Environmental Conservation 24 (1): 38–49 © 1997 Foundation for Environmental Conservation

A review of methods for the assessment of prediction errors in conservation presence/absence models

ALAN H. FIELDING[1*] AND JOHN F. BELL[2]
[1]*Department of Biological Sciences, the Manchester Metropolitan University, Manchester M1 5GD, UK and* [2]*University of Cambridge Local Examinations Syndicate, University of Cambridge, Cambridge, UK*
Date submitted: 18 December 1996 Date accepted: 1 March 1997

Summary

Predicting the distribution of endangered species from habitat data is frequently perceived to be a useful technique. Models that predict the presence or absence of a species are normally judged by the number of prediction errors. These may be of two types: false positives and false negatives. Many of the prediction errors can be traced to ecological processes such as unsaturated habitat and species interactions. Consequently, if prediction errors are not placed in an ecological context the results of the model may be misleading. The simplest, and most widely used, measure of prediction accuracy is the number of correctly classified cases. There are other measures of prediction success that may be more appropriate. Strategies for assessing the causes and costs of these errors are discussed. A range of techniques for measuring error in presence/absence models, including some that are seldom used by ecologists (e.g. ROC plots and cost matrices), are described. A new approach to estimating prediction error, which is based on the spatial characteristics of the errors, is proposed. Thirteen recommendations are made to enable the objective selection of an error assessment technique for ecological presence/absence models.

Keywords: habitat relationships, data partitioning, habitat-association model, classifier, threshold, validation

Introduction

The habitat-association approach to ecology has been used for a variety of purposes, including conservation and ecological management. In particular the approach has been used to develop predictive models for estimating population sizes and geographical ranges and for identifying the potential impacts of habitat changes (e.g. Stillman & Brown 1994). A recent conference on bird conservation recognized that habitat-based ecological studies of individual species should not be devalued since most successful conservation has been based on such studies (Anon. 1995). The conference also noted that the urgent need for conservation advice outstripped the resources available for ecologists and, as a consequence, the successful development of modelling techniques could pay great dividends.

Any approach to ecological modelling has little merit if the predictions cannot be, or are not, assessed for their accuracy using independent data (Verbyla & Litaitis 1989). Most habitat-association studies use a very restricted set of error measures, of which percentage overall accuracy is the most common (e.g., Brennan *et al*. 1986; Capen *et al*. 1986; Verbyla & Litvaitis 1989; Donázar *et al*. 1993). Because little attention appears to have been paid to the assessment of error in this type of model our aims are to review the nature of prediction errors and, subsequently, to evaluate a range of techniques that may be used to assess and compare prediction success. Part of the justification for focusing on error assessment is Chatfield's (1995) suggestion that it may be advantageous to adopt a pragmatic approach to model building in which we concentrate on the model's accuracy and usefulness, rather than testing the statistical validity of the model.

We do not consider, directly, the methods that may be used to study the habitat associations. All of them share a common approach, with data consisting of a set of positive and negative cases (stations) for which a range of habitat variables have been recorded. These data are subjected to analysis by a particular algorithm or classifier (e.g. discriminant analysis, logistic regression, decision trees and artificial neural networks) so that a 'rule' is obtained which is capable of correctly classifying cases as positive (where a species is present) or negative (where a species is absent). The usefulness of this rule is generally assessed by examining how many of the cases are predicted correctly. Some of these methods are reviewed by Manly *et al*. (1992) and Morrison *et al*. (1992). Although most of the illustrative examples refer to bird studies our observations and recommendations are more general.

Nature and measurement of prediction errors

Types of error

In a presence/absence model there are two possible prediction errors: false positives (FP) and false negatives (FN). The performance of a presence/absence model is normally sum-

* Correspondence: Dr Alan Fielding Tel: 144 161 247 1198 Fax: 144 161 247 6325 email: A.Fielding@mmu.ac.uk

		Actual	
		+	-
Predicted	+	a	b
	-	c	d

Figure 1 A confusion matrix.

marized in a confusion or error matrix (Fig. 1) that cross-tabulates the observed and predicted presence/absence patterns. Morrison *et al.* (1992) refer to FP errors as type I and FN errors as type II. Data in the confusion matrix are sometimes presented as percentages rather than counts.

Data partitioning

It is generally accepted that robust measures of prediction success make use of independent data, i.e. data not used to develop the prediction model. Table 1 describes some strategies used to obtain testing data. We refer to the two data sets needed to develop and test predictions as 'training' and 'testing' data. A variety of synonyms are used by other workers, e.g. learning and validation data. It is common practice to split or partition the available data to provide the training and testing data. Chatfield (1995) questioned the use of data partitioning for model testing, suggesting that splitting data arbitrarily is not the same as collecting new data ('prospective sampling' in Table 1). He also questioned the validity of a 'hold-out' sample (see Table 1) to choose a 'best' model. The best model, e.g. deciding which set of predictor variables to include, is normally based on some measurement of error. The ecological literature seems to have paid little attention to how the partitioning method can influence the error rates. Verbyla & Litaitis (1989) briefly reviewed a range of partitioning methods in their assessment of resampling methods for evaluating classification accuracy.

When the performance of a classifier is assessed we do not obtain its actual confusion matrix, rather it is estimated from an apparent confusion matrix based on the testing set(s) (Blayo *et al.* 1995). This is analogous to the relationship between a parameter and its statistic. Unfortunately the number of available test sets is finite and frequently small. A classifier that uses all of the available data will, on average, perform better than a classifier based on a subset. Consequently if data are partitioned the size of training set must decrease and this can reduce model accuracy. Conversely, larger test sets reduce the variance of the error estimates. There is, therefore, a trade-off between having a large test set that gives a good assessment of the classifier's performance and a small training set which is likely to result in a poor classifier. Rencher (1995) suggests that while partitioning should be used for model validation, all available data should be used to develop the eventual classification rule.

The resubstitution method (Table 1) tends to give optimistically-biased estimates of error rates because of overfitting and a loss of generality (Chatfield 1995). The resubstitution approach provides a lower boundary for the error probabilities of a particular classifier. All partitioning methods reduce the size of the training set resulting in overestimates of the actual error rates. It is possible to average the results from several partitions of the data [k-fold partitioning where $2<k<(n-1)$, Table 1] and thus make the accuracy estimate less dependent on a single partition. The *L-O-O* method and the statistically-equivalent bootstrap (Table 1) method gives an upper boundary for the error probabilities.

Table 1 Data partitioning methods for the allocation of cases to training and testing data sets.

