Journal of Applied Ecology 2001
38, 921–931

Evaluating presence–absence models in ecology: the need to account for prevalence

STÉPHANIE MANEL*, H. CERI WILLIAMS† and S.J. ORMEROD†
**Laboratoire de Biologie des Populations d'Altitude, UMR CNRS 5553, Université Joseph Fourier BP53 X, 38041 Grenoble, Cedex 09, France; and †Catchment Research Group, School of Biosciences, Cardiff University, PO Box 915, Cardiff CF1 3TL, UK*

Summary

1. Models for predicting the distribution of organisms from environmental data are widespread in ecology and conservation biology. Their performance is invariably evaluated from the percentage success at predicting occurrence at test locations.
2. Using logistic regression with real data from 34 families of aquatic invertebrates in 180 Himalayan streams, we illustrate how this widespread measure of predictive accuracy is affected systematically by the prevalence (i.e. the frequency of occurrence) of the target organism. Many evaluations of presence–absence models by ecologists are inherently misleading.
3. With the same invertebrate models, we examined alternative performance measures used in remote sensing and medical diagnostics. We particularly explored receiver-operating characteristic (ROC) plots, from which were derived (i) the area under each curve (AUC), considered an effective indicator of model performance independent of the threshold probability at which the presence of the target organism is accepted, and (ii) optimized probability thresholds that maximize the percentage of true absences and presences that are correctly identified. We also evaluated Cohen's kappa, a measure of the proportion of all possible cases of presence or absence that are predicted correctly after accounting for chance effects.
4. AUC measures from ROC plots were independent of prevalence, but highly significantly correlated with the much more easily computed kappa. Moreover, when applied in predictive mode to test data, models with thresholds optimized by ROC erroneously overestimated true occurrence among scarcer organisms, often those of greatest conservation interest. We advocate caution in using ROC methods to optimize thresholds required for real prediction.
5. Our strongest recommendation is that ecologists reduce their reliance on prediction success as a performance measure in presence–absence modelling. Cohen's kappa provides a simple, effective, standardized and appropriate statistic for evaluating or comparing presence–absence models, even those based on different statistical algorithms. None of the performance measures we examined tests the statistical significance of predictive accuracy, and we identify this as a priority area for research and development.

Key-words: Cohen's kappa, logistic regression, model performance, model testing, ROC, species, validation.

Journal of Applied Ecology (2001) **38**, 921–931

Introduction

Knowledge about factors influencing the distribution of organisms is among the most important in ecology (Gaston & Blackburn 1995, 1999; Lawton 1996). The potential applications are many (Table 1), particularly at large spatial scales (Ormerod, Pienkowski & Watkinson 1999; Caldow & Racey 2000). For example, in the ecology of nuisance species or disease, there is a need to identify other sites or species that might be at risk from attack (Venier *et al.* 1998; Ferreras & Macdonald 1999; Buchan & Padilla 2000). In bio-assessment, the use of biological indicators depends on being able to

Correspondence: Professor S. J. Ormerod, Catchment Research Group, School of Biosciences, Cardiff University, PO Box 915, Cardiff CF1 3TL, UK (fax 01222 874305; e-mail ormerod@cardiff.ac.uk).

© 2001 British Ecological Society

Table 1. Areas of applied ecology aided by the ability to predict species occurrence (see the Introduction for referenced examples)

Field of application	Use of species prediction
Conservation biology	Identify sites expected to hold important species using environmental data
	Identify sites for species reintroductions
	Guide site management by manipulating features known to favour species occurrence
	Identify gaps in distribution and diagnose their cause
	Identify locations at risk of species extinction
Biological indication	Identify major influences on species distribution, hence revealing indicator value
	Discriminate effects of habitat and pollution on species distribution to diagnose which is responsible for absence
	Predict site value for important species using other biota as predictors
Nuisance species	Predict sites at risk from outbreaks
	Guide site management by manipulating features known to reduce species occurrence
Invasion ecology	Predict sites sensitive to alien invasion
	Model negative effects of non-indigenous species on native biota
All areas of applied ecology	Predict distributional change in response to changing climate or land use

Table 2. Methods used to evaluate the performance of presence–absence models in a sample of ecological publications (1989–99). All the values are percentages

Source	No evaluation	Prediction success	Kappa	Odds ratio	ROC
Key journals (n = 33 papers)	55	36	6	0	3
All others (n = 54 papers)	50	46	2	2	0
All journals (n = 87 papers)	52	43	3	1	1

Taxonomic groups most frequently involved in presence–absence models were birds (33%), mammals (20%), invertebrates (18%), trees and other angiosperms (11%), fish (9%), amphibians (3%) and reptiles (2%). Less frequent applications were from bacteria, plankton and fungi. Modelling techniques used included logistic regression (79%), discriminant analysis (22%), artificial neural networks (4%) and other methods (3%).

Key journals were *Ecology*, *Ecological Applications*, *Oikos*, *Journal of Ecology*, *Journal of Applied Ecology*, *Journal of Animal Ecology*, *Functional Ecology* and *Biological Conservation*.

discriminate the effects of habitat structure and pollution on distribution (Utzinger, Roth & Peter 1998). Perhaps most important of all, in conservation biology, there is a need to assess from environmental data those sites that might support important taxa. This need arises because the presence or range of key species is increasingly modelled from remote data (Verlinden & Masogo 1997; Wright, Fielding & Wheater 2000) and because changes in climate or land use may require us to predict how target species might respond (Buckland & Elston 1993). Alternatively, prediction can sometimes reveal additional populations of threatened species (Pfab & Witowski 1997) or, in contrast, reveal unexpected gaps in their range (Wiser, Peet & White 1998). Equally, knowledge of the environmental factors that favour key biota can guide the management of protected areas or other environments (Li *et al.* 1999; Bradbury *et al.* 2000). Increasingly, there is a need to identify, from environmental data, those areas that might be candidate locations for species reintroductions (Yanez & Floater 2000) or that have a high risk of species extinction (Araujo & Williams 2000; Gates & Donald 2000).

In all these cases, quantitative distribution models are important (Collingham *et al.* 2000; Cowley *et al.* 2000; Milsom *et al.* 2000; Suarez, Balbontin & Ferrer 2000; Wadsworth *et al.* 2000). Typically, they use abiotic or biotic variables to predict the abundance, presence or absence of the target organism(s) (Jongman, ter Braak & van Tongeren 1995). At the large spatial scales typical of conservation biology, empirical presence–absence models are often derived from survey data using correlative univariate or multivariate techniques such as discriminant analysis, logistic regression and artificial neural networks (Manel *et al.* 1999). Ideally, such models should be tested with independent data (Fielding & Bell 1997). Our evaluation of a sample of published ecological literature shows that many users of presence–absence models make no evaluation at all, even in leading ecological journals (Table 2). Where performance is assessed, invariably it involves calculating the percentage of locations at which presence or absence is correctly predicted, in other words the prediction success or matching coefficient (Buckland & Elston 1993). However, there are indications that this measure might be affected by the frequency or prevalence of the test organism(s) being modelled (Fielding & Bell 1997; Manel *et al.* 1999). If this is true, current widespread practice in ecology would be at fault, comparison between models would be misleading, and investigators might wrongly judge poorly performing models as adequate.

