

6. Evaluation of modelling techniques: I.

Development of evaluation methodology

6.1 Introduction

As detailed in Section 2, this consultancy had two main components. The first component was a broad evaluation of surrogates, in which the performance of different types of surrogates was compared (see Sections 3,4 and 5). The second component involved a more detailed evaluation of the performance of modelling techniques in predicting species distributions. This chapter describes the methodology used to conduct the evaluation of modelling techniques. Section 7 then describes the application of this methodology to data from forested north east NSW.

Two other recent consultancies have evaluated the performance of predictive modelling techniques for Commonwealth government agencies. CSIRO Division of Wildlife and Ecology has conducted an extensive evaluation of a range of predictive modelling techniques for the Australian Nature Conservation Agency (ANCA) and the Environmental Resource Information Network (ERIN), using both real and simulated data (Austin *et al.* 1994; Austin 1994; Austin and Meyers 1995; Austin *et al.* 1995). The NSW National Parks and Wildlife Service has also conducted a separate consultancy for ANCA, testing the performance of habitat models for vertebrates and vascular plants using independent field survey data (Pearce and Ferrier 1996). Both of these consultancies are of direct relevance to the evaluation described in this report. Wherever possible we have attempted to complement, rather than duplicate, the work of the other consultancies. Readers interested in gaining a comprehensive overview of the performance of predictive species models are strongly advised to read all three consultancy reports.

The evaluation in the current consultancy was restricted to modelling techniques that extrapolate the potential distribution of species across large regions by modelling the probability (or relative likelihood) of presence in relation to remotely mapped or derived environmental variables. The evaluation did not consider modelling of species abundance nor the use of environmental variables that require direct field measurement.

Predictive modelling of species distributions can be viewed as a type of environmental surrogate. The general performance of predictive species modelling relative to other surrogates (vegetation mapping, environmental domains etc) was evaluated in Section 4. Although predictive modelling performed well in that evaluation, the analysis was restricted to a particular type of modelling, i.e. generalised additive modelling of presence/absence survey data in relation to environmental variables mapped at a relatively fine spatial resolution. The detailed evaluation of modelling described below and in Section 7 was aimed at assessing the relative performance of different types of modelling. Of particular interest were:

- the relative performance of different mathematical modelling algorithms (e.g. generalised linear modelling, decision tree modelling);
- the relative performance of models based on 'presence/absence' species data derived from well designed systematic field surveys as opposed to 'presence-only' data derived from *ad hoc* survey work;
- the relative performance of models based on fine-scaled versus coarse-scaled environmental predictors; and
- the relative performance of models for different types of species (different taxonomic groups, rare species versus common species etc).

The purpose of this chapter is to describe the methodology adopted in the consultancy for evaluating the effect of the above factors on model performance.

6.2 Selection of evaluation methodology

6.2.1 Requirements

The evaluation strategy adopted in this consultancy was to test predictions from models against real presence/absence data. The strategy required two methodological components:

- a procedure for obtaining both predicted probability (or relative likelihood) of presence and actual presence/absence at a sample of sites; and
- a technique for measuring the performance of models in terms of the correspondence between predicted and actual occurrence of species at these sites.

Both components needed to be able to accommodate models derived from either presence/absence or presence-only data, using a wide range of mathematical modelling techniques.

6.2.2 Review of potential approaches

Evaluation procedures

Procedures that can be used to compare predictions from models with actual observations include:

- Simple resubstitution, whereby performance is measured in terms of how well predictions from a model fit the data from which the model was derived. This is the easiest procedure to implement and has been widely used to evaluate predictive models of species distributions (e.g. Lindenmayer *et al.* 1990; Walker 1990). Resubstitution has, however, been shown to provide a biased estimate of model performance (Stone 1974; Efron 1982; 1983; Gong 1986). The technique yields an inflated estimate of performance (or deflated estimate of error rate) because the same data that were used in the derivation of the model are also used to test the model. The following alternatives are designed to reduce the bias inherent in resubstitution by testing predictions against data that were not used in the derivation of a model.
- Statistical resampling procedures such as cross-validation, jackknifing and bootstrapping employ resampling of a single dataset to reduce bias in the assessment of model performance (Efron 1982, 1983; Gong 1986; Van Houwelingen and Le Cessie 1990; Efron and Tibshirani 1993). In cross-validation, the dataset is randomly divided into two subsets (usually of equal size). A model is derived using one of the subsets and then tested against the other subset. Cross-validation has been employed recently by Stockwell *et al.* (1990) and Flather and King (1992) to evaluate the performance of predictive models of species distributions. Jackknifing is a refined version of cross-validation in which each survey site is tested in turn against a model derived using all the other sites. Bootstrapping uses random resampling (with replacement) of a dataset to derive a large number of models from which a mean measure of performance can be estimated. A vital assumption underlying all of these procedures is that the dataset to which resampling is applied is itself an unbiased sample. If the dataset is heavily biased toward a particular part of a region's environmental or geographical space then measures of performance derived from statistical resampling of this dataset will be inflated relative to the true performance of models across the entire region.