Method	*Examples*	*Notes*
Resubstitution	Stockwell (1992) Osborne & Tigar (1992)	No partitioning is carried out, the same data are used for training and testing. This tends to provide optimistic measures of prediction success.
Bootstrapping	Buckland & Elston (1993) Verbyla & Litaitis (1989)	Bootstrap samples (sampling with replacement) are used to assess prediction success. Accuracy is usually reported as a mean and confidence limits.
Randomization	Capen *et al.* (1986)	Random samples are obtained by sampling without replacement. Accuracy is usually reported as a mean and confidence limits.
Prospective sampling	Capen *et al*. (1986) Fielding & Haworth (1995) Morrison *et al*. (1987)	A new sample of cases is obtained after the model has been developed. These could be from a different region or time.
k-fold partitioning	Stockwell (1992)	The data are split into k ($k > 2$) sets, only one of which is used for training. The remaining $k - 1$ sets are pooled for testing purposes. Also known as the hold-out or external method. Accuracy is usually reported as a mean and confidence limits.
Special cases of k-fold partitioning		
Leave-One-Out (L-O-O)	Capen *et al*. (1986) Osborne & Tigar (1992)	Also known as jackknife sampling, n samples of 1 case are tested sequentially, the remaining $n - 1$ cases forming the training set.
$K = 2$	Smith (1994)	Data are split into one training set and one testing set. A variety of strategies may be employed to determine the split.

https://doi.org/10.1017/S0376892997000088 Published online by Cambridge University Press

Thus, the true performance of a classifier lies somewhere between these upper and lower boundaries.

Huberty (1994) provided a heuristic ('rule of thumb') for determining the ratio of training to testing cases that is based on the work of Schaafsma and van Vark (1979). This heuristic, which is restricted to presence/absence models, suggests a ratio of $[1 + (p - 1)^{1/2}]^{-1}$, where p is the number of predictors. This approximates to a training set consisting of 75% of the cases when $p > 10$.

Origin of prediction errors

Factors that lead to prediction errors can be placed into two broad categories. The first category ('algorithmic' errors) comprises limitations imposed by the classification algorithm and the data-gathering process. The second category ('biotic' errors) comprises processes arising directly from the organism's ecology. These biotic errors arise because not all of the ecologically-relevant processes have been specified in the model. Unfortunately relevant data are often inaccessible.

The greatest difficulty that ecological processes can create for classifiers is that some of the negative locations may be similar, and possibly identical, to positive locations. This will degrade the performance of the classifier and/or result in too many false positives. There are a variety of ecological processes, operating over a range of timescales, that can give rise to data of this type. For example, in conservation-based studies it is almost inevitable that the species will be restricted to few locations, thus only a small proportion of the potentially positive cases will be occupied. This is a particular example of a more general problem of unsaturated populations. Capen *et al.* (1986) noted that there is an implicit assumption in most presence/absence designs that breeding habitats are saturated. They suggest that such an assumption may be unjustified. If a habitat is unsaturated there will be negative cases that have the potential to become positive should the population expand. The classifier will probably falsely predict some of these as positive cases (given the current population size). Newton (1979) provided a good example of this when he demonstrated how in the UK Peregrine, *Falco peregrinus*, recovering populations have twice re-occupied traditional nesting crags. Thus some traditional, but currently unused, crags could have been labelled incorrectly as positive cases by a predictive model.

Intra- and interspecific interference create similar difficulties for classifiers. Austin & Gaywood (1994) and Austin *et al.* (1994) suggested that unimodal distributions, upon which canonical correspondence analysis (CCA) is predicated, may be inappropriate for some species since the curve's shape is influenced by competition. For example, Ratcliffe (1993, p. 308) described how Peregrine were excluded from cliffs in SW Scotland following the re-establishment of Golden Eagle *Aquila chrysaetos* in 1945. Subsequently these cliffs were abandoned by the Eagles and Peregrines returned to breed. Thus, depending on the status of the Golden Eagle a particular cliff could be classified as either positive or negative for Peregrine. This scenario appears to be common amongst birds of prey (Solonen 1993). Interference is not the only interaction that can result in absences from apparently suitable habitat. Reed & Dobson (1993) reviewed the importance of conspecific attraction to the selection of breeding sites. One solution to this type of problem would be to incorporate interspecific information into the classifier. Although this is possible for some applications it is difficult to see how this could be achieved if we do not have the necessary ecological data. Few papers refer to this problem, although Capen *et al.* (1986) developed their models with an explicit assumption that interspecific competition did not preclude nesting.

One of the difficulties with incorporating biotic interactions into a classifier is that interactions must have a scale context. It is well known that bird-habitat associations can have a marked scale dependence (Wiens *et al*. 1987). At the extremes no two individuals can occupy the same space and yet all coexist on the planet. Thus, the size of the sampling unit will determine whether we identify the relationship between individuals as interference or coexistence. The scale problem is symptomatic of a more general problem. The underlying theory of most classifiers assumes that the cases are discrete and unambiguous entities. This assumption will be violated when cases are arbitrarily-defined units of habitat.

Intraspecific competition, in particular territoriality, will result in a minimum distance between positive cases. If the between-positive distance is greater than the sample unit size this can result in the appearance of false positive cases adjacent to the true positive. This effect is mainly a scale problem confounded by spatial autocorrelation (see below) that could be reduced if the sample unit size matched the territory size, but there are difficulties with this. Firstly, sampling units would need to be correctly orientated with respect to the territories. Secondly, territory size is not fixed, it can be dependent on season, age and gender (e.g. Schwede *et al.* 1993; Cederlund & Sand 1994). Thirdly, territories have seldom approximated to simple and constant geometrical shapes (e.g. Weir & Picozzi 1983). Consequently we will always need to accept some compromise value for the sampling unit size and shape.

When field data are obtained the utilized habitat ('realized niche') will be influenced by a variety of spatially and temporally dynamic processes. This has significance for predictive models because the habitat utilization will vary between individuals. Aebischer *et al.* (1993) said that pooling data across radio-tracked individuals, to identify habitat utilization patterns, was justifiable only if it could be shown that the animals do not differ. Variability between sample cases is expected, and should not create difficulties for most classifiers if the variability is random and stationary with respect to location. However, Hengveld (1994) has suggested that some ecological processes are non-stationary at regional or geographical scales. Stauffer & Best (1986) concluded that their data indicated different habitat-selection patterns within a relatively small geographical area. Schooley (1994) demonstrated that failure to take account of annual variation in habitat selection could lead to misleading inferences about habitat

associations. Rotenberry's (1986) regression models failed to predict consistently the responses of birds to habitat projected through time, space or perturbation. Rotenberry (1986) suggested that this should not happen with presence/absence models since they need to sample over a larger environmental scale. However, Fielding & Haworth (1995), using a variety of presence/absence approaches, found great variability in the habitat-associations of Golden Eagle, Raven *Corvus corax* and Buzzard *Buteo buteo*.

If individuals of the same species were given an unrestrained choice would they, for example, select the same nest site? Individual variability, related to genetic and phenotypic differences, would probably mean that the answer was no. For example, a simple experiment described by Kettlewell (1955) indicated that wild-type and melanic peppered moth *Biston betularia* preferentially selected backgrounds that gave the maximum cryptic advantage. In reality most individuals will not have an unconstrained choice. They usually exist within an intraspecific and/or interspecific landscape that influences their decisions, e.g. Hohmann (1994) showed that buzzards of different social status used different habitat and this was thought to be due the effect of the population dominance hierarchy. In addition, organisms make 'decisions' at some point in time in a particular ecological landscape, thus their decisions have a historical context. It may not be possible for an individual to alter its decision at a later time. Obvious examples of this include sessile organisms, particularly plants. Harper (1977) described unpublished work that demonstrated how the distribution of *Ranunculus bulbosus* seedlings followed the outlines of cattle footprints that had since disappeared. Harper suggested that only rarely can such behaviour be traced back to a causal event. Similarly, Bocard & Legendre (1994) and Harvey (1995) noted that spatially structured historical events may be an important contributor to community structure, but one which is impossible to assess. Thus, organisms may be in their current locations because of past rather than current events. It is difficult to see how this ecologically important information could be incorporated into a classifier.