Predicting the presence or absence of organisms in ecology has parallels in other fields. For example, in remote sensing investigators assess whether remote images predict true categories of ground vegetation (Helmer, Brown & Cohen 2000). Even more widespread

© 2001 British Ecological Society,

13652664, 2001, 5, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1046/j.1365-2664.2001.00647.x, Wiley Online Library on [25/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

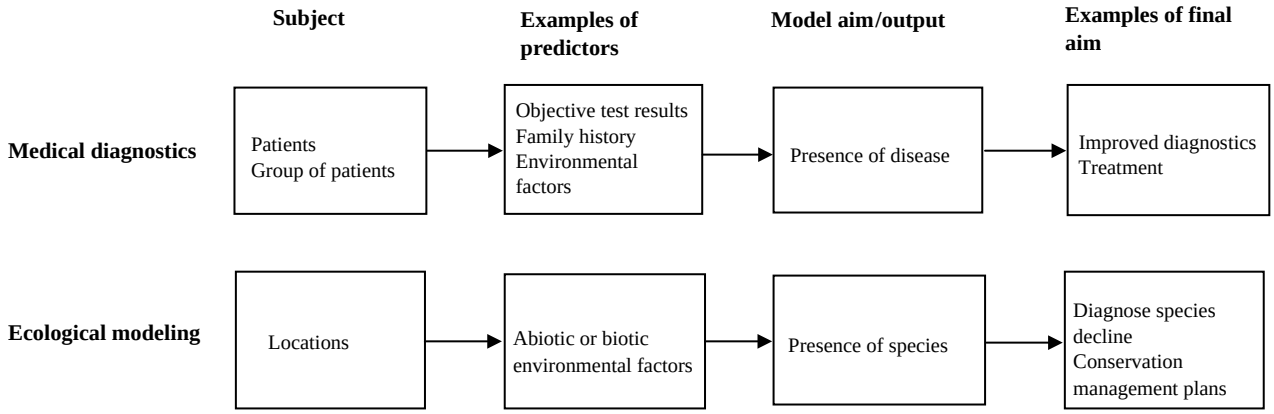

Fig. 1. The objectives of presence–absence prediction in ecology and medical diagnostics.

examples arise in medical diagnostics, where investigators detect or predict the likely presence of disease from test procedures or predisposing conditions (Fig. 1; Albert & Harris 1987; Walker, Cross & Harrison 1999). In these instances, several statistical methods have been developed to evaluate model performance (Robertson & Zweig 1981; Van Steirteghem *et al*. 1982; Zweig & Robertson 1982; Robertson, Zweig & Van Steirteghem 1983; Zweig, Broste & Reinhart 1992; Zweig & Campbell 1993). Among them are Cohen's kappa, a simply derived statistic that measures the proportion of all possible cases of presence or absence that are predicted correctly by a model after accounting for chance. They also include plots based on receiver-operating characteristics (ROC plots), which are believed to indicate model performance independently of the apparently arbitrary probability threshold required in presence–absence models at which the presence of a target feature is accepted. Fielding & Bell (1997) suggested such methods might be used to assess the performance of ecological models, but few examples are available (Titus, Mosher & Williams 1984; Monserud & Leemans 1992; Murtaugh 1996; Zimmermann & Kienast 1999; Collingham *et al*. 2000; Guisan & Zimmermann 2000; Hallgren & Pitman 2000). We know of no systematic evaluation of these alternative measures under real modelling conditions.

In this paper, we provide such a field evaluation using real data. Using logistic regression, we derived models to predict the presence and absence of 34 families of stream invertebrate families in the Himalayan mountains, and then evaluated their performance in predicting new cases. We aimed to identify performance measures that would be unaffected by the prevalence of the test organism, and would allow comparison between models from different organisms or locations. We included both conventional performance measures and also many of the alternative procedures proposed by Fielding & Bell (1997). The work forms part of a series of papers evaluating modelling procedures in ecology at coarse spatial scales (Manel, Dias & Ormerod 1999; Manel *et al*. 1999; Manel, Buckton & Ormerod 2000).

Data sources, study area and methods

The source data describe the distribution of 39 families of aquatic invertebrates in 180 Himalayan streams. All were in independent catchments spread over 4300 m of altitude and 1000 km of latitude in seven distinct regions between northern Uttar Pradesh and eastern Nepal (Manel *et al*. 1999). They form part of a data set collected to appraise systematically Himalayan aquatic biodiversity during surveys in winter (October–November) 1994–96. Chemical data were available from a full ionic analysis (Collins & Jenkins 1996) and habitat data from detailed river habitat surveys (RHS; Manel *et al*. 1999; Manel, Buckton & Ormerod 2000). The latter records 120 variables that reflect the complex physical structure of rivers. A parallel set of invertebrate and physicochemical data, collected using identical methods, was also available from 103 streams in Wales, and all the patterns assessed in this paper were checked simultaneously using this second data set (S. J. Ormerod, unpublished data).

At each site, we recorded the presence and absence of all families from the orders Plecoptera, Ephemeroptera and Trichoptera using timed kick-samples in riffles (1-min duration, net mesh size 400 μm; Manel, Buckton & Ormerod 2000). All specimens were preserved on-site and removed for identification. Although the models that follow involve family level identification, the results were also apparent at the species level as shown by the Welsh data. Samples obtained by these methods collect in excess of 70% of the families present at any site, so that the models include realistic errors in the measurement of presence–absence.

PREDICTION BY LOGISTIC REGRESSION

In all our modelling exercises, environmental predictors of presence–absence were derived from the habitat and chemical data using principal components analysis (PCA) on the correlation matrix. Because stream environments embody complex variations in hydrodynamics, hydrochemistry, geomorphology and catchment character, we separated sets of 10–40 variables,

© 2001 British Ecological Society,

13652664, 2001, 5, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1046/j.1365-2664.2001.00647.x, Wiley Online Library on [25/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

respectively, describing chemistry (ChemistryPC1–5), flow character (FlowPC1–5), channel structure (ChanPC1–5) bank structure (BankPC1–5) and riparian character (RiparPC1–5); altitude and slope were also possible predictors (x_i, $i = 1$, 27 predictor variables). While this categorized selection of variables removed some of the guaranteed orthogonality from PCA, we wished to mimic model operations where investigators aim to identify those aspects of environmental variation that best predict presence–absence (for example, chemistry, stream structure, land use). The use of so many principal components in each case is also unusual, but it was necessary in these exercises because (i) sequential variance explained by each principal component was small relative to the total eigenvalues for each predictor set and (ii) initial explorations revealed that principal components ranked as high as PC5 were sometimes highly significant predictors of presence–absence. Further exploration revealed that the number of principal components involved in modelling had no effect on our final conclusions with respect to performance measures.

