ELSEVIER

Ecological Modelling 133 (0000) 225–245

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

Evaluating the predictive performance of habitat models developed using logistic regression

Jennie Pearce*, Simon Ferrier

NSW National Parks and Wildlife Service, PO Box 402, Armidale, NSW, 2350, Australia

Accepted 3 May 2000

Abstract

The use of statistical models to predict the likely occurrence or distribution of species is becoming an increasingly important tool in conservation planning and wildlife management. Evaluating the predictive performance of models using independent data is a vital step in model development. Such evaluation assists in determining the suitability of a model for specific applications, facilitates comparative assessment of competing models and modelling techniques, and identifies aspects of a model most in need of improvement. The predictive performance of habitat models developed using logistic regression needs to be evaluated in terms of two components: reliability or calibration (the agreement between predicted probabilities of occurrence and observed proportions of sites occupied), and discrimination capacity (the ability of a model to correctly distinguish between occupied and unoccupied sites). Lack of reliability can be attributed to two systematic sources, calibration bias and spread. Techniques are described for evaluating both of these sources of error. The discrimination capacity of logistic regression models is often measured by cross-classifying observations and predictions in a two-by-two table, and calculating indices of classification performance. However, this approach relies on the essentially arbitrary choice of a threshold probability to determine whether or not a site is predicted to be occupied. An alternative approach is described which measures discrimination capacity in terms of the area under a relative operating characteristic (ROC) curve relating relative proportions of correctly and incorrectly classified predictions over a wide and continuous range of threshold levels. Wider application of the techniques promoted in this paper could greatly improve understanding of the usefulness, and potential limitations, of habitat models developed for use in conservation planning and wildlife management.

Keywords: Logistic regression; Model evaluation; Prediction; Relative operating characteristic curve

1. Introduction

The use of statistical models to predict the likely occurrence or distribution of species is becoming an increasingly important tool in conservation planning and wildlife management. Such

* Corresponding author. Present address: Landscape Analysis and Applications Section, Canadian Forest Service-Sault Ste. Marie, Great Lakes Forestry Centre, 1219 Queen Street East, PO Box 490, Sault Ste. Marie, Ont., Canada P6A 5M7. Tel.: +1-705-7595740 ext. 2306; fax: +1-705-7595700.

E-mail address: jpearce@nrcan.gc.ca (J. Pearce).

0304-3800/00/$ - see front matter
PII: S0304-3800(00)00322-7

modelling often employs logistic regression to model the presence or absence of a species at a set of survey sites in relation to environmental or habitat variables, thereby enabling the probability of occurrence of the species to be predicted at unsurveyed sites. These models are usually fitted using generalised linear modelling (McCullogh and Nelder, 1989) or generalised additive modelling (Hastie and Tibshirani, 1990). For example, logistic regression has been used widely to predict the occurrence and habitat use of endangered vertebrate species (Ferrier, 1991; Lindenmayer et al., 1991; Mills et al., 1993; Pearce et al., 1994; Mladenoff et al., 1999), game species (Straw et al., 1986; Diffenbach and Owen, 1989) and vascular plants (Austin et al., 1990), to examine the response of species to environmental perturbation (Reckhow et al., 1987), and to model the regional distributions of a large number of fauna and flora species to provide information for regional forest conservation planning (Osborne and Tigar, 1992; NSW NPWS, 1994a,b).

Evaluating the predictive performance of models is a vital step in model development. Such evaluation assists in determining the suitability of a model for specific applications. It also provides a basis for comparing different modelling techniques and competing models, and for identifying aspects of a model most in need of improvement. Although the use of statistical modelling techniques such as logistic regression is increasing, relatively little attention has been devoted to the development and application of appropriate evaluation techniques for assessing the predictive performance of habitat models.

To obtain an unbiased estimate of a model's predictive performance, evaluation is best undertaken with independent data collected from sites other than those used to develop the model. These independent sites should be sampled representatively from the region across which the model is likely to be extrapolated. If independent data are not available, then statistical resampling techniques such as cross-validation (Stone, 1974) and jackknifing (Efron, 1982), may be used to reduce bias in the measurement of predictive performance. In cross-validation the model development sites are divided into K groups of roughly equal size (in the special case of jackknifing each group consists of just one site). For each group of sites a model is fitted to the data from the other $K-1$ groups. This model is used to predict a probability of occurrence for each of the sites in the group excluded from the fitting of the model. This procedure is repeated for all groups until predicted values have been calculated for all sites. These predicted values are then used to assess the accuracy of the predictive model. Cross-validation is a less rigorous approach to model evaluation than using a truly independent dataset, particularly in situations where the model development sites are not distributed representatively across the region under consideration, in terms of both environmental and geographical coverage.

Good predictions are both reliable and discriminatory. Reliable predictions may be used at 'face value', as each predicted probability is an accurate estimate of the likelihood of detecting the species at a given site. A model with good discrimination ability is one that can correctly discriminate between occupied and unoccupied sites in an evaluation dataset, irrespective of the reliability of the predicted probabilities. The measurement of discrimination performance requires a different approach to that used to measure reliability.

This paper adopts a framework for model evaluation developed by Murphy and Winkler (1987, 1992) that explicitly defines the links between model reliability and discrimination. Techniques for measuring each aspect of predictive performance are described and their calculation demonstrated using models developed for two species in north-east New South Wales. The paper also discusses the relevance and importance of each of these measures in the application of predictive models to conservation planning and management. While the evaluation techniques presented are described primarily in relation to logistic regression, these techniques can potentially be applied to any type of model that generates predicted probabilities of species occurrence (e.g. decision trees, artificial neural networks).

This paper considers only a single component of model performance, that of the accuracy of model predictions. Other, equally important aspects of model performance, such as the rational-

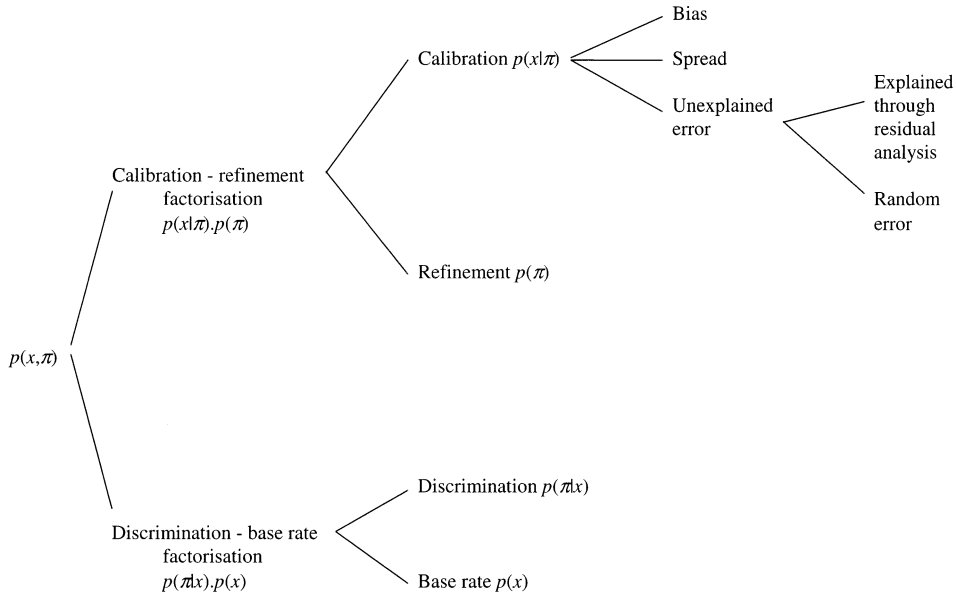

Fig. 1. Framework for assessing the predictive performance of models, describing the links between each measurable component of performance described in the text.

ity and interpretability of the explanatory variables included in the model and the validity of the predicted shape of the response curves, will not be discussed here.

2. A framework for evaluating predictive performance

Murphy and Winkler (1987, 1992) have developed a framework for assessing predictive performance of models that explicitly links model reliability and discrimination ability. This framework, summarised in Fig. 1, also allows prediction error to be partitioned between a number of sources. Their approach is based on the joint distribution of predictions, denoted π, and observations, denoted x, which can be represented as $p(\pi,x)$. The best approach to understanding this joint distribution is to consider it as a frequency table of observed and predicted values as shown in Fig. 2. For models derived from species presence/absence data the value predicted by the model for each evaluation site is the probability of occurrence of the species concerned π. The observation obtained at each evaluation site is the presence or absence of the species x. To define specific characteristics of the performance of a predictive model, Murphy and Winkler factorise the joint distribution of π and x into a conditional distribution and a marginal distribution, using two different factorisations. Each factorisation may be considered as viewing the frequency table from a different direction. The first factorisation is based on the predictions (the columns of Fig. 2), and involves the conditional distribution of the observations given the predictions $p(x/\pi)$, and the marginal distribution of the predictions $p(\pi)$:

	π_1	...	π_k	
$x=0$	n_{01}	...	n_{0k}	$n_{0.}$
$x=1$	n_{11}	...	n_{1k}	$n_{1.}$
	$n_{.1}$	...	$n_{.k}$	$n_{n..}$

Fig. 2. The framework described by Murphy and Winkler (1987, 1992) can be considered as a frequency table of observations x, and predicted values from the model π for a given set of evaluation sites.