Spatial autocorrelation, which is the tendency of neighbouring sample units to possess similar characteristics, is a potential problem for all area-based studies. If sample data are spatially autocorrelated the assumption of independence between cases will be violated leading to problems with the significance of test statistics. If the variables used by the classifier do not reflect fully the 'choices' made by an animal, residuals from a fitted model will exhibit spatial autocorrelation (Augustin *et al.* 1996). Spatial autocorrelation arising in this way is algorithmic since it results from the selection of inappropriate variables. However, spatial autocorrelation may also arise when the probability of occurrence in one sampling unit is not independent of the probability of occurrence in its neighbouring sampling units. Few papers have addressed the spatial autocorrelation problem despite warnings about its effects (for example: Bocard *et al*. 1992; Legendre 1993). Smith (1994) provided a possible solution that incorporated information about neighbours as a predictor variable. He noted that this approach makes prediction difficult since knowledge of neighbours will be missing, although it may be possible to use an iterative technique to generate the predictions. Augustin *et al*. (1996) developed a different iterative method that could be used to predict presence/absence in unsurveyed squares prior to calculating an autocorrelation term based on the occupancy of neighbouring squares.

There are, therefore, a variety of ecological processes which can create difficulties for classifiers leading to more prediction errors. In many analyses it is impossible to incorporate suitable corrective measures, e.g. the location of competitors, into the classifier because the relevant ecological data are unavailable. It is possible, but not necessarily desirable, to use *post-hoc* criteria as an aid to explaining prediction errors. For example, Pereira & Itami (1991) used local expertise to interpret the failure of their model to predict two isolated areas of squirrel activity.

Confusion matrix derived measures

A variety of error or accuracy measures can be calculated from a confusion matrix (Table 2). All of the measures described in Table 2 assume that data are counts and not percentages.

'Sensitivity' is the conditional probability that case X is correctly classified, $p(X_{Alg} \mid X_{true})$. 'Specificity' is the inverse, $p(not\ X_{Alg} \mid X_{False})$. 'Positive predictive power' assesses the probability that a case is X if the algorithm classifies the case as X, $p(X_{True} \mid X_{Alg})$. 'Negative predictive power' assesses the probability that a case is not X if the algorithm does not classify the case as X, $p(X_{False} \mid not\ X_{Alg})$. These measures have different characteristics, in particular some are sensitive to the prevalence of positive cases. Table 3 is based on the

Table 2 Confusion matrix derived measures of classification accuracy.

Measure	*Calculation*
Prevalence	(a + c)/N
Overall diagnostic power	(b + d)/N
Correct classification rate	(a + d)/N
Sensitivity	a/(a + c)
Specificity	d/(b + d)
False positive rate	b/(b + d)
False negative rate	c/(a + c)
Positive predictive power (PPP)	a/(a + b)
Negative predictive power (NPP)	d/(c + d)
Misclassification rate	(b + c)/N
Odds-ratio	(ad)/(cb)
Kappa	[(a + d) − (((a + c)(a + b) + (b + d)(c + d))/N)]/[N − (((a + c)(a + b) + (b + d)(c + d))/N)]
NMI n(s)	[−a.ln(a)−b.ln(b)−c.ln(c)−d.ln(d)+(a+b).ln(a+b)+(c+d).ln(c+d)]/[N.ln N −((a+c).ln (a+c) + (b+d).ln(b+d))]

https://doi.org/10.1017/S0376892997000088 Published online by Cambridge University Press

Table 3 The effect of prevalence on the predictive power of a habitat association model using three hypothetical examples that assume different prevalence representing three levels of study (local fieldwork, regional survey and national survey). (*a*) Confusion matrices; (*b*) summary statistics derived from the confusion matrices.

a

		Observed		
Prevalence	*Prediction*	*Present*	*Absent*	*Total*
0.5 (local fieldwork)	Present	70	5	75
	Absent	30	95	125
	Total	100	100	200
0.1 (regional survey)	Present	70	45	115
	Absent	30	855	885
	Total	100	900	1000
0.011 (national survey)	Present	70	450	520
	Absent	30	8550	8580
	Total	100	9000	9100

b

	Prevalence		
	0.5	*0.1*	*0.011*
Summary statistic			
Overall diagnostic power	0.500	0.900	0.989
Sensitivity	0.700	0.700	0.700
Specificity	0.950	0.950	0.950
PPP	0.933	0.610	0.130
NPP	0.760	0.970	0.997
Misclassification rate	0.175	0.075	0.053
Odds ratio	44.33	44.33	44.33
Kappa	0.650	0.610	0.210
NMI	0.371	0.360	0.264

work of Baldessarini *et al.* (1983) and illustrates how prevalence affects some of the measures.

Only three measures given in Table 2 (odds-ratio, Kappa K and the normalized mutual information NMI) make full use of the information contained in the confusion matrix. The odds-ratio has an unfortunate characteristic of being infinite when either b or c are 0. Thus, it has the same value when the algorithm is perfect or lacks one type of error (Forbes 1995). Forbes (1995) suggests that a suitable confusion matrix-based measure should meet four requirements and obey six additional constraints. In particular, it should measure agreement and not association. A classifier that got everything wrong would have a highly significant association but no agreement. Kappa is the proportion of specific agreement and meets most of Forbes's constraints. Forbes notes that K may be sensitive to the sample size and fails when the size of one class far exceeds the other. He introduced the NMI measure, which obeys all of his requirements and constraints while being the most conservative of the measures tested. However, the NMI measure has non-monotonic behaviour under conditions of excessive errors.

The measures described in Table 2 serve different purposes and a measure should be selected to reflect its intended use. If the aim is to assess the effectiveness of the classifier a measure that assesses improvement over chance is appropriate, e.g. K. This is important because it is possible to obtain high overall accuracy using trivial rules when, for example, prevalence is low. Indeed overall accuracy, measured by the correct classification rate (Table 2), is dependent on the prevalence (p) since it can be rewritten as [(p.sensitivity) + $(1-p)$. specificity)] (Ruttimann 1994). For example, if prevalence is 5% it is possible to achieve a 95% correct classification rate by labelling all cases as negative. Landis & Koch (1977) have suggested the following ranges of agreement for the Kappa statistic: poor $K < 0.4$; good $0.4 < K < 0.75$ and excellent $K > 0.75$. The Tau coefficient is related to K, but its calculation is based on *a priori* probabilities of group membership instead of the *a posteriori* probabilities used for the estimation of K. Ma & Redmond (1995) suggested that Tau is a better measure of classification accuracy for use with remote-sensing data.