We next used multiple logistic regression, involving a logit link and binomial error distribution (McCullagh & Nelder 1989; Jongman, ter Braak & van Tongeren 1995), to model the presence–absence of each family. Previous comparisons with discriminant analysis and artificial neural networks showed that logistic regression has several advantages for these purposes (Manel *et al*. 1999) but the results will apply equally to any presence–absence procedure. The logit transformation of the probability of presence–absence (p) produced linear function according to the equation:

$$\text{logit}(p) = \log\frac{p}{1-p} = b_0 + \sum_{i=1}^{27} b_{1i}x_i$$

in which b_0 and b_i are the regression constants. Models were fitted using a maximum likelihood method (McCullagh & Nelder 1989). We used backwards elimination to select the variables in the final models (Green, Osborne & Sears 1994; Austin & Meyers 1996; Manel, Dias & Ormerod 1999) using Akaike's information criterion (AIC) and changes in scaled deviance. The latter is approximately distributed like χ^2 (McCullagh & Nelder 1989; Collett 1991). The output variables (predicted values) in each case have a value between 0 and 1, and presence for all families was initially accepted at a threshold probability of 0·5 (i.e. a fixed cut-off of $p = 0{\cdot}5$). We have investigated elsewhere the effects of varying this probability threshold on the performance of presence–absence modelling (Manel, Dias & Ormerod 1999) and below we evaluate one possible method for selecting alternative threshold values. Other suggested methods have been explored previously (Buckland & Elston 1993; Huntley *et al*. 1995).

Performance in each model was assessed both with the calibration data (i.e. 180 sites) and through a jack-knife procedure that isolated calibration sites (179) from independent test sites ($n = 1$), the latter iterated for each separate observation (i.e. 180 times; Manel *et al*. 1999; Manel, Dias & Ormerod 1999). Model tests with this jack-knife procedure provided results consistent with more rigorous procedures where calibration and test data were in geographically separate regions (Manel *et al*. 1999; Manel, Dias & Ormerod 1999).

EVALUATION OF MODEL PERFORMANCE AND EFFECTS OF PREVALENCE

The evaluation of performance measures for each family first required the derivation of matrices of confusion that identified true positive (a), false positive (b), false negative (c) and true negative (d) cases predicted by each model (Table 3; Fielding & Bell 1997). Conventional statistics of association on these 2×2 tables are inappropriate for assessing model performance, for example because highly significant values of χ^2 would arise just as strongly from wholly inaccurate models (high values of b and c) as from accurate models (high values of a and d). Nor would any individual value from the matrix give a synoptic view of overall model performance. From the values in the matrix of confusion, we therefore calculated alternative performance measures including overall prediction success (matching coefficient; Buckland & Elston 1993), sensitivity, specificity, the odds ratio, the normalized mutual information statistic (NMI) and Cohen's kappa (Table 4; Fielding 1999). The latter three measures share the proposed advantage of allowing an assessment of the extent to which models correctly predict occurrence at rates that are better than chance expectation (Forbes 1995; Fielding & Bell 1997). For the NMI, values range from 0 where models are completely inaccurate, to 1 where presence–absence is perfectly predicted (Forbes 1995). However, the NMI cannot be applied directly where any value in the matrix of confusion is 0 due to the dependence on logarithmic data; the NMI could not be applied to some families in our data set. The odds ratio is also affected by zero values. For kappa, values of 0·0–0·4 are considered in medical applications to indicate slight to fair model performance, values of 0·4–0·6 moderate, 0·6–0·8 substantial and 0·8–1·0 almost perfect (after Landis & Koch 1977). In each set of applications, we plotted values for each performance measure against prevalence.

Table 3. The derivation of the confusion matrix used as a basis for performance measures in presence–absence models. The table cross-tabulates observed (actual) presence/absence patterns against those predicted: a, true positive values; b, false positives; c, false negatives; d, true negatives

		Actual	
		+	–
Predicted	+	*a*	*b*
	–	*c*	*d*

13652664, 2001, 5, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1046/j.1365-2664.2001.00647.x, Wiley Online Library on [25/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 4. Possible measures for assessing the importance of presence–absence models (after Fielding & Bell 1997). The formulae are applied to assessments of correctly predicted positive occurrences (a), falsely predicted positive occurrences (b), falsely predicted negative occurrences (c) and correctly predicted negative cases (d). n is the overall number of cases

Performance measure and definition		Formula
Overall prediction success	Percentage of all cases correctly predicted (S)	$a + d/n$
Sensitivity	Percentage of true positives correctly predicted (Sn)	$a/(a + c)$
Specificity	Percentage of true negatives correctly predicted (Sp)	$d/(b + d)$
Odds ratio	Ratio of correctly assigned cases to incorrectly assigned cases	ad/cb
Negative predictive power	Percentage of predicted absences that were real	$d/(c + d)$
NMI		$\dfrac{-a \times \ln(a) - b \times \ln(b) - c \times \ln(c) - d \times \ln(d) + (a+b) \times \ln(a+b) + (c+d) \times \ln(c+d)}{n \times \ln(n) - ((a+c) \times \ln(a+c) + (b+d) \times \ln(b+d))}$
Kappa	Proportion of specific agreement	$\dfrac{[(a+d) - (((a+c)(a+b) + (b+d)(c+d))/n)]}{[n - (((a+c)(a+b) + (b+d)(c+d))/n)]}$

Methods involving ROC curves have so far been infrequently applied to ecological data (Murtaugh 1996; Manel, Dias & Ormerod 1999; Guisan & Zimmermann 2000; Pearce & Ferrier 2000). They assess performance from model output at all possible probability thresholds at which family presence might be accepted (i.e. $p > 0$ to $p < 1$). They were developed from signal-detection theory (Kraemer 1988) but have been adapted for several areas of medical diagnostics (Robertson & Zweig 1981; Van Steirteghem *et al.* 1982; Zweig & Robertson 1982; Robertson, Zweig & Van Steirteghem 1983; Zweig, Broste & Reinhart 1992; Zweig & Campbell 1993). The curve is obtained by plotting sensitivity vs. (1 – specificity) for varying probability thresholds. Good model performance is characterized by a curve that maximizes sensitivity for low values of (1 – specificity), in other words when the curve passes close to the upper left corner of the plot (Robertson, Zweig & Van Steirteghem 1983). High performance models are indicated by large areas under the ROC curves (i.e. large areas under the curve; AUC). Usually AUC values of 0·5–0·7 are taken to indicate low accuracy, values of 0·7–0·9 indicate useful applications and values of > 0·9 indicate high accuracy (Swets 1988). The approach does not place restrictive assumptions on the distribution of response variables (Kraemer 1988). It is possible to weight the probability threshold at which presence is accepted to favour sensitivity or specificity (Kraemer 1988; Forbes 1995), but in our application both were given equal weight *a priori* to simulate a case aiming to predict both presence and absence with equal success.