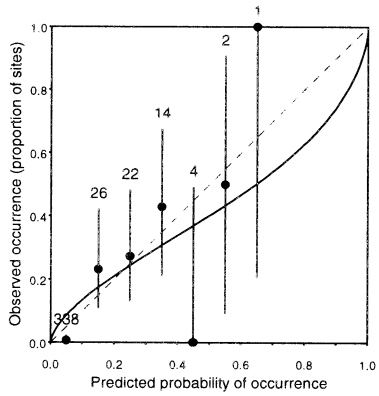

Fig. 3. Relationship between the predicted probability of occurrence and the observed proportion of evaluation sites occupied for the Yellow Box *E. melliodora* predictive model developed for north-east New South Wales. The graph is developed by plotting the proportion of evaluation sites found to be occupied within each of ten equi-interval predicted probability classes. These proportions are then plotted against the median value of each class. The number of evaluation sites and the confidence interval of the observed proportion of observations are shown for each class. The regression line describes the overall relationship between the observed and predicted values.

$$p(\pi,x) = p(x/\pi)\cdot p(\pi) \quad (1)$$

The marginal distribution of the predictions describes the frequency distribution of predicted probability values within the evaluation sample (shown by the bottom row of the frequency table in Fig. 2). The conditional distribution of the observations given the predictions describes the distribution of observed values (presence or absence) obtained for each unique predicted probability (the individual columns of Fig. 2).

The second factorisation is based on the observations (the rows of Fig. 2), and involves the conditional distribution of the predictions given the observations $p(\pi/x)$ and the marginal distribution of the observations $p(x)$:

$$p(\pi,x) = p(\pi/x)\cdot p(x) \quad (2)$$

The marginal distribution of the observations describes the distribution of observed values (presence or absence) within the evaluation sample (shown by the third column of the frequency table in Fig. 2). The conditional distribution of the predictions given the observations describes the distribution of predicted values for each unique observed value (the individual rows of Fig. 2). For presence-absence data, the observations have only two unique values, and so the conditional distribution can be thought of as consisting of two frequency distributions, the first describing the distribution of predicted values for evaluation sites at which the species was observed, and the second describing the distribution of predicted values for sites at which the species was not observed.

2.1. Factorisation based on predictions

Murphy and Winkler (1987) describe the factorisation based on predictions as the calibration-refinement factorisation, where the conditional distribution $p(x/\pi)$ reflects model calibration, and the marginal distribution $p(\pi)$ reflects model refinement.

Refinement relates to the range of predictions produced by the model for a given set of sites. A model is well refined if predictions cover the full probability range, with predicted values near both one and zero. The variance of the predictions (σ_π) provides a measure of model refinement, with large values indicating a greater level of refinement than small values. It is necessary to have at least a moderate level of refinement in order to be able to examine model performance further.

Calibration relates to the level of agreement between predictions generated by a model and actual observations. This can be examined graphically by breaking the predicted probability range up into classes, and plotting the proportion of evaluation sites that are observed to be occupied within each of these classes against the median predicted value of each class, as shown in Fig. 3. If the model is well calibrated then the points should lie along a 45° line. The lack of agreement may be partitioned into three components: bias, spread, and unexplained error. These components and their implications for species modelling are described in detail later in this paper.

2.2. Factorisation based on observations

The factorisation based on the observations is described by Murphy and Winkler (1987) as the discrimination-base rate factorisation, where the conditional distribution $p(\pi/x)$ measures the discrimination ability of a model, and the marginal distribution $p(x)$ is the base rate.

The base rate indicates how often different values of x (presence or absence) occur and, in species modelling, therefore describes the prevalence of a species at a sampled set of sites (i.e. species rarity or prior probability of occurrence). Murphy and Winkler call this distribution the base rate because it provides information on the probability of a species being observed as present at a randomly selected site, in the absence of a predictive model. The base rate is measured as the proportion of evaluation sites at which a species is observed to be present. This value needs to be moderately large in order for the predictive performance of a model to be examined.

The conditional distribution $p(\pi/x)$ indicates how often different values of π are predicted for sites at which a particular value of x is observed. For presence/absence observations of a species, $p(\pi/x=1)$ describes the proportion of occasions that each value of π is predicted for sites at which the species has been recorded as present, and $p(\pi/x=0)$ describes the proportion of occasions that each π value is predicted for sites at which the species has been recorded as absent. These conditional probabilities indicate how well a model discriminates between sites at which the species has been observed and those at which it has not been observed. The discrimination ability of a model can be examined graphically by plotting a frequency distribution of the predicted values for occupied sites, and comparing this with a frequency distribution of the predicted values associated with unoccupied sites (see Fig. 8a). A model with good discrimination ability will show little overlap between these two distributions, with the predicted values of occupied sites being greater on average than those of unoccupied sites. The measurement of discrimination capacity is described later in this paper.

2.3. Relationship between calibration/refinement and discrimination

The two factorisations described by Murphy and Winkler (1987) are equivalent,

$$p(x/\pi)\cdot p(\pi) = p(\pi/x)\cdot p(x) \quad (3)$$

so that a model which has good calibration and refinement must also have good discrimination (as the base rate is fixed and constant). However, the converse is not necessarily true, as calibration and refinement may both be poor, yet still combine in a manner which gives good discrimination.

As suggested by the above equation, components of the calibration-refinement factorisation influence the discrimination capacity of a model. Refinement affects how well a model will potentially discriminate between observed presence and absence, as the likelihood of good discrimination increases as the predictions span a larger range of the 0 to 1 probability distribution. For example, a model that always predicts a value of 0.5, whether a site is occupied or not, lacks refinement and will have no discriminatory power, even though it may exhibit excellent calibration. As the range of predicted probabilities increases, the model has an increased likelihood of being able to discriminate between positive and negative outcomes. That is, good refinement does not necessarily imply good discrimination, but improves the potential for good discrimination. A model has good discriminatory power if the predicted probability range associated with occupied sites is higher than, and has little overlap with, the probability range associated with unoccupied sites.

However, although these two factorisations are related, they tell us about different aspects of the predictive performance of a model. The calibration/refinement factorisation tells us about the reliability of predictions; that is, how closely predicted probabilities match observed proportions of occurrence. The discrimination/base rate factorisation tells us about how well predictions can discriminate between observed presence and absence, regardless of the absolute value of those predictions. The relative importance of each performance measure — calibration, refinement and discrimination — depends on the use of the

model and the experience of the user. If the predictions are to be used as an absolute measure of probability of occurrence, then knowledge of prediction calibration (reliability) and refinement (sharpness) is essential, otherwise the predictions may be misleading. If, on the other hand the predictions from a model are to be used only as a relative measure of the likelihood of occurrence (for example, to rank areas for reservation), then only discrimination ability need be examined. However, if this relative index is to be broken into two or more classes (for example, suitable versus unsuitable habitat, or areas with high, moderate and low conservation potential), then model calibration and refinement need to be examined (at least visually) to select appropriate class thresholds.

3. Measuring discrimination performance

The discrimination performance of wildlife habitat models derived by logistic regression is often assessed by examining the agreement between predictions and actual observations, using a 2×2 classification table as shown in Fig. 4 (Lindenmayer et al., 1990; Pearce et al., 1994). A species is predicted to be present or absent at a site based on whether the predicted probability for the site is higher or lower than a specified threshold probability value.

The table can be used to calculate four indices describing predictive performance of models. Two

	Recorded present	Recorded absent	
Predicted present	A	B	$A+B$
Predicted absent	C	D	$C+D$
	$A+C$	$B+D$	$A+B+C+D$

Fig. 4. The classification table describing the agreement between the observed presence and absence of a species and the predicted presence or absence of a species. Each of the values A, B, C and D, represent numbers of observations, so that their sum equals the sample size of the evaluation sample $(A+B+C+D)$.

of these indices — sensitivity (or the true positive fraction) and specificity (or the true negative fraction) — measure the proportion of sites at which the observations and the predictions agree. The other two indices — the false positive fraction and the false negative fraction — measure the proportion of sites at which the observations and the predictions disagree. These four indices are defined as follows (Fig. 4):

Sensitivity

$$= \frac{\text{Number of positive sites correctly predicted}}{\text{Total number of positive sites in sample}}$$

$$= \frac{A}{(A+C)} \tag{4}$$

Specificity

$$= \frac{\text{Number of negative sites correctly predicted}}{\text{Total number of negative sites in sample}}$$

$$= \frac{D}{(B+D)} \tag{5}$$

False positive fraction

$$= \frac{\text{Number of false positive predictions}}{\text{Total number of negative sites in sample}}$$

$$= \frac{B}{(B+D)} \tag{6}$$

False negative fraction

$$= \frac{\text{Number of false negative predictions}}{\text{Total number of positive sites in sample}}$$

$$= \frac{C}{(A+C)} \tag{7}$$

Using these indices, the accuracy (i.e. the total fraction of the sample that is correctly predicted by the model) can be calculated as:

$$\text{Accuracy} = \frac{A+D}{A+B+C+D} \tag{8}$$

However, this measure of accuracy can be misleading, as its interpretation depends on a knowledge of the prior probability of occurrence (or base rate) of the species in question, or $p(x)$. For example, if the species occurs at only 5% of sites surveyed, a high predictive accuracy can be ob-