Huberty (1994) describes other methods for assessing if the predicted success of a classifier exceeds that expected by chance. As with the Tau coefficient these calculations depend upon the values of the prior probabilities of positive and negative cases (p_+ and p_-). These values are not necessarily the proportions observed in the training set, which may have been fixed to match some constraints of the classifier. The p_+ and p_- values may be based on an *a priori* criterion. Huberty (1994) describes the calculation of a standard normal statistic

Table 4 Use of a z-test to determine if the observed correct classification rate (o) exceeds that expected by chance (e_i). Calculations are shown for two confusion matrices, I and II, each of which is tested for two assumed prevalence proportions (p_+). Expected frequencies are calculated for two criteria: proportional chance (e_p) and maximum chance (e_m). Details of the z-test are given in the text, no test (NT) is carried out if $o < e_i$. (*a*) Example confusion matrices; (*b*) e_i and z values.

a

Matrix	*Predicted*	*Observed*	
		present	*absent*
I	present	60	5
	absent	20	95
II	present	70	450
	absent	30	8550

b

Matrix	o	p_+	e_p (z)	e_m (z)
I	155	0.40	92 (9.39)	100 (0.84)
		0.50	90 (9.69)	100 (9.19)
II	8620	0.01	8911 (NT)	9009 (NT)
		0.05	8555 (2.87)	8645 (NT)

https://doi.org/10.1017/S0376892997000088 Published online by Cambridge University Press

to test a null hypothesis of no difference between observed and expected correct classifications. This is a one-tailed test since there is no necessity for a test if the observed rate is less than the expected.

$$\text{observed correct classification (o)} = a + d \quad (1)$$

$$\text{expected correct classification (e)} = p_+ (a + c) + p_-(b + d) \quad (2)$$

$$z = \frac{o - e}{\sqrt{e(N - e)/N}} \quad (3)$$

If prevalence (p) is low, an expected maximum chance value (e_m) can be calculated by assuming a trivial classification rule of assigning absence to each case. In this situation the expected correct classification rate is equal to $(1 - p)N$. This allows a test of the hypothesis that there is no improvement over the maximum chance value. Example calculations are shown in Table 4. It is apparent from Table 4 that the z values are dependent on the possibly subjective assignment of values to p_+.

All of the measures described in this section depend on the values assigned to *a,b,c* and *d* in the confusion matrix. These values are obtained by application of a threshold criterion to a continuous variable generated by the classifier. Typically, the classifier generates a variable that has values within the range 0 − 1 to which a 0.5 threshold is applied. Thus, a continuous, or at least ordinal, variable is dichotomized. If the threshold criterion is altered the values in the confusion matrix will change. Often, the raw scores are available so it is relatively easy to examine the effect of changing the threshold. Even with techniques such as decision trees, which appear to use dichotomous variables, the software will have dichotomized a continuous variable.

There are a variety of reasons why the threshold value may need to be examined. For example, unequal group sizes

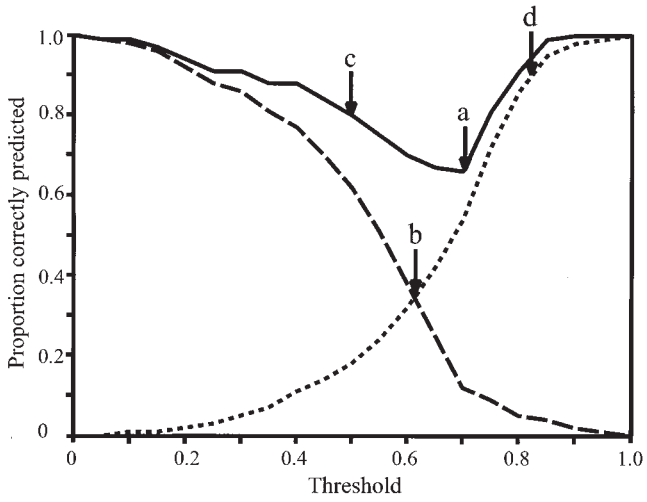


Figure 2 The effect of the cut-point threshold on the three error rates (total misclassifications – solid line; positives – dotted line; negatives – dashed line). Four possible cut-points are marked: (a) minimum total misclassifications; (b) cross-over of misclassification rates; (c) 0.5 cut-point; and (d) minimum acceptable error (90% of positives).

Table 5 Predition success from an unpublished model on bat distributions. Three potential cut-points are labelled to satisfy different criteria. [1] 'Best' cut-point to identify sites where bats can be studied (few false positives); [2] 'best' improvement over chance (Kappa); [3] eliminates sites with no bats.

Cut Point	*True* + a	*False* + b	*True* − c	*False* − d	*Sensitivity*	*Specificity*	*Kappa*
0.79[1]	15	4	93	235	0.14	0.98	0.16
0.64	67	33	41	206	0.62	0.86	0.49
0.62[2]	83	43	25	196	0.77	0.82	0.56
0.21	100	106	8	133	0.93	0.56	0.39
0.15	105	134	3	105	0.97	0.44	0.31
0.05[3]	108	190	0	49	1.00	0.21	0.14
0.00	108	239	0	0	1.00	0.00	0.00

(prevalence) can influence the scores for many of the classifier methods. This is particularly true for logistic regression which produces scores biased towards the larger group (Hosmer & Lemeshow 1989). Similarly, if we have decided that FN errors are more serious than FP errors the threshold can be adjusted to decrease the FN rate at the expense of an increased FP error rate. The effect of the threshold on three error rates is shown in Fig. 2. Few ecological models appear to have addressed this problem. Pereira & Itami (1991) used sensitivity analysis and comparisons with predictive im-

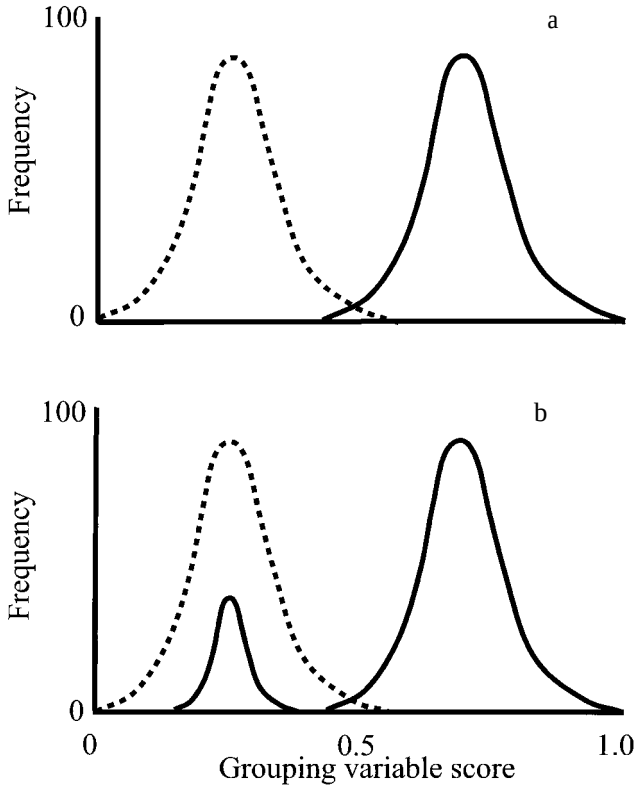


Figure 3 (*a*) Expected overlap in group scores. (*b*) Bimodality in group scores. Solid line – negative cases; dotted line – positive cases.