We produced ROC plots for each family using S-plus software (B. Atkinson, personal communication; Fig. 2) based in turn on the non-parametric method of DeLong, DeLong & Clarke-Pearson (1988). For each ROC curve we calculated the AUC, with bootstrapped confidence intervals, as a measure of model performance in its own right (Zweig & Robertson 1982). We assessed whether the AUC values were independent of invertebrate prevalence, and also assessed whether AUC correlated with

Fig. 2. An example of a receiver-operating characteristic (ROC) curve for the Hydropsychidae.

© 2001 British Ecological Society,

13652664, 2001, 5, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1046/j.1365-2664.2001.00647.x, Wiley Online Library on [25/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the other more straightforward and easily computed measures of model performance derived above.

While the AUC measure from an ROC curve is considered useful for comparing the performance of presence–absence models in a threshold-independent fashion (Fielding & Bell 1997), truly predictive modelling might require some probability at which to accept the presence of the target organism. The ROC procedure offers a way of identifying an optimum probability threshold by simply reading the point on the curve at which the sum of sensitivity and specificity is maximized (Albert & Harris 1987; Zweig & Campbell 1993). Ecologists sometimes use threshold optimization procedures (Collingham *et al*. 2000). We therefore wished to assess whether markedly increasing or reducing probability thresholds for accepting presence, as optimized from the ROC plots in this way, had any effects on the predicted frequency of occurrence of organisms during real model applications. Thus, we calibrated logistic models for predicting presence using the five western-most regions in the data set and applied them to the geographically distinct regions to the east. We have used this procedure previously as a recommended method for testing any presence–absence model on fully independent data (Manel *et al*. 1999; Guisan & Zimmermann 2000). We carried out this analysis on 20 families representing a wide range of prevalence from scarce to widespread. In each case, we compared the actual occurrence of the target invertebrate family at the test sites with occurrence predicted by the models.

Results

Thirty-nine families formed the data, with prevalence ranging from 1 to 176 out of 180 possible streams. Initial applications using all the sites gave significant logistic regression models for 34 families, four others being too scarce for meaningful prediction (Prosopistomatidae, Ecnomidae, Beraeididae and Calanoceratidae, all at one to three sites) and one being too common (Baetidae, at 176 sites). Significant effects of stream chemistry, riparian habitat character, flow character, bank character, channel structure, altitude and slope were all involved in prediction, but in different combinations between families.

PREDICTION AND PERFORMANCE WITH FIXED PROBABILITY THRESHOLDS

Overall success at predicting presence and absence always exceeded 50%, but varied curvi-linearly and highly systematically with prevalence (Fig. 3a). This reflected the composite effects of prevalence on sensitivity and specificity, positive occurrences being more effectively

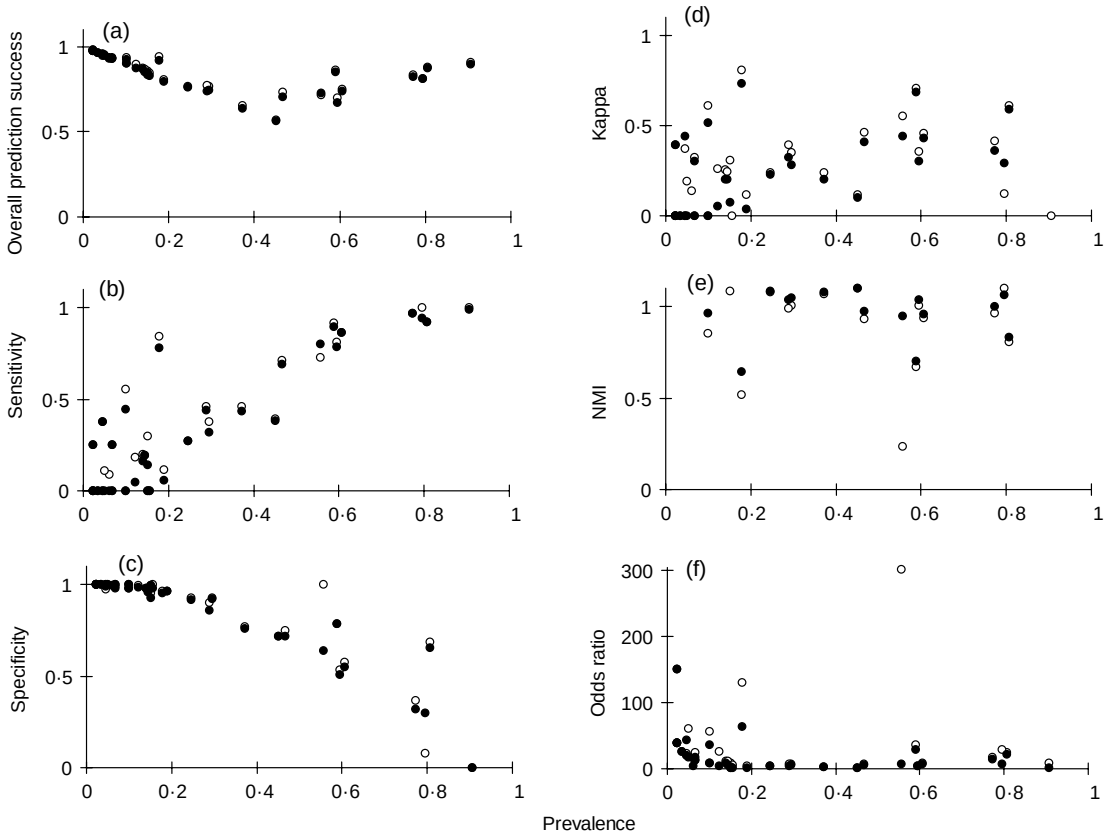

Fig. 3. The overall prediction success (a), sensitivity (b; $r_s = 0{\cdot}82$, $P < 0{\cdot}001$ in jack-knife data), specificity (c; $r_s = -0{\cdot}95$, $P < 0{\cdot}001$), kappa (d; $r_s = 0{\cdot}38$, $P < 0{\cdot}03$), NMI (e; $r_s = -0{\cdot}23$ NS, $n = 21$) and odds ratio (f; $r_s = -0{\cdot}47$, $P < 0{\cdot}007$) of presence–absence models for 34 aquatic invertebrate families in relation to their occurrence in 180 streams in the Indian and Nepali Himalaya. Results are shown for all calibration data (open symbols) and during a jack-knife application (solid symbols).