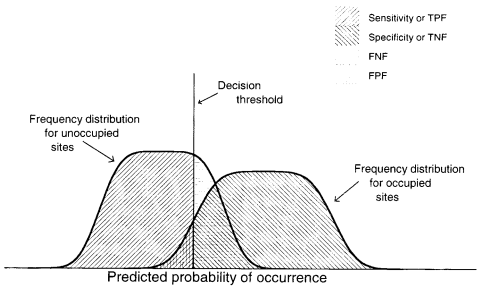

Fig. 5. An example of the model that underlies ROC analysis. The curves represent the frequency distribution of probabilities predicted by a model for occupied and unoccupied sites within a data set for which the real distribution of the species is known. A threshold probability, represented by the vertical line, separates sites predicted to be occupied from sites predicted to be unoccupied.

tained if the species is predicted to never occur. The predictions will then be correct 95% of the time. A true accuracy measure should not be sensitive to the relative frequency of the species within the test sample. This measure is also a poor indicator of relative predictive accuracy, as two models may have the same accuracy, but perform differently with respect to the types of correct and incorrect predictions they provide. The incorrect predictions from one model may be all false negatives, while for another model they might be all false positives. The nature of incorrect predictions must be examined further to properly interpret the performance of a model, as different types of error have different implications for how a model can be applied.

As described by Murphy and Winkler (1987), the discrimination capacity of a model can be evaluated graphically by plotting the two conditional distributions $p(\pi/x=1)$ and $p(\pi/x=0)$ on the same axis and examining the degree of overlap. Two such distributions are depicted in Fig. 5. The horizontal axis represents the predicted probability of a site being occupied, as derived from the model. The two curves represent the frequency distribution of predicted probabilities for two sets of evaluation sites: sites at which the species was observed as present (occupied sites) and sites at which the species was recorded as absent (unoccupied sites). The distribution of predicted values for the occupied sites, $p(\pi/x=1)$, should lie to the right of that for the unoccupied sites, $p(\pi/x=0)$, if a model is to be of any use in terms of discrimination ability. Unless this discrimination capacity is perfect, the two distributions will overlap, and some predicted values may be associated with either occupied or unoccupied sites. To generate a 2×2 classification table, a decision threshold (the vertical line) must be identified to separate sites predicted to be occupied from those predicted to be unoccupied. For a given decision threshold the proportion of the distribution of occupied sites falling to the right of this threshold defines the sensitivity (true positive fraction) of a model, while the proportion of unoccupied sites falling to the left of the thresholds defines the specificity (or true negative fraction), as indicated in Fig. 5 (Metz, 1978). The proportion of the unoccupied distribution to the right of the threshold is the false positive fraction, while the proportion of the occupied distribution to the left of the threshold is the false negative fraction. The sensitivity, specificity, false positive rate and false negative rate will vary as the decision threshold is changed.

These traditional measures of discrimination capacity depend on the arbitrary choice of a decision threshold, which introduces a further complication into the interpretation of the classification statistics. The choice of a decision threshold in habitat modelling is usually based partly on knowledge of the prior probability of occurrence of the species of interest, i.e. its rarity, and partly on value judgements regarding the consequences of various kinds of correct and incorrect decisions (Metz, 1986; Fielding and Bell, 1997). For example, if a species is endangered and the model is intended to identify potential re-introduction sites, then it is important that the habitat chosen is indeed suitable for the species, to minimise the risk of failure. This would require the choice of a relatively high threshold probability that would result in the identification of only those sites with a high predicted probability of occurrence. The number of true positive and false positive values would be correspondingly low, as few sites would be predicted to be occupied. Conversely, if the model is used to identify areas within which proposed development may impact the species (e.g. as part of an environmental im-

pact assessment), then it is important to be precautionary by identifying all potentially suitable habitats. The decision threshold would therefore need to be more lenient, and the true positive rate and the false positive rate would be expected to be high, as the model predicts more sites to be occupied. The choice of an actual threshold value in both of these cases is essentially arbitrary, and different practitioners may choose different values. However, this choice strongly affects the relative frequencies of correct and incorrect predictions, and thus the measurement of discrimination performance. A true measure of discrimination capacity should be valid for all situations in which a model will be employed, with any decision threshold that may be appropriate for a given application.

Metz (1986), and more recently from an ecological perspective, Fielding and Bell (1997) have reviewed several of the most commonly used discrimination indices, including those traditionally employed in wildlife habitat studies. They found most indices to be unsuitable as an unbiased measure of accuracy, being dependent on species rarity and/or the choice of a threshold probability. However, both Metz (1986), and Fielding and Bell (1997) found that one index did meet the requirements of an unbiased discrimination index. This index is derived from the area under a relative operating characteristic curve.

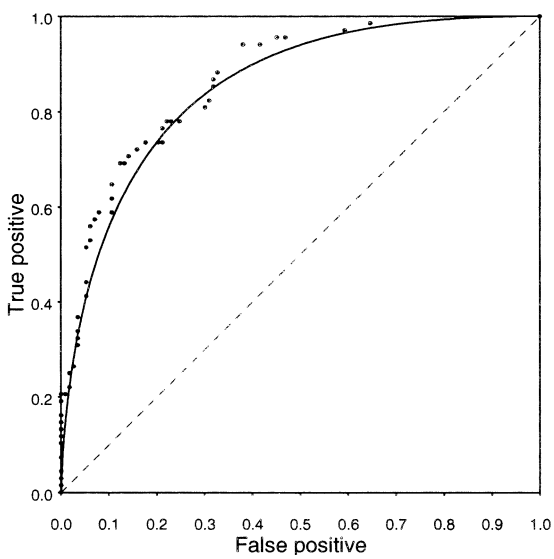

Fig. 6. The ROC graph in which the sensitivity (true positive proportion) is plotted against the false positive proportion for a range of threshold probabilities. A smooth curve is drawn through the points to derive the ROC curve. The 45° line represents the sensitivity and false positive values expected to be achieved by chance alone for each decision threshold.

3.1. The relative operating characteristic curve

From Fig. 5 it can be seen that the true positive fraction (sensitivity) plus the false negative fraction must add to 1, as do the true negative fraction (specificity) and false positive fraction. That is, for a given observed state (positive and negative), the number of correct predictions plus the number of incorrect predictions must equal the number of observations with that state. Thus, the various indices defined above are related by the following equations:

$$\text{sensitivity} + \text{false negative fraction} = 1 \quad (9)$$

$$\text{specificity} + \text{false positive fraction} = 1 \quad (10)$$

In the notation of Murphy and Winkler (1987) these equations can also be written as:

$$p(P/x = 1) + p(A/x = 1) = 1 \quad (11)$$

$$p(A/x = 0) + p(P/x = 0) = 1 \quad (12)$$

where P is a predicted value greater than or equal to the threshold probability, and A is a predicted value less than the threshold probability.

Because of these constraints, it is only necessary to specify one fraction from each equation above to describe all four performance measures. Typically, the sensitivity and the false positive fraction are specified, as these fractions describe the performance of the positive predictions, and vary in the same direction as the decision threshold is varied. Varying the decision threshold incrementally across the predicted probability range of a model will generate a series of pairs of sensitivity and false positive values. Each pair of sensitivity and false positive values can be plotted as the y and x coordinates respectively on a graph, such as that shown in Fig. 6. This series of points defines a smooth curve, which is called the relative operating characteristic (ROC) curve (Metz, 1978).

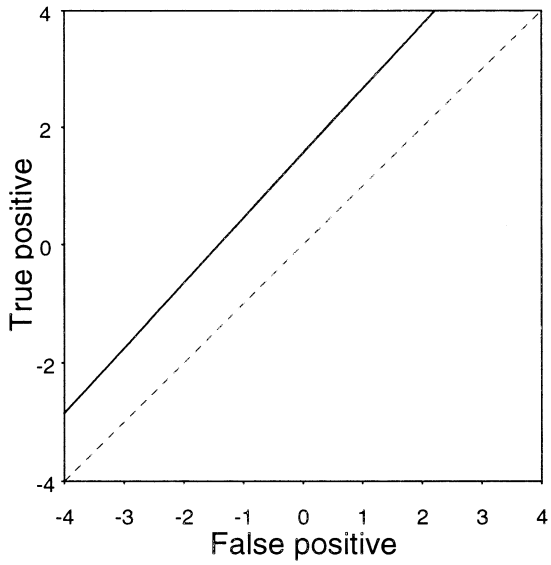

Fig. 7. The ROC curve plotted on binormal axes.