https://doi.org/10.1017/S0376892997000088 Published online by Cambridge University Press

provement over random classification to determine the optimum threshold value for the assessment of model accuracy. Capen *et al.* (1986) chose unspecified thresholds to balance the correct classification rates between presence and absence plots, this is equivalent to an equal weighting of FP and FN errors. Fielding & Haworth (1995) used a threshold which was calculated as the mid-point between the mean probabilities of occupancy for the present and absent groups to reduce the FN error rates. Other thresholds could be applied, e.g. a 'minimum acceptable error' (a FN criterion) could be defined that depended on the intended application of the classifier. For example, we could tolerate more false-positives for a particularly endangered species. If the purpose of the model was to identify experimental sites where we could be certain of finding a species, the threshold would be adjusted to minimize the FP error rates. Table 5 shows the effect of amending the threshold and demonstrates how different thresholds are appropriate for different aims.

If a classifier produces prediction errors there must be overlap in group scores (Fig. 3*a*). Usually it is expected that this overlap will be in the tails of two unimodal distributions. However, if ecological processes are generating false predictions we may observe bimodal score distributions (Fig. 3*b*). This bimodality will reduce the effectiveness of a threshold adjustment. An alternative solution to threshold adjustments is to make use of all the information contained within the original continuous variable and calculate threshold independent measures (see below).

Cost matrices

Measures derived from confusion matrices assume that both error types are equivalent. There are situations, particularly in conservation-based models, where this assumption can be questioned. If a model is used to define protected areas, failure to correctly predict positive locations will be more 'costly' (in conservation-terms) than would the prediction of false positives, i.e. 'FN cost' (FNC) > 'FP cost' (FPC). Although these inequalities can be compensated for partly by the choice of error measure and threshold, it is possible to adopt other approaches. One method that has been used by artificial intelligence workers is the concept of a cost matrix that weights errors prior to the calculation of model accuracy. In the absence of clear economic gains and losses, the allocation of weights must be subjective. In a conservation-based model we may be able to assign weights by taking into account perceived threats to the species. Lynn *et al.* (1995) used a matrix of misclassification costs to evaluate the performance of a decision-tree model for the prediction of landscape levels of potential forest vegetation. Their cost structure was based on the amount of compositional similarity between pairs of groups.

Threshold-independent measures

One problem with the threshold dependent measures is their failure to use all of the information provided by the classifier. Although dichotomous classifications are convenient for decision making they can introduce distortions (Deleo & Campbell 1990). Altman *et al.* (1994) showed that the dichotomization of a continuous variable, to give an 'optimal *p*-value', using groupings determined by the data, will result in bias. The medical literature has recognized these problems and other measures have been introduced. In particular, the use of threshold-independent receiver operating characteristic (ROC) plots has received considerable attention.

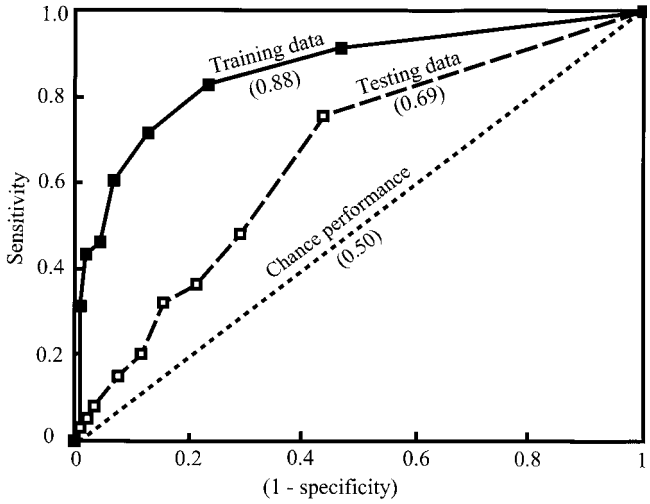


Figure 4 An example ROC plot. The data relate to Golden Eagle distribution using a logistic regression model. Figures in parentheses are the areas under the curves.

The ROC technique developed in signal processing and the term 'receiver operating characteristic' refers to the performance (the 'operating characteristic') of a human or mechanical observer (the 'receiver') engaged in assigning cases into dichotomous classes (Deleo 1993). Zweig & Campbell (1993) provide a review of the use of ROC methodology in clinical medicine; they include details of software that may be used for ROC analyses.

A ROC plot is obtained by plotting all sensitivity values (true positive fraction) on the *y* axis against their equivalent (1 − specificity) values (false positive fraction) for all available thresholds on the *x* axis, as in the example shown in Fig. 4. The area under the ROC function (AUC) is usually taken to be an important index because it provides a single measure of overall accuracy that is not dependent upon a particular threshold (Deleo 1993). The value of the AUC is between 0.5 and 1.0. If the value is 0.5, the scores for two groups do not differ, while a score of 1.0 indicates no overlap in the distributions of the group scores (Fig. 4). Typically, values of the AUC will not achieve these limits. A value of 0.8 for the AUC means that for 80% of the time a random selection from the positive group will have a score greater than a random selection from the negative class (Deleo 1993).

The ROC plot does not provide a rule for the classification of cases. However, there are strategies that may be used to develop decision rules (Deleo 1993; Zweig & Campbell

