McClean, C., Miles, L., Segurado, P., Dawson, T.E. & Lees, D.C. (2006) Model-based uncertainty in species' range prediction. *Journal of Biogeography*, DOI: 10.1111/j.1365_2699.2006.01460x.

Peterson, A.T. & Robins, C.R. (2003) Using ecological-niche modelling to predict barred owl invasions with implications for spotted owl conservation. *Conservation Biology*, **17**, 1161–1165.

Petit, S., Chamberlain, D., Haysom, K., Pywell, R., Vickery, J., Warman, L., Allen, D. & Firbank, L. (2003) Knowledge-based models for predicting species occurrence in arable conditions. *Ecography*, **26**, 626–640.

Reese, G.C., Wilson, K.R., Hoeting, J.A. & Flather, C.H. (2005) Factors affecting species distribution predictions. A simulation modelling experiment. *Ecological Applications*, **15**, 556–564.

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

Rushton, S.P., Ormerod, S.J. & Kerby, G. (2004) New paradigms for modelling species distributions? *Journal of Applied Ecology*, **41**, 193–200.

Sanchez-Cordero, V., Cirelli, V., Munguial, M. & Sarkar, S. (2005) Place prioritization for biodiversity content using species ecological niche modelling. *Biodiversity Informatics*, **2**, 11–23.

Saseendran, S.A., Singh, S.V., Rathore, L.S. & Das, S. (2002) Characterization of weekly cumulative rainfall forecasts over meteorological subdivisions of India using a GCM. *Weather and Forecasting*, **17**, 832–844.

Schmidt, M., Kreft, H., Thiombiano, A. & Zizka, G. (2005) Herbarium collections and field data-based plant diversity maps for Burkina Faso. *Diversity and Distributions*, **11**, 509–516.

Segurado, P. & Araujo, M.B. (2004) An evaluation of methods for modelling species' distributions. *Journal of Biogeography*, **31**, 1555–1568.

Seoane, J., Carrascal, L.M., Alonso, C.L. & Palomino, D. (2005) Species-specific traits associated to prediction errors in bird habitat suitability modelling. *Ecological Modelling*, **185**, 299–308.

Shao, G. & Halpin, P.N. (1995) Climatic controls of eastern North American coastal tree and shrub distributions. *Journal of Biogeography*, **22**, 1083–1089.

Skov, F. & Svenning, J.C. (2004) Potential impact of climatic change on the distribution of forest herbs in Europe. *Ecography*, **27**, 366–380.

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

Stockwell, D.R.B. & Peterson, A.T. (2002) Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, **148**, 1–13.

Thuiller, W. (2003) BIOMOD: optimising predictions of species distributions and projecting potential future shifts under global change. *Global Change Biology*, **9**, 1353–1362.

Thuiller, W., Lavorel, S. & Araújo, M.B. (2005) Niche properties and geographical extent as predictors of species sensitivity to climate change. *Global Ecology and Biogeography*, **14**, 347–357.

Thuiller, W., Lavorel, S., Araujo, M.B., Sykes, M.T. & Prentice, I.C. (2005a) Climate change threats to plant diversity in Europe. *Proceedings of the National Academy of Sciences USA*, **102**, 8245–8250.

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

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

Tsuji, N. & Tsubaki, Y. (2004) Three new algorithms to calculate the irreplaceability index for presence/absence data. *Biological Conservation*, **119**, 487–494.

Vaughan, I.P. & Ormerod, S.J. (2003) Improving the quality of distribution models for conservation by addressing shortcomings in the field collection of training data. *Conservation Biology*, **17**, 1601–1611.

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

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

Received 4 December 2005; final copy received 24 May 2006
Editor: Jack Lennon

Appendix

We describe the generalization of TSS for a $k \times k$ contingency table, adapting Doswell, Daviesjones & Keller (1990). Let us denote the k categories as $C_i \ldots C_k$. Let n_{ij} be the number of cases that C_i was forecasted and C_j was observed. Let $n_{i.}$ be the total number of cases that C_i was forecasted, $n_{.j}$ be the total number of cases that C_j was observed, and $n_{..}$ the total number of cases. The value expected to occur by chance for the ijth cell is given by $E_{ij} = (n_{i.})(n_{.j})/n_{..}$ and should be subtracted from the observed ijth element n_{ij} of the matrix n to remove success due to random guessing.

Let us now denote by R the matrix whose elements are $R_{ij} = n_{ij} - E_{ij}$, and construct a standard matrix R^*, that will be compared to R. R^* is a matrix of perfect forecasts, accounting for random guessing. A matrix n^* of perfect forecasts has $n_{.i}$ in the iith diagonal element and all zeros in the off-diagonal elements. The number of cases expected to occur by chance is given by $E^*_{ij} = (n^*_{i.})(n^*_{.j})/n^*_{..}$ and thus $R^* = n^* - E^*$. The trace of R (the sum of the elements in the main diagonal) gives the number of correct forecasts beyond those attributable to chance. Using the trace of R^* as a standard we define the generalized version of TSS as

$$\text{TSS} = \frac{tr(R)}{tr(R^*)} = \frac{tr(n - E)}{tr(n^* - E^*)} \qquad \text{eqn 5}$$

Using the above notation Cohen's Kappa is defined by:

$$\text{Kappa} = \frac{tr(n - E)}{tr(n^* - E)} \qquad \text{eqn 6}$$

It is now clear that TSS = Kappa whenever E^* is replaced with E.