The ROC curve describes the compromises that are made between the sensitivity and false positive fractions as the decision threshold is varied. Furthermore, the sensitivity and false positive values are independent of the prevalence of a species because they are expressed as a proportion of all sites with a given observed state (Swets, 1988). ROC analysis is therefore independent of both species prevalence and decision threshold effects.

3.2. Deriving the ROC curve

Several techniques are available for fitting the ROC curve to the sensitivity and false positive data. Any number of sensitivity and false positive pairs may be used to define and fit the ROC curve, although most applications use at least five points. However, depending on the curve fitting technique used, larger numbers of sensitivity and false positive pairs will result in a better fit. We have generally employed thresholds spaced at 0.01 intervals across the predicted probability range.

The simplest technique for fitting the ROC curve is to plot the points manually with error bars, calculated using the following formulae (Metz, 1978):

$$\text{Standard error of sensitivity fraction (TPF)} = \sqrt{\frac{\text{TPF} \times (1 - \text{TPF})}{(\text{No. occupied sites} - 1) \times (\text{No. occupied sites})}} \quad (13)$$

$$\text{Standard error of false positive fraction (FPF)} = \sqrt{\frac{\text{FPF} \times (1 - \text{FPF})}{(\text{No. unoccupied sites} - 1) \times (\text{No. unoccupied sites})}} \quad (14)$$

A smooth curve is then drawn through the points and error bars. More objective and reliable methods require that some assumptions be made regarding the functional form of the ROC curve. Many functional forms have been proposed, but the binormal form is used most widely. According to the binormal model, each ROC curve is assumed to have the same functional form as that generated by two Gaussian distributions, or variables that can, by some monotonic transformation, be transformed into Gaussian form. The binormal assumption allows the ROC curve to be transformed to a straight line when plotted on normal deviate axes, as shown in Fig. 7 (Metz, 1986). Other distributional forms, including logistic, triangular and exponential, also yield approximately straight lines on a binormal graph (Swets, 1986b). The Gaussian approach is therefore applicable even if the form of the underlying distributions is not Gaussian.

The binormal ROC curve can be described by two parameters, a and b; a is the y-intercept, and b is the gradient of the straight line representing the ROC curve when plotted against normal deviate axes. Assuming a Gaussian distributional form, a can be interpreted as the distance between the means of the two distributions, $(\mu_{x=1} - \mu_{x=0})/\sigma_{x=0}$, and b can be interpreted as the ratio of the standard deviations of the two distributions $\sigma_{x=1}/\sigma_{x=0}$.

For a binormal graph the task of curve-fitting becomes one of choosing numerical values for a and b that best represent the measured data. This is usually done using maximum likelihood estimation algorithms (Metz, 1986).

If the two underlying decision distributions have equal variance then the ROC curve is symmetrical about the minor axis, as shown in Fig. 6.

However, in practice the curve tends to be skewed either toward the origin or toward the top right hand corner of the graph. On binormal axes, the symmetrical ROC curve has a gradient of 1. If the curve is asymmetrical, the gradient will tend to be less than 1 if the curve is skewed toward the origin, and greater than 1 if skewed toward the top right hand corner of the graph. These differences occur due to the distributions of the two decision variables not having equal variance. If the curve is skewed toward the origin, then the variance of the distribution for the unoccupied decision variable is larger than that for the occupied sites.

In practice, we have found the Gaussian approach to be a rapid and easily applied technique for fitting the ROC curve to the sensitivity and false positive pairs. However, we also recommend displaying the actual sensitivity and false positive points on the ROC curve, so that departures from the Gaussian assumption, if present, may be observed.

The ROC curve provides a graphical approach to assessing discrimination capacity over a range of threshold probabilities. A model that has no discrimination ability will generate an ROC curve that follows the 45° line. That is, for all decision threshold values, the sensitivity equals the false positive fraction. Perfect discrimination is indicated when the ROC curve follows the left hand and top axes of the unit square. That is, for all threshold values, the true positive fraction equals one and the false positive fraction equals zero.

The ROC curve can also be used to identify an appropriate threshold value for a given application, by balancing the cost that would arise from an incorrect decision against the benefit to be gained by a correct decision (Metz, 1978; Hilden, 1991). Murtaugh (1996) further describes, with examples, the application of ROC methodology in assessing the accuracy of ecological indicators.

3.3. Deriving a discrimination index from the ROC curve

Once an ROC curve is developed, a single index is required to describe the discrimination capacity of the model. Hilden (1991) argues that a summary measure of discrimination accuracy must incorporate cost-benefit information. However, in ecological applications the costs and benefits associated with correct and incorrect predictions are often quite intangible, making their quantification extremely difficult and, at best, arbitrary. It is therefore most practicable to develop a summary measure that assumes that the costs and benefits are equal.

Swets (1986a) has reviewed several discrimination measures that are consistent with the use of a variable decision threshold (as employed in ROC curves) and do not require cost-benefit information. Unfortunately, most of these measures assume that the two underlying decision distributions have equal variance, an assumption that is rarely met in ecological applications. Swets concluded that the best discrimination index in a range of situations appears to be the area under the ROC curve expressed as a proportion of the total area of the unit square defined by the false positive and true positive axes. This index ranges from 0.5 for models with no discrimination ability, to 1 for models with perfect discrimination.

This index can also be interpreted in terms of the true positive and false positive values used to create the curve. Areas between 0.5 and 0.7 indicate poor discrimination capacity because the sensitivity rate is not much more than the false positive rate. Values between 0.7 and 0.9 indicate a reasonable discrimination ability appropriate for many uses, and rates higher that 0.9 indicate very good discrimination because the sensitivity rate is high relative to the false positive rate (Swets, 1988). Hanley and McNeil (1982) have shown that the ROC index can also be interpreted as the probability that a model will correctly distinguish between two observations, one positive and the other negative. In other words, if a positive observation and a negative observation are selected at random the index is an estimate of the probability that the model will predict a higher likelihood of occurrence for the positive observation than for the negative observation.

The simplest technique to measure the area under the ROC curve is the direct approach, which uses the trapezoidal rule to calculate the area directly from the points on the graph. This

approach, however, can underestimate the true area under the curve if the sample is not well refined.

If the curve is derived using the parametric Gaussian approach, then the area under the curve, A, can be calculated using the two parameters a and b describing the straight line on binormal axes (Swets and Pickett, 1982). Alternatively, Brownie et al. (1986) have demonstrated how the mean and variance of the predicted values for occupied and unoccupied sites can be used to directly calculate the area under the ROC curve if the predicted values are normally distributed. This approach can result in more precise estimates of the area under the ROC curve, because the predictions and observations do not need to be first converted to true positive and false positive counts based on a series of probability thresholds. Instead the raw predicted probabilities are used directly to calculate the area under the ROC curve. Both of these statistical techniques depend on the degree to which the data meet the binormal assumption, but this assumption may not be met with ecological data.

Bambar (1975) recognised that the area under the ROC curve is intimately connected with the statistic W calculated in the Wilcoxon or Mann–Whitney statistical test of the difference between two samples (Sokal and Rohlf, 1981). The Mann–Whitney statistic therefore provides a means by which the area under an ROC curve may be calculated without the need to assume normality. The Mann–Whitney statistic is based on a comparison between the ranks of predicted values associated with positive observations and the rank of predicted values associated with negative observations. Hanley and McNeil (1982) further describe this approach, and also provide a formula by which the standard error of the area may be calculated. Bootstrapping may also be used to calculate the standard error of W (Efron and Tibshirani, 1993). The bootstrap is performed by randomly selecting a site, with replacement, n times from the total set of n evaluation sites (i.e. a given site can be represented more than once in the sample). Each sample selected in this manner is used to calculate a W value. This is repeated a large number of times, and the generated sample of W values is then used to estimate the standard error of the original Mann–Whitney value. The bootstrapping approach is preferable to that presented by Hanley and McNeil (1982).

Of the four techniques for calculating an ROC index of discrimination ability, the Mann–Whitney technique (with bootstrapping to provide a standard error) is the most reliable approach for ecological applications as it makes no distributional assumptions. This is the approach that will be used below. However, if it is known that the two decision distributions are binormal, then the parametric approach of Brownie et al. (1986) is recommended as probably providing the most accurate results.

The following example illustrates the evaluation of discrimination performance:

3.4. Example 1

The distribution of Yellow Box *Eucalyptus melliodora* in north-east New South Wales has been modelled as a function of environmental and geographical attributes using field survey data from 2223 sites distributed throughout the region (NSW NPWS 1994b). Yellow Box was recorded as present at 80 of these sites.

The following logistic regression model was fitted to the Yellow Box survey data using forward stepwise generalised additive modelling (Hastie and Tibshirani, 1990):

$$\begin{aligned}\log\text{it}(p) &= s(\text{soil depth},\ \text{df}=4)\\ &+ s(\text{moisture index},\ \text{df}=3)\\ &+ s(\text{rainfall},\ \text{df}=2) + s(\text{temperature},\ \text{df}=2)\end{aligned} \tag{15}$$

The predictive accuracy of this model was then tested using independent evaluation data collected at a further 407 sites within the region (NSW NPWS 1995a). Yellow Box was recorded at 22 of these evaluation sites.