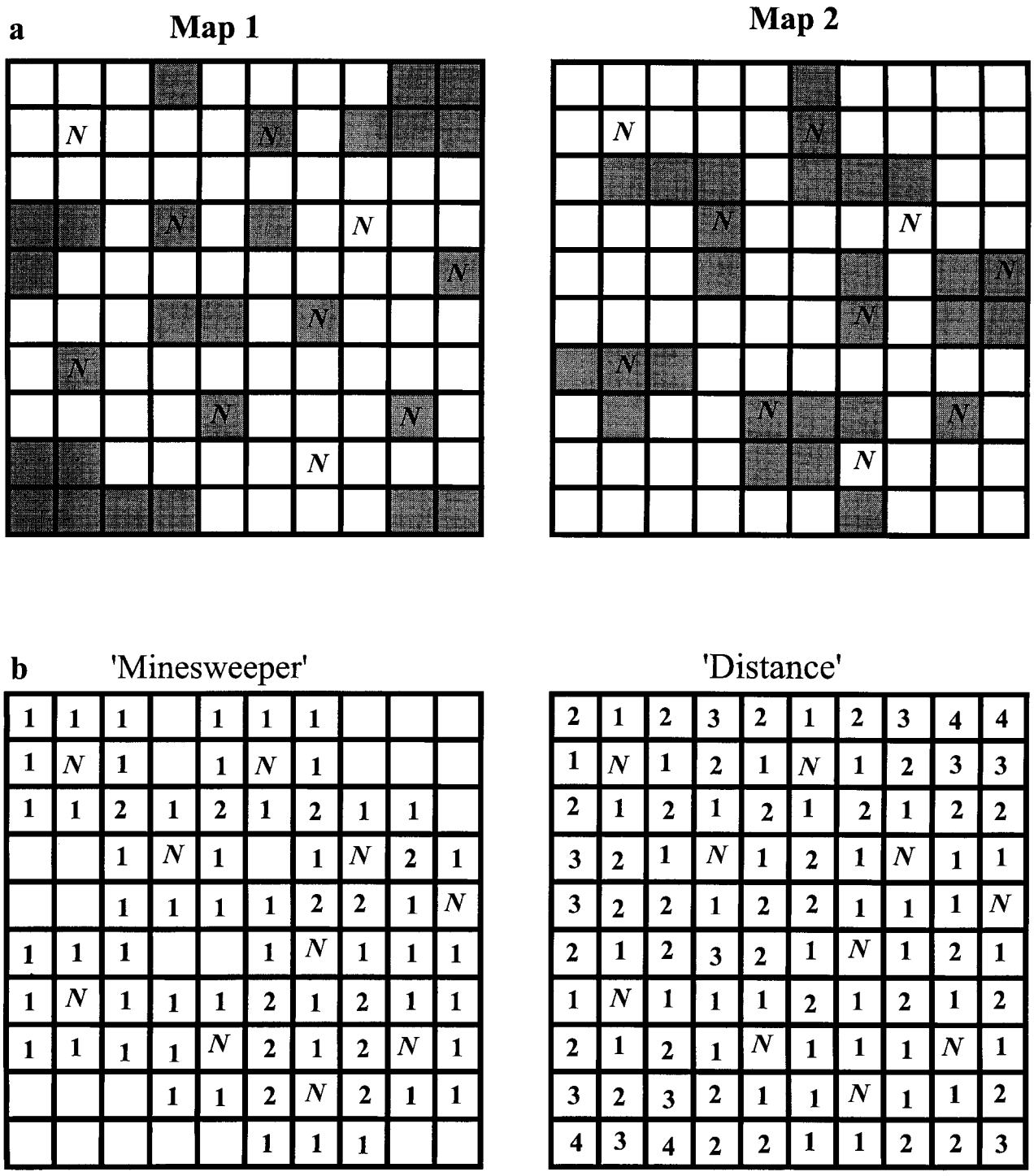
https://doi.org/10.1017/S0376892997000088 Published online by Cambridge University Press

1993). Two elements are required to identify the appropriate threshold (Zweig & Campbell 1993). The first is the relative cost of FP and FN errors. Assigning values to these costs is complex and subjective and dependent upon the context within which the classification rule will be used. As a guideline Zweig & Campbell (1993) suggest that if FPC > FNC the threshold should favour specificity, while sensitivity should be favoured if FNC > FPC. The second is the prevalence (*p*) of positive cases. Combining these factors allows the calculation of a slope (Zweig & Campbell 1993).

$$m = (FPC/FNC) \times ((1\text{-}p)/p) \quad (4)$$

If the ROC plot is a smooth and parametric curve, *m* de-

Figure 5 (*a*) Two hypothetical predictions for the presence/absence of nests. The locations of nests are marked N. Squares predicted to contain a species are shaded. (*b*) Weights for the calculation of spatially-corrected error measures using the Minesweeper and Distance algorithms.



https://doi.org/10.1017/S0376892997000088 Published online by Cambridge University Press

scribes the slope of a tangent to this curve. The point at which the tangent touches the curve identifies a particular sensitivity/specificity pair. If the ROC plot is a stepped non-parametric curve the equivalent sensitivity/specificity pair is found by moving a line, with slope *m*, from the top left of the ROC plot. The sensitivity/specificity pair is found where the line and the curve first touch (Zweig & Campbell 1993). The desired threshold is the value which gives the selected sensitivity/specificity pair.

Kraemer (1988) suggests some caution is necessary when using ROC methods with biological data since biological cases may not be directly equivalent to the original definition. In particular, the original ROC model assumes that the group allocation is absolutely reliable and each signal is homogeneously presented and processed. Kraemer (1988) provides a modified method based on the relative costs of FN and FP cases; these costs are used in conjunction with a plot of the sensitivity quality against the specificity quality.

Spatially-corrected measures

One of the difficulties of basing error assessment on confusion matrices and other summary statistics is that they do not take into account the spatial context of the errors. Figure 5 shows two sets of predictions for the same positive locations that will yield identical values for all previously described measures. Depending on the context and purpose of the predictions we could, by examining prediction maps, place different ecological interpretations on the results. Buckland & Elston (1993) discussed how patterns in prediction errors could be used to infer spatial patterns of habitat suitability.

There is some justification for calculating error measures that take into account the spatial characteristics of the data. In this section we present two simple methods that weight false positive errors by reference to their proximity to actual positive cases. These measures are related to the technique developed by Augustin *et al.* (1996) for incorporating explicit autocorrelation into general linear presence/absence models. The rationale for the spatial weighting is that FPs close to real positives may be less serious errors than FPs distant from a real positive. Two approaches to spatial weighting are illustrated (Fig. 5*b*). The first approach is based on the ubiquitous Minesweeper computer game. The weight is calculated as the number of adjacent positive squares

$$w_M = 1 - (\text{neighbours})/9. \quad (5)$$

In the second approach, errors are weighted by their 'city-block' distance from the nearest positive case. A weight can then be calculated as

$$w_D = 1 - 1/(2.\text{distance}) \quad (6)$$

with a threshold constraint such that any $w_D > 0.85$ is rounded to 1. Using these weights an adjusted confusion matrix may be constructed from which adjusted error measures are calculated. If the ratio of adjusted errors to actual errors is calculated it will provide information about the spatial characteristics of the prediction errors. Table 6 shows example calculations using the maps in Fig. 5. As expected, the adjustments have their greatest effect on Map 2 predictions. Table 6 also shows that the Minesweeper algorithm provides the most conservative adjustment.

If it is considered that some of the FP errors are a consequence of interference, it would also be possible to use information about territory size and spacing to weight some of the errors.

Comparing error rates

The results presented in most papers are generally those produced by a 'best' model. A variety of methods may have been used to decide between candidate models. It is not always certain that appropriate questions have been asked. Similarly, if different classifiers (e.g. discriminant analysis and logistic regression) have been tested we may wish to decide which is the 'best'. Judgements of this type usually depend on an error-based comparison. Chatfield (1995) suggested that when comparing forward stepwise regression models we should ask the question 'does it (the extra variable) provide value for money in improving predictions?' rather than 'does it lead to a significant improvement of fit?'. In other words when considering relative performance we should consider both accuracy and costs. More detailed aspects of cost-sensitive classifications are described in Turney (1995).

Huberty (1994) describes a range of techniques, that do not incorporate costs, for the comparison of results from different classifiers. The results from two classifiers can be compared by constructing the following 2 × 2 table that cross-tabulates prediction success:

		Classifier 1	
		correct	incorrect
	correct	*a*	*b*
Classifier 2	incorrect	*c*	*d*

The test must compare if $(a + b)/N$ is significantly different from $(a + c)/N$. Huberty (1994) suggests using McNemar's test to calculate a measure,

$$(b - c)^2 / (b + c) \quad (7)$$

that has an approximate χ^2 value with 1 degree of freedom if $b + c$ is 'large'. Multiple comparisons can be carried out using Cochran's *Q* test. Huberty (1994) has additional details.