13652664, 2001, 5, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1046/j.1365-2664.2001.00647.x, Wiley Online Library on [25/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 5. The mean performance of 34 significant logistic regression models for predicting the presence–absence of aquatic invertebrate families in Himalayan streams. All values are means (with SD) resulting from applications to jack-knife data

Sensitivity	Specificity	Odds ratio	Kappa	NMI	Overall percentage success
0·365 (0·365)	0·833 (0·247)	18·42 (27·96)	0·221 (0·222)	1·20 (0·32)	0·845 (0·106)

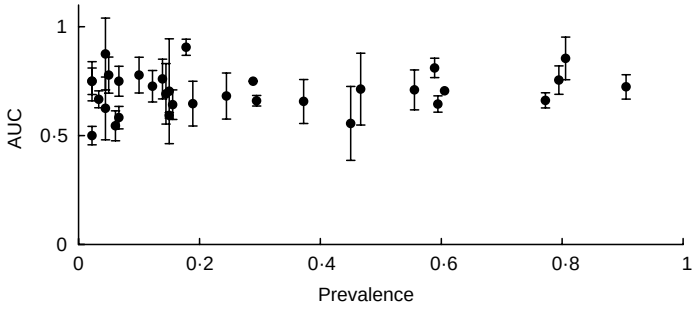

Fig. 4. Areas under the curve (AUC ± boot-strapped 95% confidence interval) from logistic regression models applied using receiver-operating characteristic (ROC) curve procedures to 34 aquatic invertebrate families in relation to their occurrence in 180 streams in the Indian and Nepali Himalaya.

predicted as prevalence increased (Fig. 3b) and negative occurrences as prevalence declined (Fig. 3b).

The effects of prevalence on the odds ratio were also highly significant ($P < 0{\cdot}01$; Fig. 3f), with patterns similar in the calibration and jack-knife data. In contrast, the NMI was unaffected by prevalence, but was incalculable in 13 families due to the occurrence of zero values in the confusion matrix. Kappa was always calculable, was only marginally affected by prevalence, and was more informative than the odds ratio in providing a range of performance values (Fig. 3d). Presence–absence models for 26/34 families gave kappa values of less than 0·4 when applied to the jack-knife data, and hence would be considered weak. Only 2/34 families (Perlidae and Taeniopterygidae) gave excellent values, while models for 6/34 families gave moderate agreement between observed and expected data (Limnephilidae, Lepidostamatidae, Helicopsychiae, Ephemerellidae, Capniidae and Hydropsychiae). This apparently poor modelling performance contrasted with the indications from overall prediction success, in which values were always > 50%, with an overall mean of 84·5% (± 10·6% SD; Table 5).

PREDICTION AND PERFORMANCE WITH ROC PROCEDURES AND VARIABLE PROBABILITY THRESHOLDS

Application of the ROC procedure produced AUC values that were wholly independent of prevalence (Fig. 4). Six families had values typical for low accuracy models (AUC < 0·7), five (Perlidae, Capniidae, Hydropsychidae, Helicopsychiae and Taeniopterygidae) had values typical for accurate models (AUC ≥ 0·9), and the remaining 23 had values between these extremes (AUC 0·7–0·9). Overall, sensitivity and specificity derived from optimized probability thresholds gave the impression that model accuracy was markedly increased over models derived at the conventional probability threshold of $p = 0{\cdot}5$. Moreover, sensitivity and specificity values produced using the ROC were independent of prevalence (Fig. 5). AUC,

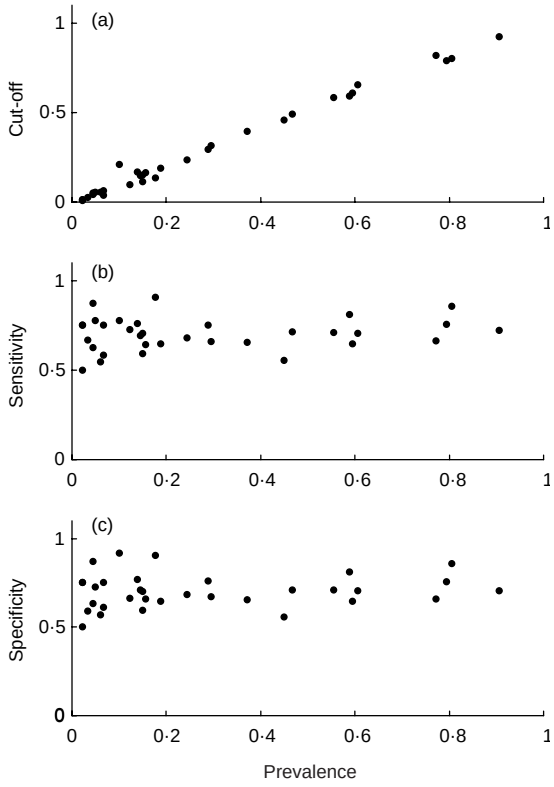

Fig. 5. Optimized probability cut-off (a), sensitivity (b) and specificity (c) for presence–absence models for 34 aquatic invertebrate families in relation to their occurrence in 180 streams in the Indian and Nepali Himalaya. In each case the models have been derived using receiver-operating characteristic (ROC) procedures and applied with the optimized threshold for accepting presence shown in (a) (see text for the methods).

© 2001 British Ecological Society,

13652664, 2001, 5, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1046/j.1365-2664.2001.00647.x, Wiley Online Library on [25/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

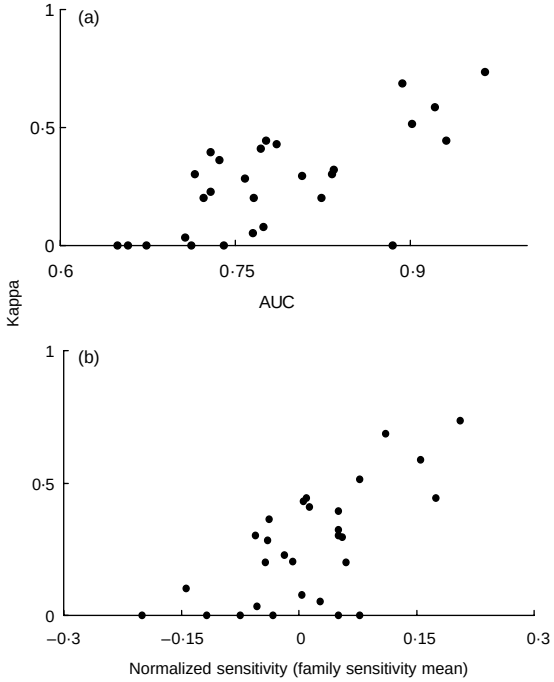

Fig. 6. The relationship between area under the curve (AUC) (a) and normalized sensitivity (b) models derived using receiver-operating characteristic (ROC) procedures and kappa (i.e. derived at a presence–absence threshold of $p = 0{\cdot}5$) for 34 families of aquatic invertebrates. Correlations were highly statistically significant in both cases.

and increases in sensitivity relative to the mean, both correlated highly significantly with values of kappa derived at $p = 0{\cdot}5$ (Fig. 6). Kappa thus provided a robust indicator of model performance when compared with the more sophisticated ROC procedure.