Before calculating an ROC curve, the discrimination ability of the model was assessed visually by comparing the distribution of predicted probabilities for occupied sites with the distribution of the predicted probabilities for unoccupied sites, as shown in Fig. 8a. The graph indicates that predicted values for sites at which the species was

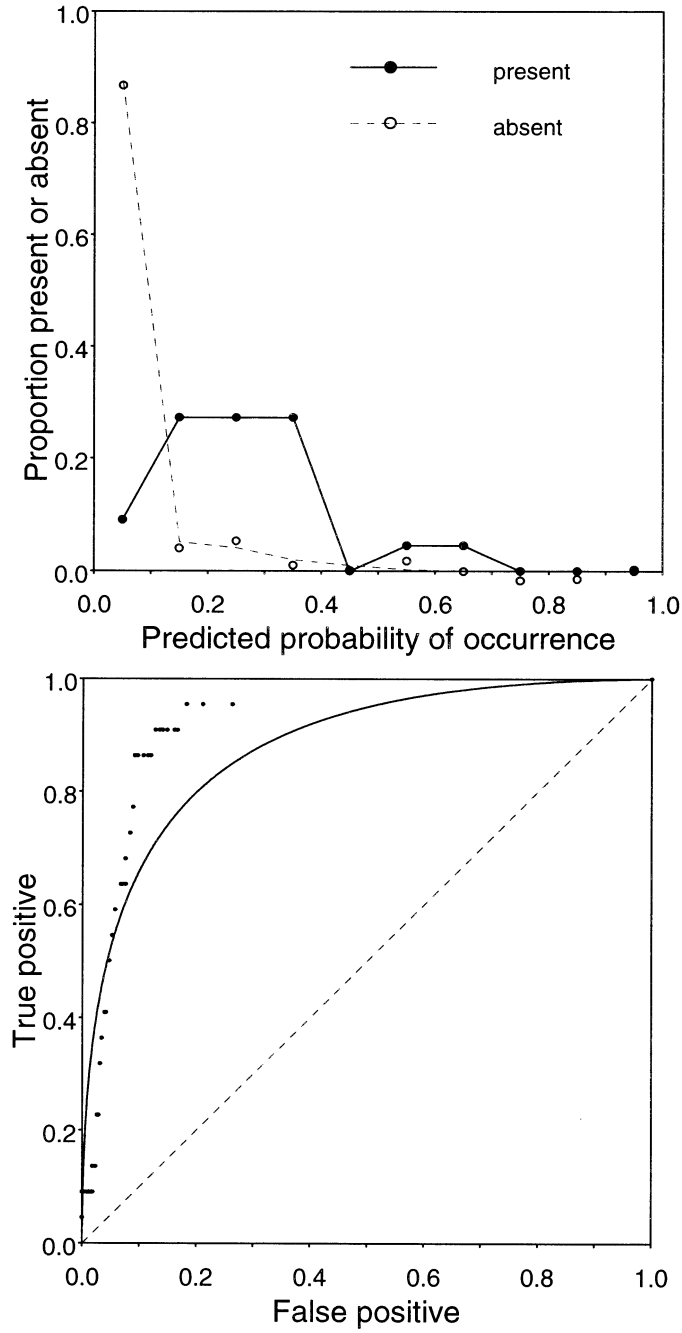

Fig. 8. Discrimination capacity of the distribution model developed for Yellow Box *E. melliodora* in north-east New South Wales. A: Distribution of predicted probability values associated with either occupied sites or unoccupied sites. B: The ROC curve.

recorded are, on average, higher than those for unoccupied sites, suggesting the model has good discrimination ability. The refinement of values predicted by the model is also good, with predictions ranging from 0 to approximately 0.7.

To examine the discrimination performance of the model over a range of threshold levels, the proportion of evaluation sites correctly predicted to be occupied (sensitivity or true positive rate), and the proportion of sites incorrectly predicted to be occupied (false positive rate) were calculated for 70 threshold values spread evenly across the range of available predicted values (from 0.0 to 0.7). These sensitivity and false positive values were then plotted against each other, and a smooth curve drawn through the points (using the Gaussian assumption), to produce the ROC curve shown in Fig. 8b.

To obtain a summary measure of discrimination capacity, the area under the ROC curve was calculated using the non-parametric approach based on the Mann–Whitney statistic. The non-parametric approach was employed because Fig. 8b indicates that the Gaussian assumption does not provide a good fit for this species, as the fitted curve underestimates the area under the observed ROC curve. The standard error of the Mann–Whitney statistic was calculated using a bootstrap sample of 200. The discrimination capacity of the model was calculated to be 0.916 ± 0.033 indicating that the model can correctly discriminate between occupied and unoccupied sites 91.6% of the time. In other words, if a pair of evaluation sites (one occupied and the other unoccupied) is chosen at random, then there is a 0.916 probability that the model will predict a higher likelihood of occurrence for the occupied site than for the unoccupied site.

4. Measuring model calibration

While discrimination deals with the ability of a model to distinguish between occupied and unoccupied sites, calibration describes the agreement between observations and predicted values (or goodness-of-fit), and therefore describes the reliability with which a model predicts the probability of a site being occupied. The deviance has traditionally been used to evaluate logistic regression models by calculating the amount by which a model reduces null deviance (or the percentage of null deviance explained by the model). The deviance, however, in uninformative about the goodness-of-fit of a model developed from binary data as it depends only on the fitted probabilities (Collett, 1991 p. 64). However, we can examine the agreement between observations and prediction to examine how and why the predictions depart from the observations. As described earlier (Fig. 1), calibration, can be separated into two measurable components, bias and spread, and a third component, unexplained error.

These components of calibration can be best understood by re-examining Fig. 3. If a regression line is fitted to the logits of the predicted probabilities and the observed proportions, then bias and spread can be thought of as the intercept and slope, respectively, of this line. The regression line will be a straight line if both the predicted and observed axes are transformed to a logit scale, but will be a curved logistic line when plotted against untransformed probability axes as shown in Fig. 3. Perfect calibration is represented by a 45° line (the dotted line in Fig. 3). Bias describes a consistent overestimate or underestimate of the probability of occurrence of a species resulting in a consistent upward or downward shift in the regression line across the entire predicted probability range. In other words, the intercept of the regression line is either too high or too low. It occurs because the prevalence of the species in the evaluation data is higher or lower than that in the original model development data due, for example, to the use of different survey techniques, or seasonal variation in the abundance or detectability of the species. Bias may also arise when a model, developed in one region or environment, is applied to another region or environment where the species is more or less prevalent.

Spread describes the systematic departure of the regression line, fitted to the predicted and observed values, from a gradient of 45°. On logit axes, a slope of 1, in the absence of bias, implies the predictions follow the 45° line. A slope greater than 1 indicates that predicted values greater than

0.5 are underestimating the occurrence of the species and that predicted values less than 0.5 are overestimating the occurrence of the species. A gradient between 0 and 1 implies that predicted values less than 0.5 are underestimating the occurrence of the species and that predicted values greater than 0.5 are overestimating the occurrence of the species. A gradient significantly different from 1 indicates misspecification of a model.

The third component of calibration, unexplained error, is due to the variability of individual records around the regression line fitted to predicted and observed values, and describes variation not accounted for by the bias and spread of a model. Some of this variation arises because particular covariate patterns or habitat types were not well represented in the development the model, and may be identified by analysing deviance, bias or spread residuals (Miller et al., 1991). The rest of the unexplained error is due to random variation, the source of which is not identifiable through residual analysis. In wildlife habitat studies this component of the deviance is expected to be quite large due to error in measurement of species presence and environmental predictors, and to factors not included in the model that may influence habitat selection or species distribution (such as competition, predation, or biogeographical barriers to dispersal).

A reliable model (well calibrated) should be able to correctly predict the actual proportion of sites occupied by the species of interest. That is, $E(x/\pi) = \pi$. Cox (1958) formalised an approach to detecting bias and spread by using logistic regression to model the relationship between the logit of the predicted probabilities (π_i) for the evaluation sites and the observed occurrences ($p(x_i = 1)$) at these sites. This modelled relationship takes the form:

$$\ln\left(\frac{p(x_i = 1)}{p(x_i = 0)}\right) = a + b \ln\left(\frac{\pi_i}{1 - \pi_i}\right) \quad (16)$$

The coefficients a and b in this relationship represent bias and spread respectively. In a perfectly calibrated model a will have a value of zero and b will have a value of one.

If $a = 0$, then the observed proportion ($p(x_i = 1)$) at $\pi_i = 0.5$ will be 0.5, regardless of the value of b. If $b > 1$ (Fig. 9(A)) then predicted values less than 0.5 are overestimating the observed proportion of occurrence and predicted values greater than 0.5 are underestimating observed occurrence. If $0 < b < 1$, then the reverse is occurring, with predicted values less than 0.5 underestimating and those above 0.5 overestimating (Fig. 9(B)). If $b < 0$, then the overall trend in predicted probabilities is wrong, with values less than 0.5 being associated with a higher observed proportion of occurrence than those above 0.5.