One of the difficulties with this type of approach is that it depends on the dichotomization of two continuous variables. An alternative approach is to compare ROC plots. If one of the curves is consistently above the other, clearly one is better. Zweig & Campbell (1993) provide full details of methods for comparing the areas under ROC plots for both paired and unpaired data.

Conclusions and recommendations

The quality of a predictive model is usually judged by its

https://doi.org/10.1017/S0376892997000088 Published online by Cambridge University Press

accuracy. This review has demonstrated that there are many routes to the calculation of predictive accuracy and that some objective consideration should be given to the choice of measure. The results of model tests must be interpreted in the context of how the model will be applied, some applications may be able to tolerate less accuracy or precision (Schamberger & O'Neil 1986). If the objective was to purchase habitats with high opportunity costs the model should accurately predict species presence, however if the model was to be used to predict impacts for endangered species false positives may be of greater concern (Morrison *et al.* 1992). Since there are many aspects to consider when assessing prediction error we would like to suggest the following approach.

(1) Decide which data are to be used for the estimation of error. Do not rely on an estimate based on resubstitution of the training data. A more robust estimate will be obtained from independent testing data.
(2) If predictions are to be restricted to a homogeneous region consider a data-partitioning technique. If the predictions are to be tested for their generality use a prospective sample selected via temporal or geographical criteria.
(3) If data-partitioning is to be used consider using more than one approach, ideally including *k*-fold partitioning or jack-knifing. When deciding on a size for the training set use a heuristic such as that suggested by Schaafsma & van Vark (1979), but also take into account any cases:variables constraints imposed by your classifier.
(4) Understand the nature of any error measures that are used. In particular, take account of the effect of prevalence. Overall accuracy may be a very poor guide to the value of your predictions. Use prediction error cost criteria to guide your choice, e.g. if false positive errors are more costly than false negatives, use sensitivity.
(5) If you wish to determine if a classifier predicts better than chance, use a measure such as Kappa or NMI. Recall that NMI is less affected by prevalence.
(6) ROC plots avoid the problems associated with threshold effects. If error is to be based solely on confusion-matrix-derived measures consider adjusting the threshold. It is desirable to use *a priori* criteria for deciding on a threshold.
(7) If classifiers are to be ranked, comparisons based on ROC plots are likely to be more robust since they are independent of the values in a confusion matrix.
(8) If the aim is to improve within-region accuracy consider using spatial analysis methods that incorporate the almost inevitable spatial autocorrelation.
(9) If the aim is to improve the predictive success with prospective samples, based on a different region, an attempt should be made to remove the spatial structure from the models. Bocard *et al.* (1992) and Okland & Eilertsen (1994) describe how this may be accomplished using canonical correspondence analysis.
(10) If appropriate for your data examine the spatial pattern of the errors and consider using, with caution, *post-hoc* hypotheses to interpret the patterns.
(11) Consider weighting errors if there are ecological or economic justifications.
(12) Be cautious of any statement of model accuracy that does not justify the choice of error measure.
(13) If, after model validation, the aim is to derive a robust classification rule, all of the available data should be used.

Acknowledgements

We would like to thank all of our colleagues who commented on earlier versions of this manuscript. We are very grateful for comments made by the referees.

References

Aebischer, N.J., Robertson, P.A. & Kenward, R.E (1993) Compositional analysis of habitat use from animal radio-tracking data. *Ecology* 74: 1313–25.
Altman, D.G., Lausen, B., Sauerbrei, W. & Schumacher, M. (1994) Dangers of using 'optimal' cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute* 86: 829–35.
Anon. (1995) Bird conservation: the science and the action preface. Conclusions and recommendations. *Ibis* 137: S3–S7.
Augustin, N.H., Mugglestone, M.A. & Buckland, S.T. (1996) An autologistic model for the spatial distribution of wildlife. *Journal of Applied Ecology* 33: 339–47.
Austin, M.P. & Gaywood, M.J. (1994) Current problems of environmental gradients and species response curves in relation to continuum theory. *Journal of Vegetation Science* 5: 473–82.
Austin, M.P., Nicholls, A.O., Doherty, M.D. & Meyers, J.A. (1994) Determining species response functions to an environmental gradient by means of a beta function. *Journal of Vegetation Science* 5: 215–28.
Baldessarini, R.J., Finklestein, S. & Arana, G.W. (1983) The predictive power of diagnostic tests and the effect of prevalence of illness. *Archives of General Psychiatry* 40: 569–73.
Blayo, F., Chevenal, Y., Guérin-Dugué, A., Chentouf, R., Aviles-Cruz, C., Madrenas, J., Moreno, M. & Voz, J.L. (1995) Enhanced learning for evolutive neural architecture. Deliverable R3-B4-P Task B4: Benchmarks. ESPIRIT Basic Research Project Number 6891. Available from ftp.dice.ucl.ac.be/pub/neural-nets/ELANA/databases.
Bocard, D., Legendre, P. & Drapeau, P. (1992) Partialling out the spatial component of ecological variation. *Ecology* 73: 1045–55.
Bocard, D. & Legendre, P. (1994) Environmental control and spatial structure in ecological communities: an example using oribatid mites (Acari, Oribata). *Environmental and Ecological Statistics* 1: 37–62.
Brennan, L.A., Block, W.M. & Gutiérrez, R.J. (1986) The use of multivariate statistics for developing habitat suitability index models. In: *Wildlife 2000: Modelling Habitat Relationships of Terrestrial Vertebrates*, ed. J.A. Verner, M.L. Morrison & C.J. Ralph, pp. 177–82. Madison, WI, USA: University of Wisconsin Press.
Buckland, S.T. & Elston, D.A. (1993) Empirical models for the spatial distribution of wildlife. *Journal of Applied Ecology* 30: 478–95.
Capen, D.E., Fenwick, J.W., Inkley, D.B. & Boynton, A.C. (1986)

Multivariate models of songbird habitat in New England forests. In: *Wildlife 2000: Modelling Habitat Relationships of Terrestrial Vertebrates*, ed. J.A. Verner, M.L. Morrison and C.J. Ralph, pp. 171–75. Madison, WI, USA: University of Wisconsin Press.

Cederlund, G. & Sand, H. (1994) Home-range size in relation to age and sex in moose. *Journal of Mammology* 75: 1005–12.

Chatfield, C. (1995) Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series A.* 158: 419–66.

Deleo, J.M. (1993) Receiver operating characteristic laboratory (ROCLAB): software for developing decision strategies that account for uncertainty. In: *Proceedings of the Second International Symposium on Uncertainty Modelling and Analysis*, pp. 318–25. College Park, MD: IEEE Computer Society Press.

Deleo, J.M. & Campbell, G. (1990) The fuzzy receiver operating characteristic function and medical decisions with uncertainty. In: *Proceedings of the First International Symposium on Uncertainty Modelling and Analysis*, pp. 694–9. College Park, MD: IEEE Computer Society Press.