There were some drawbacks with the ROC procedure, however, when the models were required to predict occurrence in independent data: the derivation of optimum probability thresholds in ROC was strongly dependent on prevalence. Thus, scarce organisms were allocated low thresholds for accepting presence (sometimes $p < 0{\cdot}05$) by ROC plots, while large thresholds ($p > 0{\cdot}9$) resulted where organisms were widespread (Fig. 5). Scarce organisms might therefore be predicted erroneously as being widespread in test data when their

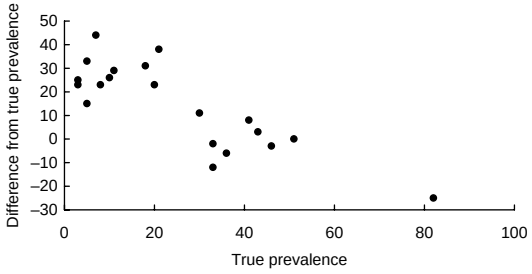

Fig. 7. The performance of logistic regression models with thresholds for accepting presence optimized using ROC curves. The value for each family is the difference between true prevalence and that apparent when models from calibration sites were applied to a subset of 61 test sites from geographically separate regions (see also Table 6 for example data).

true prevalence was low. This problem was confirmed when we applied models with optimized probability thresholds to real data from different Himalayan regions, and example data are given in full in Table 6: in widespread families, simulated prevalence in ROC model outputs underestimated true values by comparison with real data but substantially inflated the presence of scarcer families (Fig. 7). Results of this type would clearly be misleading in real model use.

Discussion

At geographical scales beyond experimentation, empirical models provide one of the only ways to develop and test hypotheses about ecological features affecting distribution (Gaston & Blackburn 1999; Manel, Buckton & Ormerod 2000; Ormerod & Watkinson 2000). Effective and correct model assessment therefore has real significance to fundamental ecology. In addition, models for predicting presence–absence are used increasingly in applied ecology and particularly in conservation biology as the numbers of threatened species rise (Table 1). However, the frequency of occurrence of organisms inevitably varies between the data sets used to produce models, and in turn there are major effects on some measures of model performance. Overall prediction success, for example, is in widespread use in ecology, even in the world's leading

Table 6. The performance of logistic regression models with conventional thresholds for accepting presence (i.e. $p = 0{\cdot}5$, left-hand columns) and with thresholds optimized using ROC curves to maximize the sum of sensitivity and specificity (right-hand columns). The values are the prevalence of each example family in reality, and in modelling predictions at all sites or at a subset of 61 test sites from geographically separate regions (see also Fig. 7 for all 20 families included in this analysis)

	All sites (n = 180)		Test regions only (n = 61 sites)	
Family	True prevalence	Model prevalence (jack-knife at $p = 0{\cdot}5$)	True prevalence	Model prevalence (p threshold optimized by ROC)
Ephemerellidae	56	41	43	46
Leptophlebiidae	14	4	20	43
Nemouridae	77	89	82	57
Chloroperlidae	12	3	3	26
Peltoperlidae	15	8	5	38

13652664, 2001, 5, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1046/j.1365-2664.2001.00647.x, Wiley Online Library on [25/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ecological journals (Table 2). Our illustration reveals that this measure is inherently misleading in failing to take account of prevalence effects (Fielding & Bell 1997; Manel *et al*. 1999). This effect is well known to statisticians but apparently not to ecologists (Fielding & Bell 1997): many users of ecological models are assuming good performance because their data are well fitted statistically, and because they can predict many occurrences correctly. More valuable are measures of performance and accuracy that account for chance success in correctly predicting cases.

Fields such as vegetation classification in remote sensing and medical diagnostics, where there are parallels with the prediction of species occurrence, have a long history of performance measures that account for prevalence. Among the suggested measures, the odds ratio and NMI are limited by difficulties with zero values in any category of the confusion matrix (Table 3), a problem that characterized almost one-third of the models we derived. Correction procedures are possible but were not explored here. The ROC method has had some application in ecology (Manel *et al*. 1999) because of assumed advantages in either assessing model performance in a threshold-independent fashion, or because it allows variation in the probability at which presence is accepted (Fielding & Bell 1997). Our analysis indicates that the AUC of an ROC plot will be independent of the prevalence of the organism being measured, and as such is a useful measure of how well a model is parameterized and calibrated. However, clear problems arose when ROC-optimized models were used to make true predictions. In this mode, model operation is no longer threshold independent, and in our evaluation the thresholds that maximized sensitivity and specificity were linearly related to the occurrence of the target organism. Scarce organisms were thus erroneously predicted to be widespread in test data. In addition to the complex procedures involved in calculation, ROC model accuracy will therefore be poor if used in this way for exactly those organisms where conservation interest is often greatest (Wiser, Peet & White 1998; Manel *et al*. 1999; Strayer 1999). The occurrence of scarce organisms might be overestimated in locations that are not surveyed, or candidate locations for reintroduction might be falsely identified. Both these errors would be costly in conservation management. Methods for setting probability thresholds that aim to recreate true prevalence in model output are almost certainly superior, while there are also advantages in specifying output as expected probabilities of species occurrence (Buckland & Elston 1993). Nevertheless, by virtue of their restricted occurrence, and the difficulty of calibrating suitable models where occupied sites are few, scarce organisms are likely to continue to pose problems in presence–absence modelling using any algorithm.

Another widespread statistic in other fields, kappa, had some advantages as a model performance measure in our application. Despite its simplicity, values correlated with performance measures derived from the more sophisticated and computationally demanding ROC procedures (i.e. the AUC). Kappa is also less affected by zero values in the confusion matrix than the NMI or odds ratio. Evidence about the effects of prevalence in our application was slightly less clear. In other biological fields there are concerns that kappa is affected by low prevalence (Ridenour & Heath 1999), and there was some weak evidence of this in our Himalayan data. However, we could find no similar effect in parallel studies undertaken in Wales at either species or family levels. The effects of prevalence on kappa therefore appear to be negligible, and certainly not sufficient to discourage widespread use in medicine.