At a predicted (π) value of 0.5, $\ln[p(x_i = 1)/p(x_i = 0)] = a$, and $a = 0$ implies $p(x_i = 1) = 0.5$. Thus a reflects the overall bias of the model if $b = 1$. However, if $b \neq 1$, then a describes the model bias at $\pi = 0.5$. This occurs because a is a function of b:

$$a = \mu_\pi - b\mu_x \quad (17)$$

The predicted probabilities are generally too low if $a > 0$, and too high if $a < 0$ (Fig. 9(C, D)). Fig. 9(E, F) illustrate the relationship between model predictions and observations if both spread error and bias are present.

Cox (1958) devised score tests to evaluate the bias and spread of binary regression models. Miller et al. (1991) provide equivalent tests based on likelihood ratio statistics that are more accurate and easier to apply. These tests will be used here.

The first of these likelihood ratio tests examines the hypothesis that $a = 0$, $b = 1$ and is calculated as the difference between two deviance values. The first deviance value, $D(0,1)$, is simply the deviance of observations at the evaluation sites in relation to raw predictions from the original model. This is equivalent to forcing $a = 0$ and $b = 1$ in Eq. (16), i.e. by assuming that the model has perfect calibration. The second deviance value, $D(a,b)$, is the deviance of observations in relation to the adjusted predictions, obtained by applying Eq. (16) using fitted values for a and b. In other words, these values are used to correct predictions for bias and spread. The test evaluates the significance of this correction by comparing the difference between the two deviance values:

$$D(0,1) - D(a,b) \quad (18)$$

to a χ^2 distribution with two degrees of freedom.

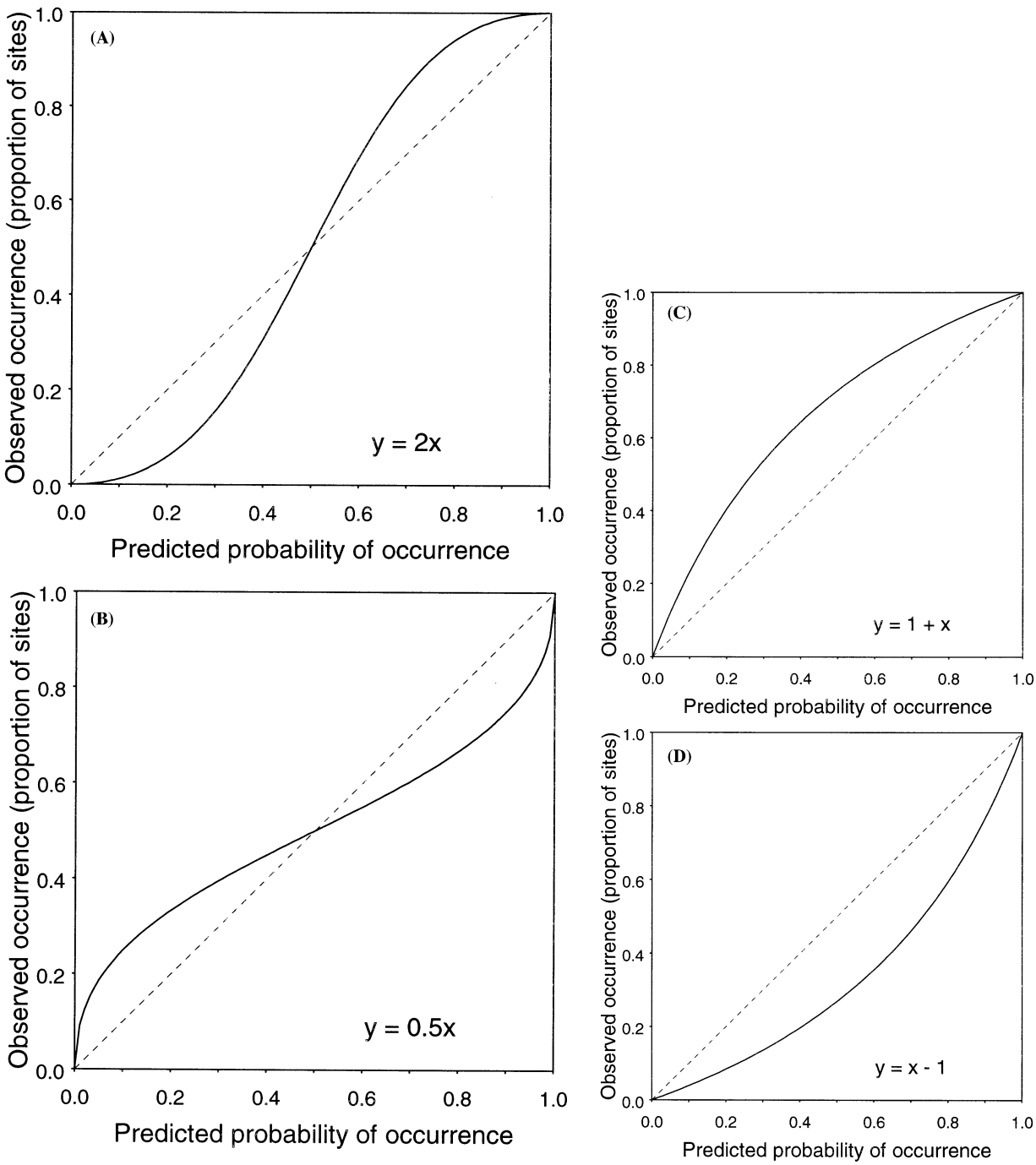

Fig. 9. Calibration diagrams showing effect of bias and spread on the agreement between model predictions and observations. (A) No bias with spread error (gradient greater than 1); (B), no bias with spread error (gradient less than 1); (C) positive bias with correct spread; (D) negative bias with correct spread; (E) positive bias with spread error (gradient greater than 1); (F) negative bias with spread error (gradient greater than 1).

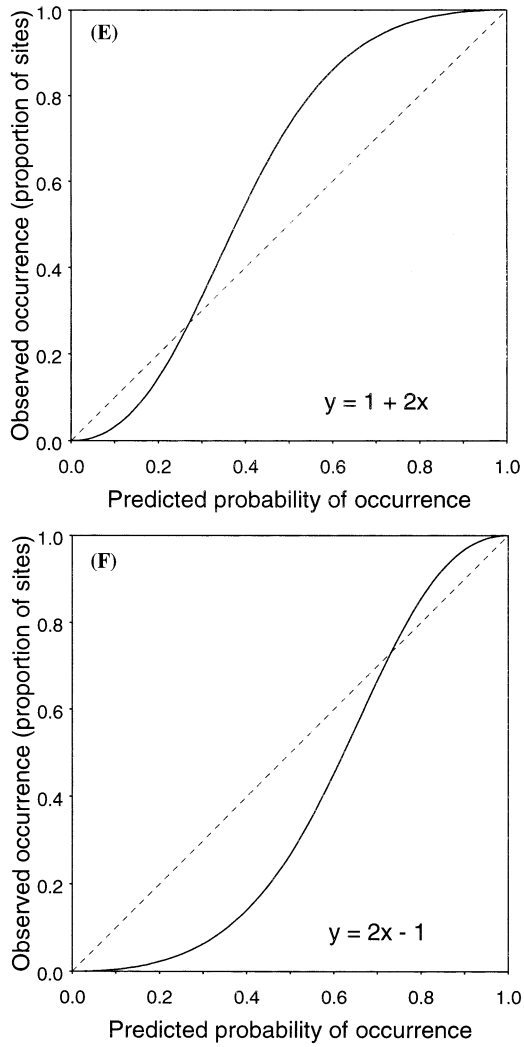

Fig. 9. (*Continued*)

The second likelihood ratio test examines the hypothesis that $a = 0/b = 1$, and is a test for bias given appropriate spread. The following test statistic is compared to a χ^2 distribution with one degree of freedom:

$$D(0,1) - D(a,1) \tag{19}$$

where $D(a,1)$ is the deviance of observations in relation to predictions adjusted by applying Eq. (16) using a fitted value for a but forcing b to equal 1.

The third likelihood ratio test examines the hypothesis that $b = 1/a$ and is a test for incorrect spread given no bias. The following test statistic is again compared to a χ^2 distribution with one degree of freedom:

$$D(a,1) - D(a,b) \tag{20}$$

The following example illustrates the evaluation of model calibration.

4.1. Example 2

As described in Example 1 the distribution of Yellow Box *E. melliodora* has been modelled within north-east New South Wales as a function of several regional scale environmental variables using field survey data from 2223 sites. The occurrence of the species has also been surveyed at an additional 407 sites within the region. These additional data are now used to evaluate the reliability of the model's predictions in terms of calibration, bias and spread.