Donázar, J.A., Hiraldo, F. & Bustamante, J. (1993) Factors influencing nest site selection, breeding density and breeding success in bearded vulture (*Gypaetus barbatus*). *Journal of Applied Ecology* 30: 504–14.

Fielding, A.H. & Haworth, P.F.(1995) Testing the generality of bird-habitat models. *Conservation Biology* 9: 1466–81.

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

Harper, J.L. (1977) *Population Biology of Plants*. London, UK: Academic Press: 892 pp.

Harvey, H.J. (1995) The National Trust and nature conservation: Prospects for the future. *Biological Journal of the Linnean Society* 56, Suppl. A: 231–48.

Hengveld, R. (1994) Biogeographical ecology. *Journal of Biogeography* 21: 341–51.

Hohmann, U. (1994) Status specific habitat use in the common buzzard *Buteo buteo*. In: *Raptor Conservation Today*, ed. B.U. Meyburg & R.D. Chancellor, pp. 359–66. Berlin, Germany: WWGBP / Pica Press.

Hosmer, D.W. & Lemeshow, S. (1989) *Applied Logistic Regression.* New York, USA: Wiley: 307 pp.

Huberty, C.J. (1994) *Applied Discriminant Analysis*. New York, USA: Wiley Interscience: 466 pp.

Kettlewell, H.B.D. (1955) Recognition of appropriate backgrounds by pale and dark phases of Lepidoptera. *Nature* 175: 943–44.

Kraemer, H.C. (1988) Assessment of 2 × 2 associations: Generalisation of signal detection methodology. *The American Statistician* 42: 37–49.

Landis, J.R. & Koch, G.C. (1977) The measurement of observer agreement for categorical data. *Biometrics* 33: 159–74.

Legendre, P. (1993) Spatial autocorrelation: trouble or new paradigm. *Ecology* 74: 1659–73.

Lynn, H., Mohler, C.L., DeGloria, S.D. & McCulloch, C.E. (1995) Error assessment in decision-tree models applied to vegetation analysis. *Landscape Ecology* 10: 323–35.

Ma, Z. & Redmond, R.L.(1995) Tau coefficients for accuracy assessment of classifications of remote sensing data. *Photogrammetric Engineering & Remote Sensing* 61: 435–39.

Manly, B.F.J., McDonald, L.L. & Thomas, D.L. (1992*) Resource Selection by Animals: Statistical Design and Analysis for Field Studies*. London, UK: Chapman & Hall: 177 pp.

Morrison, M.L., Marcot, B.G. & Mannan, R.W. (1992) *Wildlife Habitat Relationships. Concepts and Applications*. Madison, WI, USA: University Wisconsin Press: 341 pp.

Morrison, M.L., Timoss, I.C. & With, K.A. (1987) Development and testing linear regression models predicting bird-habitat relationships. *Journal of Wildlife Management* 51: 247–53.

Newton, I. (1979) *Population Ecology of Raptors.* Berkhamstead, UK: Poyser: 399 pp.

Okland, R.H. & Eilertsen, O. (1994) Canonical correspondence analysis with variation partitioning: some comments and an application. *Journal of Vegetation Science* 5: 117–26.

Osborne, P.E. & Tigar, B.J. (1992) Interpreting bird atlas data using logistic models: an example from Lesotho, Southern Africa. *Journal of Applied Ecology* 29: 55–62.

Pereira, J.M.C. & Itami, R.C. (1991) GIS-based habitat modelling using logistic multiple regression: a study of the Mt. Graham red squirrel. *Photogrammetric Engineering & Remote Sensing* 57: 1475–86.

Ratcliffe, D. (1993) *The Peregrine Falcon.* Second Edition. London, UK: T. & A.D. Poyser: 454 pp.

Reed, J.M. & Dobson, A.P. (1993) Behavioural constraints and conservation biology. *Trends in Ecology and Evolution* 8: 253–5.

Rencher, A.C. (1995) *Methods of Multivariate Analysis*. New York, USA: Wiley: 627 pp.

Rotenberry, J.T. (1986) Habitat relationships of shrubsteppe birds: even 'good' models cannot predict the future. In: *Wildlife 2000: Modelling Habitat Relationships of Terrestrial Vertebrates*, ed. J.A. Verner, M.L. Morrison & C.J. Ralph, pp. 217–21. Madison, WI, USA: University of Wisconsin Press.

Ruttimann, U.E. (1994) Statistical approaches to development and validation of predictive instruments. *Critical Care Clinics* 10: 19–35.

Schaafsma, W. & van Vark, G.N. (1979) Classification and discrimination problems with applications. Part IIa. *Statistica Neerlandica* 33: 91–126.

Schooley, R.L. (1994) Annual variation in habitat selection: patterns concealed by pooled data. *Journal of Wildlife Management* 58: 367–74.

Schamberger, M.L. & O'Neil, L.J. (1986) Concepts and constraints of habitat-model testing. In: *Wildlife 2000: Modelling Habitat Relationships of Terrestrial Vertebrates*, ed. J.A. Verner, M.L. Morrison & C.J. Ralph, pp. 5–10. Madison, WI, USA: University of Wisconsin Press.

Schwede, G., Hubert, H. & McShea, W. (1993) Social and spatial organization of female white-tailed deer, *Odocoileus virginianus*, during the fawning season. *Animal Behaviour* 45: 1007–17.

Smith, P.A. (1994) Autocorrelation in logistic regression modelling of species' distributions. *Global Ecology and Biogeography Letters* 4: 47–61.

Solonen, T. (1993) Spacing of birds of prey in Southern Finland. *Ornis Fennica* 70: 129–43.

Stauffer, D.F. & Best, L.B. (1986) Effects of habitat type and sample size on habitat suitability index models. In: *Wildlife 2000: Modelling Habitat Relationships of Terrestrial Vertebrates*, ed. J. A. Verner, M. L. Morrison & C. J. Ralph, pp. 71–77. Madison, WI, USA: University of Wisconsin Press.

Stillman, R.A. & Brown, A.F. (1994) Population sizes and habitat sizes of upland breeding birds in the South Pennines, England. *Biological Conservation* 69: 307–14.

Stockwell, D.R.B. (1992) Machine learning and the problem of pre-

diction and explanation in ecological modelling. Ph.D. Thesis, Australian National University.

Turney, P. (1995) Cost-sensitive classification:empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of Artificial Intelligence Research* 2: 369–409.

Verbyla, D.L. & Litaitis, J.A. (1989) Resampling methods for evaluating classification accuracy of wildlife habitat models. *Environmental Management* 13: 783–7.

Weir, D. & Picozzi, N. (1983) Dispersion of buzzards in Speyside. *British Birds* 76: 66–78.

Wiens, J.A., Rotenberry, J.T. & Van Horne, B. (1987) Habitat occupancy patterns of North American shrubsteppe birds: the effects of spatial scale. *OIKOS* 48: 132–47.

Zweig, M.H. & Campbell, G. (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry* 39: 561–77