As a measure of the proportion of all possible cases of presence or absence that are predicted by a model after accounting for chance effects, values of kappa offer a meaningful numerical variable for intercomparison between models. For example, in other work we have used kappa to assess the comparative power of stream chemistry, habitat structure and altitudinal relief for predicting invertebrate distribution in mountain streams (S. Manel & S. J. Ormerod unpublished data). At the same time, however, the categorization of performance indicated by kappa into fair, moderate, substantial and almost perfect, as suggested by Landis & Koch (1977), is clearly arbitrary. Such categories offer a useful way of benchmarking model performance and are used still in medical diagnostics. However, they are not defined on statistical criteria and offer no option for testing the statistical significance of predictive accuracy. Moreover, some commentators suggest that Kappa overestimates the degree of chance agreement (Foody 1992). We suggest that the development of test statistics for application in ecological presence–absence modelling is a priority research area, for example by bootstrapping variance estimates around predicted probabilities of occurrence. As a first step, however, our strongest recommendation is that ecologists reduce their reliance on prediction success as an indicator of model performance in favour of measures unaffected by the prevalence of target organisms, such as kappa.

One other interesting aspect of these data stems from the effectiveness of the models we derived to predict the distribution of organisms. Marked environmental variation and altitudinal range in the Himalayan mountains would suggest, intuitively, that ecological effects on organisms should be strong, and hence the potential for effective modelling should be large. We have shown this to be the case with Himalayan river birds (Manel *et al*. 1999). An examination of prediction success, and of the number of significant effects detected by regression, would lead to the view that presence–absence modelling was also effective for invertebrates. However, values of kappa in test performances indicated that models were excellent in only 3% of cases, good in 21% and poor in 76%. This outcome illustrates the importance of testing properly the predictive accuracy

© 2001 British Ecological Society,

13652664, 2001, 5, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1046/j.1365-2664.2001.00647.x, Wiley Online Library on [25/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

of models with independent data. It also illustrates that presence–absence models will sometimes provide a challenge to ecologists. This is despite the detailed and exhaustive environmental data used in modelling, which in this case should have captured many of the important physicochemical influences on aquatic invertebrates.

Acknowledgements

These data were collected under funding from the DEFRA Darwin Initiative for the Survival of Species (Himalayan data), and from the Welsh Assembly, DEFRA Air Quality Division, Environment Agency, NERC and Countryside Council for Wales (Welsh data). All the analyses and modelling presented here were made possible by a grant from the Royal Society European Science Exchange Programme. We are extremely grateful to two referees and Dr Gill Kerby for their helpful comments, and for the insight of Dr Mark Hill who took editorial responsibility for this manuscript.

References

Albert, A. & Harris, E.K. (1987) *Multivariate Interpretation of Clinical Laboratory Data*. Marcel Dekker, New York, NY.

Araujo, M.B. & Williams, P.H. (2000) Selecting areas for species persistence using occurrence data. *Biological Conservation*, **96**, 331–345.

Austin, M.P. & Meyers, J.A. (1996) Current approaches to modelling the environmental niche of eucalyptus: implications for management of forest biodiversity. *Forest Ecology and Management*, **85**, 95–106.

Bradbury, R.B., Kyrkos, A., Morris, A.J., Clark, S.C., Perkins, A.J. & Wilson, J.D. (2000) Habitat selection and breeding success of yellowhammers on lowland farmland. *Journal of Applied Ecology*, **37**, 789–805.

Buchan, L.A.J. & Padilla, D.K. (2000) Predicting the likelihood of Eurasian water milfoil presence in lakes, a macrophyte monitoring tool. *Ecological Applications*, **10**, 1442–1455.

Buckland, S.T. & Elston, D.A. (1993) Empirical models for the spatial distribution of wildlife. *Journal of Applied Ecology*, **30**, 478–495.

Caldow, R.W.G. & Racey, P.A. (2000) Introduction: large-scale processes in ecology and hydrology. *Journal of Applied Ecology*, **37** (Supplement 1), 6–12.

Collett, D. (1991) *Modelling Binary Data*. Chapman & Hall, London, UK.

Collingham, Y.C., Wadsworth, R.A., Huntley, B. & Hulme, P.E. (2000) Prediction of the spatial distribution on non-indigenous weeds: issues of spatial scale and extent. *Journal of Applied Ecology*, **37** (Supplement 1), 13–27.

Collins, R. & Jenkins, A. (1996) The impact of agricultural land use on stream chemistry in the Middle Hills of the Himalayas, Nepal. *Journal of Hydrology*, **185**, 71–86.

Cowley, M.J.R., Wilson, R.J., Leon-Cortes, J.L., Gutierrez, D., Bulman, C.R. & Thomas, C.D. (2000) Habitat-based statistical models for predicting the spatial distribution of butterflies and day-flying moths in a fragmented landscape. *Journal of Applied Ecology*, **37** (Supplement 1), 60–72.

DeLong, E.R., DeLong, D.M. & Clarke-Pearson, D.I. (1988) Comparing the areas under two or more correlated receiver operating characteristics curves – a non-parametric approach. *Biometrics*, **44**, 837–845.

Ferreras, P. & Macdonald, D.W. (1999) The impact of American mink *Mustella vison* on water birds in the upper Thames. *Journal of Applied Ecology*, **36**, 701–708.

Fielding, A.H. (1999) *Machine Learning Methods for Ecological Applications*. Kluwer Academic Publishers, Norwell, MA.

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.

Foody, G.M. (1992) On the compensation for chance agreement in image classification accuracy assessment. *Photogrammetric Engineering and Remote Sensing*, **58**, 1459–1460.

Forbes, A.D. (1995) Classification-algorithm evaluation: five performance measures based on confusion matrices. *Journal of Clinical Monitoring*, **11**, 189–206.

Gaston, K.J. & Blackburn, T.M. (1995) Mapping biodiversity using surrogates for species richness – macro-scales and new world birds. *Proceedings of the Royal Society of London, Series B, Biological Sciences*, **262**, 335–341.

Gaston, K.J. & Blackburn, T.M. (1999) A critique for macroecology. *Oikos*, **84**, 353–368.

Gates, S. & Donald, P.F. (2000) Local extinction of British farmland birds and the prediction of further loss. *Journal of Applied Ecology*, **37**, 806–820.

Green, R.E., Osborne, P.E. & Sears, E.J. (1994) The distribution of passerine birds in hedgerows during the breeding season in relation to characteristics of hedgerows and adjacent farmlands. *Journal of Applied Ecology*, **31**, 677–692.

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

Hallgren, W.S. & Pitman, A.J. (2000) The uncertainty in simulations by a global biome model (BIOMES) to alternative parameter values. *Global Change Biology*, **6**, 483–495.

Helmer, E.H., Brown, S. & Cohen, W.B. (2000) Mapping montane tropical forest successional stage and land-use with multi-date Landsat imagery. *International Journal of Remote Sensing*, **21**, 2163–2183.

Huntley, B., Berry, P.M., Cramer, W. & McDonald, A.P. (1995) Modelling present and potential future ranges of some European higher plants using climate response surfaces. *Journal of Biogeography*, **22**, 967–1001.

Jongman, R.H.G., ter Braak, C.J.F. & van Tongeren, O.F.R. (1995) *Data Analysis in Community and Landscape Ecology*, 2nd edn. Cambridge University Press, Cambridge, MA.