The overall agreement between model predictions and the observed occurrence of Yellow Box within the evaluation data was examined graphically. This was undertaken by breaking the predicted probability range into ten classes, and calculating the proportion of occupied sites within each class. The proportion of occupied sites, and the median predicted probability for each class were then plotted against each other, and a smooth curve drawn through the points, as shown in Fig. 3. If there were perfect agreement between predicted probabilities from the model and the observed occurrence of the species, the points in this graph would lie along the 45° line. Significant systematic departure of the points from this 45° line indicates the existence of calibration bias and/or spread error. Examination of the curve in Fig. 3 suggests that the model may lack calibration due to both bias and incorrect spread within the predictions.

To determine whether this lack of calibration is statistically significant, the observed data were modelled as a function of the logit of the predicted values to produce the following logistic regression model:

$$\ln\left(\frac{\mathrm{p}(x_i=1)}{\mathrm{p}(x_i=0)}\right) = -0.3987 + 0.6659\ln\left(\frac{\pi_i}{1-\pi_i}\right) \quad (21)$$

This model was used to fit the smooth curve to the points in Fig. 3. If the observations agree with the predictions, then the intercept and gradient of this regression line should not differ significantly from 0 and 1, respectively.

To determine whether the a and b coefficients in the above model differ significantly from 0 and 1, thereby suggesting significant bias and/or spread error, the likelihood ratio tests reported by Miller et al. (1991) were applied. $D(a,b)$, the deviance of observed values in relation to the logistic regression model presented in Eq. (21) was calculated to be 112.451 and the Null deviance from this model $D(0,1)$ as 115.697 with 405 degrees of freedom. Therefore, using Eq. (18)

$$D(0,1) - D(a,b) = 115.697 - 112.451 = 3.246 \quad (22)$$

This value was not significant when compared to a χ^2 distribution with two degrees of freedom. Therefore, the a and b coefficients do not differ significantly from 0 and 1, respectively. Consequently, the curve in Fig. 3 does not depart significantly from the 45° line and the model accurately predicts the observed probability of occurrence of Yellow Box at the evaluation sites. If this test had rejected the hypothesis that $a=0$, $b=1$ then the alternative hypothesis that $a=0/b=1$ (bias given appropriate spread) or $b=1/a$ (incorrect spread in the absence of bias) could have been evaluated further using the likelihood ratio tests presented in Eqs. (19) and (20).

5. Comparing the performance of two or more models

It is often necessary to compare the performance of two or more models for a species, developed using different explanatory variables or different modelling techniques, in order to choose the most appropriate model for a given application. The reliability of predictions generated by two or more models can be compared by examining differences in residual deviance, $D(0,1)$, of evaluation data in relation to predictions, and by comparing levels of prediction error in terms of calibration bias and spread. The model providing the most reliable predictions, is that with the smallest residual deviance and the least error in prediction spread and bias.

The difference between the area under two ROC curves generated by two or more models provides a measure of comparative discrimination capacity of these models when applied to indepen-

dent evaluation data. The best model is that generating the largest ROC area. However, it is important to note that, if two ROC curves intersect, then neither model performs consistently better than the other across the entire range of decision thresholds (Brownie et al., 1986). This information needs to be considered when interpreting differences in the summary index of average discrimination performance.

The significance of the difference between the areas under two ROC curves (A_1 and A_2) generated using independent data can be calculated as a critical ratio test:

$$Z = \frac{A_1 - A_2}{\sqrt{\mathrm{SE}_{A_1}^2 + \mathrm{SE}_{A_2}^2}} \quad (23)$$

The value of Z is then tested against a table of areas under the normal probability density function. If this value is significant at a specified probability threshold this is evidence that the true ROC areas are different.

However, if the two ROC curves are generated using the same evaluation data, the two areas are likely to be positively correlated and thus estimation of the standard error of the difference in area under the two curves, the denominator in Eq. (23), needs to account for this correlation. Hanley and McNeil (1983) have developed the following equation to incorporate r, the correlation between predictions for the two models:

$$\mathrm{SE}(A_1 - A_2) = \sqrt{\mathrm{SE}_{A_1}^2 + \mathrm{SE}_{A_2}^2 - 2r\mathrm{SE}_{A_1}\mathrm{SE}_{A_2}} \quad (24)$$

The right side of this equation replaces the denominator in the critical ratio test in Eq. (23).

To calculate r, it is necessary to determine two intermediate correlation coefficients, the correlation between predictions from the two models for the positive events and the correlation between predictions for the negative events. These are calculated independently using either the Pearson product moment correlation coefficient or the Spearman Rank correlation coefficient. The average of these two correlation values, and the average ROC area of the two models, are then compared to the table provided by Hanley and McNeil (1983) to determine r. (The average correlation value can be used to approximate r if the table is not available.)

When comparing two ROC curves it is important to consider not only the area under each curve, as discussed, but also the shape of the curves and whether they intersect. The area under the curve provides a summary measure of model discrimination accuracy. Consequently, the ROC curve with the larger area is, on average, more accurate. However, the shape of the ROC curve is also important, and describes the trade-off between true positive and false positive rates as the threshold probability is altered. Therefore, two ROC curves that intersect provide different levels of accuracy depending on which threshold probability level is selected. This trade-off may affect the choice of a model for a given application.

The following example demonstrates the analysis of the difference in discrimination capacity between two models derived from the same evaluation data.

5.1. Example 3

The distribution of the small reptile *Calyptotis scutirostrum* in north-east New South Wales has been modelled using presence/absence data from 836 sites as a function of a number of climatic, topographical and contextual variables (NSW NPWS, 1994a). The discrimination capacity of this distributional model was determined by calculating the area under an ROC curve developed using data from 125 independent evaluation sites within the region (NSW NPWS, 1995b; Clode and Burgman, 1997). One hundred threshold probabilities were used to calculate the curve. The model was found to have good discrimination ability, with an ROC area of 0.811 ± 0.038.

In an attempt to further improve the discrimination capacity of the model, microhabitat variables measured at each of the 125 evaluation sites were added to the model. To evaluate this refined model, a jackknife procedure was employed to calculate independent predicted probability values for each of the 125 validation sites. These values were then used to develop a second ROC curve. The discrimination capacity of this refined model was calculated to be 0.888 ± 0.028.

To determine whether the addition of microhabitat variables provides a significant improvement in discrimination capacity over the original model, the significance of this improvement was calculated using a critical ratio test, modified to control for correlation between the two discrimination indices. This correlation exists because the same 125 evaluation sites were used to calculate the area under the ROC curve for both predictive models.

Following the procedure of Hanley and McNeil (1983), the degree of correlation between the two ROC area measurements was calculated to be 0.982 using the many-ties version of the Spearman Rank correlation test (Conover, 1980). This value was calculated from the mean of the correlation between the predictions for occupied sites for each model ($r = 0.979$), and the predictions for unoccupied sites for each model ($r = 0.984$). The average correlation value of 0.982 was incorporated into the calculation of a critical ratio test statistic, using Eqs. (23) and (24):

$$Z = \frac{A_1 - A_2}{\sqrt{\mathrm{SE}_{A_1}^2 + \mathrm{SE}_{A_2}^2 - 2r\mathrm{SE}_{A_1}\mathrm{SE}_{A_2}}} = \frac{0.888 - 0.811}{\sqrt{0.038^2 + 0.028^2 - 2 \times 0.982 \times 0.038 \times 0.028}} = 6.55 \quad (25)$$

The resulting Z-value was significant at the 1% level when compared to the normal distribution. Therefore, adding the microhabitat information to the predictive model significantly improved the ability of the model to discriminate between occupied and unoccupied sites.

6. Conclusion

Wider application of the techniques described in this paper could improve understanding of the usefulness, and potential limitations, of habitat models developed for use in conservation planning and wildlife management. Evaluation of predictive performance can also assist in determining the suitability of a model for specific applications.

There are three main ways in which predictions of species occurrence derived from logistic regression models may be used. First, predictions can be used as an absolute estimate of the probability of a species occurring at a site. That is, the predicted probabilities are used at face value. Second, predictions can be used merely as a relative index of likelihood of occurrence, where higher values indicate sites more likely to be occupied by the species. Third, predicted probabilities can be converted to predicted presence/absence by applying a threshold to the predicted probability range. Each of these uses requires knowledge of different components of predictive performance.

If the predictions are to be used at face value, for example to estimate the total population size for a species by predicting the probability of the species occurring at a large set of sites within a region, then knowledge of model calibration, bias and spread is essential to interpret the predicted probability values. It is important to know something about the expected magnitude and nature of differences between predicted probabilities and observed proportions of occurrence.

However, in many applications of logistic regression models, exact estimation of the probability of species occurrence is not required. All that is required is an index of relative suitability of sites within a region so that areas may be ranked according to their importance as habitat or their likelihood of containing the species of interest. Examples of these applications include maps of relative habitat suitability to aid species management, or the ranking of priority areas within a region for selection of conservation reserves. This type of application requires knowledge of the degree to which higher predicted probabilities are associated with the presence of a species. The discrimination index provides a summary measure of this capability. A graph relating observed occurrence to predicted probability, such as that shown in Fig. 3, can also provide useful information on the rank order relationship between predictions and observations.