- The most rigorous method for assessing the performance of predictive models of species distributions is to conduct independent field validation surveys in areas not sampled during the original model derivation. Pearce and Ferrier (1996) have applied this approach in a separate consultancy conducted by NSW NPWS for ANCA.

Measures of performance

All of the above procedures generate predictions and actual observations for a sample of sites. Predictions are expressed as probabilities derived from modelling of presence/absence data, or relative likelihoods of occurrence derived from modelling of presence-only data. Actual observations, in this study, record the presence or absence of species at sites. Measures of performance can be derived by comparing these actual observations with predictions generated by modelling.

We first need to define more precisely what we mean by ‘performance’. The literature on evaluation of models has employed a diverse, and often confusing, array of terms relating to model performance such as ‘accuracy’, ‘reliability’, ‘calibration’, ‘sensitivity’, ‘specificity’ and ‘prediction error’. The problem is further aggravated by the fact that most of the literature deals with modelling of continuous response variables rather than binary variables of the type used in modelling of species distributions. Nevertheless, useful literature on the evaluation of probabilistic modelling of binary response variables exists within non-ecological disciplines such as weather forecasting and medical diagnosis (e.g. Mason 1982; Yates and Curley 1985; Swets 1988; Hadorn *et al.* 1992).

The terminology used to describe model performance in this report is based on that proposed by Murphy and Winkler (1992) for evaluation of probabilistic models. These authors define ‘accuracy’ as the overall degree to which predictions correspond to (or fit) actual observations. Accuracy is then partitioned into three main components. ‘Calibration’ is the degree to which predictions are, on average, either too high or too low relative to actual observations. ‘Discrimination’ is the degree to which predictions from the model discriminate between actual presences and absences or, in other words, the extent to which actual presences have higher predicted probabilities than actual absences. ‘Refinement’ is related to discrimination and measures the degree to which predicted probabilities are spread across the entire probability range (from 0 to 1) rather than being concentrated in only part of the range.

All three components of accuracy can be measured for species models that predict probability of presence by modelling presence/absence survey data in relation to environmental predictors. For models derived from presence-only data, and therefore predicting relative likelihood of presence rather than probability, only the discrimination component of accuracy can be measured.

Specific measures of accuracy that have been, or could be, used to evaluate the performance of predictive models of species distributions include:

- Measures of predictive accuracy derived from a 2 by 2 classification table in which the rows represent predicted presence or absence and the columns represent actual observed presence or absence. Classification tables can be used to calculate a variety of accuracy indices including percentage correct, sensitivity, specificity, false positive fraction, false negative fraction, the Kappa statistic, the Phi coefficient, Yule’s Q and the log odds ratio. Swets (1986) provides a comprehensive review of accuracy indices derived from classification tables, while examples of their use in evaluation of predictive species models are given by Lindenmayer *et al.* (1990), Flather and King (1992) and Pearce *et al.* (1994). Most indices derived from classification tables are essentially measures of discrimination. They are measuring how well a model can discriminate between sites where a species is recorded as present and sites where the species is absent. To apply these indices to probabilistic models, such as those commonly used to model species distributions, an arbitrary threshold value (or rule)

must be specified for splitting the continuous range of predicted probabilities into predicted presences and absences. Measures of accuracy calculated using this approach have been shown to be very sensitive to the location of the presence/absence threshold and therefore lack robustness (Metz 1986; Swets 1986, 1988).

- Recent development of accuracy measures based on receiver operating characteristic (ROC) curves has provided a robust alternative to measures based on classification tables. Instead of employing a single arbitrary threshold for dividing predicted probabilities into presences and absences ROC based techniques vary the threshold continuously between 0 and 1. This results in a smooth ROC curve relating the true positive fraction to the false positive fraction as the threshold is varied. An ROC curve describes the inherent discrimination capacity of the underlying model. Indices of model discrimination are calculated from the area under the ROC curve. A detailed explanation of ROC methodology is beyond the scope of this report. Interested readers are referred to the extensive literature now available on the subject, including Swets and Pickett (1982), Mason (1982), Metz (1986) and Swets (1988). Hanley and McNeil (1982,1983) have demonstrated that a close approximation of the area under the ROC curve can be derived from a nonparametric Mann-Whitney U (or Wilcoxon W) statistic comparing the distribution of predicted values in the actual presence and absence samples.
- The classification table and ROC approaches described above measure only one component of accuracy, i.e. discrimination. They do not measure the calibration or refinement of predictions. The overall accuracy of a probabilistic model derived from presence/absence data can be assessed using a goodness-of-fit statistic summarising the deviations of individual predicted probabilities from actual presence/absence outcomes (e.g. Hosmer and Lemeshow 1989; Van Houwelingen and Le Cessie 1990; Miller *et al.* 1991). The most commonly used measure is the percentage of null deviance explained by the model. A method for partitioning a goodness-of-fit statistic for probabilistic predictions into the different components of accuracy, including calibration and refinement, was proposed by Cox (1958) and refined by Miller *et al.* (1991).