Kraemer, H.C. (1988) Assessment of 2×2 associations: generalization of signal-detection methodology. *American Statistical Association*, **42**, 37–49.

Landis, J.R. & Koch, G.G. (1977) The measurements of observer agreement for categorical data. *Biometrics*, **33**, 159–174.

Lawton, J. (1996) Patterns in ecology. *Oikos*, **75**, 145–147.

Li, W.J., Wang, Z.J., Ma, Z.J. & Tang, H.X. (1999) Designing the core zone in a biosphere reserve based on suitable habitats: Yanchang Biosphere Reserve and the red-crowned crane (*Grus japonensis*). *Biological Conservation*, **90**, 167–173.

McCullagh, P. & Nelder, J.A. (1989) *Generalized Linear Models*, 2nd edn. Monographs on Statistics and Applied Probability. Chapman & Hall, London, UK.

Manel, S., Buckton, S.T. & Ormerod, S.J. (2000) Problems and possibilities in large-scale surveys: the effects of land use on the habitats, invertebrates and birds of Himalayan rivers. *Journal of Applied Ecology*, **37**, 756–770.

Manel, S., Dias, J.M., Buckton, S.T. & Ormerod, S.J. (1999) Alternative methods for predicting species distribution: an illustration with Himalayan river birds. *Journal of Applied Ecology*, **36**, 734–747.

© 2001 British Ecological Society,

13652664, 2001, 5, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1046/j.1365-2664.2001.00647.x, Wiley Online Library on [25/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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.

Milsom, T.P., Langton, S.D., Parkin, W.K., Peel, S., Bishop, J.D., Hart, J.D. & Moore, N.P. (2000) Habitat models of bird species distribution: an aid to the management of coastal grazing marshes. *Journal of Applied Ecology*, **37**, 706–728.

Monserud, R.A. & Leemans, R. (1992) Comparing global vegetation maps with the kappa statistic. *Ecological Modelling*, **62**, 275–293.

Murtaugh, P.A. (1996) The statistical evaluation of ecological indicators. *Ecological Applications*, **6**, 132–139.

Ormerod, S.J. & Watkinson, A.R. (2000) Large-scale ecology and hydrology: an introductory perspective from the editors of the *Journal of Applied Ecology*. *Journal of Applied Ecology*, **37** (Supplement 1), 1–5.

Ormerod, S.J., Pienkowski, M.W. & Watkinson, A.R. (1999) Communicating the value of ecology. *Journal of Applied Ecology*, **36**, 847–855.

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

Pfab, M.F. & Witowski, E.T.F. (1997) Use of geographical information systems in the search for additional populations, or sites suitable for re-establishment, of the endangered Northern Province endemic *Euphorbia civicola*. *South African Journal of Botany*, **63**, 351–355.

Ridenour, T.A. & Heath, A.C. (1999) A note on issues in meta-analysis for behavioural genetic studies using categorical phenotypes. *Behavioural Genetics*, **29**, 155–162.

Robertson, E.A. & Zweig, M.H. (1981) Use of receiver operating curves to evaluate the clinical performance of analytical systems. *Clinical Chemistry*, **27**, 1569–1574.

Robertson, E.A., Zweig, M.H. & Van Steirteghem, M.D. (1983) Evaluating the clinical accuracy of laboratory tests. *American Journal of Clinical Pathology*, **79**, 78–86.

Strayer, D.L. (1999) Statistical power of presence absence data to detect population declines. *Conservation Biology*, **13**, 1034–1038.

Suarez, S., Balbontin, J. & Ferrer, M. (2000) Nesting habitat selection by booted eagles *Hieraaetus pennatus* and implications for management. *Journal of Applied Ecology*, **37**, 215–223.

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

Titus, K., Mosher, J.A. & Williams, B.K. (1984) Chance-corrected classification for use in discriminant analysis: ecological applications. *American Midland Naturalist*, **111**, 1–7.

Utzinger, J., Roth, C. & Peter, A. (1998) Effects of environmental parameters on the distribution of bullhead *Cottus gobio* with particular consideration of the effects of obstructions. *Journal of Applied Ecology*, **35**, 882–892.

Van Steirteghem, A.C., Zweig, M.H., Robertson, E.A., Bernard, R.M., Putzeys, G.A. & Bieva, C. (1982) Comparison of the effectiveness of four clinical chemical assays in classifying patients' chest pain. *Clinical Chemistry*, **28**, 1319–1324.

Venier, L.A., Hopkin, A.A., McKenny, D.W. & Wang, Y. (1998) A spatial, climate-determined risk rating for *Scleroderris* disease of pines in Ontario. *Canadian Journal of Forest Research*, **28**, 1398–1405.

Verlinden, A. & Masogo, R. (1997) Satellite remote sensing of habitat suitability for ungulates and ostrich in the Kalahari of Botswana. *Journal of Arid Environments*, **35**, 563–574.

Wadsworth, R.A., Collingham, Y.C., Willis, S.G., Huntley, B. & Hulme, P.E. (2000) Simulating the spread and management of alien riparian weeds: are they out of control? *Journal of Applied Ecology*, **37** (Supplement 1), 28–32.

Walker, A.J., Cross, S.S. & Harrison, R.F. (1999) Visualisation of biomedical datasets by use of growing cell structure networks: a novel diagnostic classification technique. *Lancet*, **354**, 1518–1521.

Wiser, S., Peet, R.K. & White, P.S. (1998) Prediction of rare-plant occurrence: a southern Appalachian example. *Ecological Applications*, **8**, 909–920.

Wright, A., Fielding, A.H. & Wheater, C.P. (2000) Predicting the distribution of European badger (*Meles meles*) setts over an urbanized landscape: a GIS approach. *Photogrammetric Engineering and Remote Sensing*, **66**, 423–428.

Yanez, M. & Floater, G. (2000) Spatial distribution and habitat preference of the endangered tarantula *Brachypelma klaasi* (Araneae: Theraphosidae) in Mexico. *Biodiversity and Conservation*, **9**, 795–810.

Zimmermann, N.E. & Kienast, F. (1999) Predictive mapping of alpine grasslands in Switzerland: species versus community approach. *Journal of Vegetation Science*, **10**, 469–482.

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

Zweig, M.H. & Robertson, E.A. (1982) Why we need better test evaluations. *Clinical Chemistry*, **28**, 1272–1276.

Zweig, M.H., Broste, S.K. & Reinhart, R.A. (1992) ROC curve analysis: an example showing the relationships among serum lipid and apolipoprotein concentration in identifying patients with coronary artery disease. *Clinical Chemistry*, **38**, 1425–1428.

Received 28 March 2000; *revision received 24 April 2001*

© 2001 British Ecological Society,

13652664, 2001, 5, Downloaded from https://besjournals.onlinelibrary.wiley.com/doi/10.1046/j.1365-2664.2001.00647.x, Wiley Online Library on [25/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License