An understanding of discrimination capacity is particularly important if a model is to be used to delineate areas predicted to be occupied, or to contain suitable habitat, from areas predicted to be unoccupied, or to contain unsuitable habitat. An understanding of model calibration is also

important in this case to inform selection of an appropriate threshold probability to distinguish sites predicted to be occupied from sites predicted to be unoccupied.

Most studies utilising the threshold probability approach (Lindenmayer et al., 1990; Pearce et al., 1994) have assumed that the selection of a threshold, such as 0.5, implies that sites with a probability greater than 0.5 will be occupied greater than 50% of the time. However, if a model is not well calibrated and has significant bias or spread error, then a predicted value of 0.5 will not relate to an observed value of 0.5, but a higher or lower observed rate. Therefore, choosing a threshold probability without any information on bias and spread will greatly reduce the confidence that can be placed in any map that is produced. If predictions from a model tend to overestimate the occurrence of a species, then the predicted distribution will include not only areas of suitable habitat (true positive sites) but also substantial areas of unsuitable habitat (false positive sites). However, if a model tends to underestimate the occurrence of a species, then areas of potentially occupied habitat will remain unidentified (false negative sites). For a rare species this may have devastating ramifications because some populations may fail to be identified and protected. The ROC curve can also provide useful information for selecting an appropriate probability threshold by describing the trade-off between correctly predicting the occurrence of a species (true positive) and incorrectly predicting the presence of the species (false positive).

The evaluation techniques described in this paper can also be used to identify aspects of a model most in need of improvement. Each of the statistics provides information on possible reasons for a lack of agreement between predictions and observations. Calibration bias suggests that the prevalence of the species in the model development data and the evaluation data are different, and can arise because of differences in methodology between the surveys that collected these data sets (e.g. different detection techniques, different seasons), or differences between the regions or environments covered by these surveys. It is important to develop models using data that are representative of the situations in which a model is to be applied. Lack of model refinement and the presence of spread error suggest that important explanatory variables are missing from the model, that non-discriminatory variables have been included in the model, or that weighting of variables within the model places too much emphasis on variables that are only weakly related to the occurrence of the species in the validation data. Lack of discrimination ability usually arises because the explanatory variables in the model are not strongly associated with the presence of the species. Fielding and Bell (1997) provide a detailed discussion of potential sources of classification error in ecological applications, and their effect on the classification ability of a model.

The model evaluation techniques described in this paper can play an important role in the development and application of models for conservation planning and wildlife management. Such evaluation needs to be included routinely as part of the model development process. Models cannot be applied confidently without knowledge of their predictive accuracy and the nature and source of prediction errors.

Acknowledgements

The work described in this paper was performed as part of two consultancies funded by the Australian Nature Conservation Agency and Environment Australia. We thank Andrew Taplin and Dave Barrett from these agencies for their support and encouragement. We also thank David Meagher for commenting on an early draft of the manuscript.

References

Austin, M.P., Nicholls, A.O., Margules, C.R., 1990. Measurement of the realised qualitative niche: environmental niches of five *Eucalyptus* species. Ecol. Monogr. 60, 161–177.

Bambar, D., 1975. The area above the ordinal dominance graph and the area below the receiver operating graph. J. Math. Psych. 12, 387–415.

Brownie, C., Habicht, J., Cogill, B., 1986. Comparing indicators of health or nutritional status. Am. J. Epidemiol. 124, 1031–1044.

Clode, D., Burgman, M., 1997. Joint Old Growth Forests Project: Summary Report. NSW National Parks and Wildlife Service and State Forests of NSW, Sydney.

Collett, D., 1991. Modelling Binary Data. Chapman and Hall, London.

Conover, W.J., 1980. Practical Nonparametric Statistics, 2nd ed. Wiley, New York.

Cox, D.R., 1958. Two further applications of a model for binary regression. Biometrika 45, 562–565.

Diffenbach, D.R., Owen, R.B., 1989. A model of habitat use by breeding American Black Ducks. J. Wildl. Manage. 53, 383–389.

Efron, B., 1982. The jackknife, the bootstrap, and other resampling plans. Volume 38 of CBMS-NSF Regional conference series in applied mathematics. SIAM.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman and Hall, New York.

Ferrier, S., 1991. Computer-based extension of forest fauna survey data: current issues, problems and directions. In: Lunney, D. (Ed.), Conservation of Australia's Forest Fauna. Royal Zoological Society of New South Wales, Sydney, pp. 221–227.

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

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.

Hanley, J.A., McNeil, B.J., 1983. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148, 839–843.

Hastie, T.J., Tibshirani, R., 1990. Generalized Additive Models. Chapman and Hall, London.

Hilden, J., 1991. The area under the ROC curve and its competitors. Med. Dec. Making 11, 95–101.

Lindenmayer, D.B., Cunningham, R.B., Tanton, M.T., Smith, A.P., Nix, H.A., 1990. Habitat requirements of the mountain brushtail possum and the greater glider in the montane ash-type forests of the central highlands of Victoria. Aust. Wildl. Res. 17, 467–478.

Lindenmayer, D.B., Cunningham, R.B., Tanton, M.T., Nix, H.A., Smith, A.P., 1991. The conservation of arboreal marsupials in the montane ash forests of the Central Highlands of Victoria, south-east Australia. III. The habitat requirements of Leadbeater's Possum, *Gymnobelideus leadbeateri* and models of the diversity and abundance of arboreal marsupials. Biol. Cons. 56, 295–315.

McCullogh, P., Nelder, J.A., 1989. Generalized Linear Models, 2nd ed. Chapman and Hall, London.

Metz, C.E., 1978. Basic principles of ROC analysis. Sem. Nuclear Med. 8, 283–298.

Metz, C.E., 1986. ROC methodology in radiologic imaging. Invest. Rad. 21, 720–733.

Miller, M.E., Hui, S.L., Tierney, W.M., 1991. Validation techniques for logistic regression models. Stat. Med. 10, 1213–1226.

Mills, L.S., Fredrickson, R.J., Moorhead, B.B., 1993. Characteristics of old-growth forests associated with northern spotted owls in Olympic National Park. J. Wildl. Manage. 57, 315–321.

Mladenoff, D.J., Sickley, T.A., Wydeven, A.P., 1999. Predicting gray wolf landscape recolonisation: logistic regression models vs. new field data. Ecol. Appl. 9, 37–44.

Murphy, A.H., Winkler, R.L., 1987. A general framework for forecast verification. Monthly Weather Rev. 115, 1330–1338.

Murphy, A.H., Winkler, R.L., 1992. Diagnostic verification of probability forecasts. Int. J. Forecasting 7, 435–455.

Murtaugh, P.A., 1996. The statistical evaluation of ecological indicators. Ecol. Appl. 6, 132–139.

NSW NPWS, 1994a. Fauna of north-east NSW forests, North East Forests Biodiversity Study Report No. 3. NSW National Parks and Wildlife Service, Sydney, New South Wales, Australia.

NSW NPWS, 1994b. Flora of north-east NSW forests, North East Forests Biodiversity Study Report No. 4. NSW National Parks and Wildlife Service, Sydney, New South Wales, Australia.

NSW NPWS, 1995a. Vegetation Survey and Mapping of Upper North-eastern New South Wales. NSW National Parks and Wildlife Service, Sydney, New South Wales, Australia.

NSW NPWS, 1995b. Vertebrates of Upper North East New South Wales. NSW National Parks and Wildlife Service, Sydney, New South Wales, Australia.

Osborne, P.E., Tigar, B.J., 1992. Interpreting bird atlas data using logistic models: an example from Lesotho, Southern Africa. J. Appl. Ecol. 29, 55–62.

Pearce, J.L., Burgman, M.A., Franklin, D.C., 1994. Habitat selection by helmeted honeyeaters. Wildl. Res. 21, 53–63.

Reckhow, K.H., Black, R.W., Stockton, T.B. Jr, Vogt, J.D., Wood, J.G., 1987. Empirical models of fish response to lake acidification. Can. J. Fish. Aquat. Sci. 44, 1432–1442.

Sokal, R.R., Rohlf, F.J., 1981. Biometry, 2nd ed. W.H. Freeman, New York.

Stone, M., 1974. Cross-validation choice and assessment of statistical predictions. J. R. Stat. Soc. B 36, 111–147.

Straw, J.A. Jr, Wakely, J.S., Hudgins, J.E., 1986. A model for management of diurnal habitat for American woodcock in Pennsylvania. J. Wildl. Manage. 50, 378–383.

Swets, J.A., 1986a. Indices of discrimination or diagnostic accuracy: their ROCs and implied models. Psych. Bull. 99, 100–117.

Swets, J.A., 1986b. Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. Psych. Bull. 99, 181–198.

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

Swets, J.A., Pickett, R.M., 1982. Evaluation of Diagnostic Systems. Academic Press, New York.