6.2.3 Approach adopted in DEST consultancy

Two types of biological data from forested north east NSW were used to evaluate model performance (details are provided in Section 7):

- presence/absence data generated by planned systematic field surveys; and
- presence-only data collated from *ad hoc* unplanned surveys.

Models fitted to the presence/absence data were evaluated using jackknifing, whereby each survey site was tested in turn against a model derived using all the other presence/absence sites. Models fitted to the presence-only data were evaluated using the presence/absence dataset (these two datasets are independent).

Two measures of accuracy were employed in the evaluation. A measure derived from the Mann-Whitney U statistic was used (as an approximation of the ROC statistic) to evaluate the discrimination ability of both presence-absence and presence-only models. The overall accuracy of probabilistic predictions from presence/absence models was measured using a goodness-of-fit statistic, the percentage of deviance explained by the model.

The evaluation approach adopted in this consultancy was less rigorous than that used in a separate consultancy conducted by NSW NPWS evaluating the performance of modelling techniques for ANCA (Pearce and Ferrier 1996). The ANCA consultancy employed independent field survey data to evaluate the predictive performance of presence/absence models. Accuracy was assessed using full ROC analysis and a decomposition of goodness-of-fit statistics into individual components of accuracy.

6.3 Adopted procedures for evaluating model performance

6.3.1 Evaluation of ‘presence/absence’ models using jackknifing

‘Presence/absence’ models are derived from planned biological surveys in which each species is recorded as either present or absent at each survey site. Because both presences and absences are recorded in the dataset, models derived from these data can be used to predict the probability of a species occurring at unsurveyed sites.

Jackknifing is used to evaluate the predictive performance of presence/absence models in the following way. Say, for example, a survey of diurnal birds had been conducted at 500 sites scattered throughout north east NSW. At each of these sites, species of birds were recorded as either present or absent. The probability of occurrence of one of these species (e.g. Olive Whistler) was then modelled in relation to a number of environmental variables held in a GIS database covering the region. To apply jackknifing to this model, 500 different versions of the model are derived, by excluding each survey site in turn from the dataset used to model the species. The probability of the species occurring at each survey site is then predicted from the model fitted without that site.

The data generated by the jackknifing consist of a predicted probability and an observed presence or absence for each survey site (in the above example there would be 500 pairs of predicted and observed values). These predicted and observed values are then evaluated using the performance measures described in Section 6.4. For a more detailed discussion of jackknifing and its relationship with other statistical resampling techniques (e.g. bootstrapping) see Efron (1982) or Efron and Tibshirani (1993).

6.3.2 Evaluation of ‘presence-only’ models using independent ‘presence/absence’ data

‘Presence-only’ models are derived from datasets in which only the locations of known presences of a species are recorded. No record is kept of those sites or areas that were surveyed without detecting the species, i.e. ‘absent’ sites. Presence-only models are used to predict the relative likelihood, rather than probability, of a species occurring at unsurveyed sites.

Because presence-only datasets are usually derived from opportunistic or *ad hoc* surveys they are particularly prone to bias in geographical and/or environmental coverage. The extent of such bias is difficult to assess due to a lack of recorded absences. Such datasets are not well suited to evaluation by statistical resampling (jackknifing, bootstrapping etc).

The approach adopted in this consultancy to evaluating presence-only models is to employ an independent presence/absence dataset derived from planned biological surveys. The data used in the evaluation therefore consist of a predicted relative likelihood and an observed presence or absence for each independent survey site. These data are then used to measure the ability of the presence-only model to discriminate between sites with known presences and absences (see Section 6.4.1). As discussed in Section 6.2.2 discrimination is the only component of accuracy that can be evaluated for presence-only models.

6.4 Adopted measures of model performance

6.4.1 Measure of discrimination: Mann-Whitney U test

Discrimination is measured using a statistic derived from the Mann-Whitney U test, which is equivalent to the Wilcoxon test for unmatched samples. Hanley and McNeil (1982) have demonstrated that the result of this test, when expressed in the manner described below, approximates the area under an ROC curve derived from the same data. The

statistic employed by Hanley and McNeil, and adopted in this consultancy, is calculated as:

$$\frac{\sum_1^{n_p} \sum_1^{n_a} S(x_p, x_a)}{n_p \times n_a}$$

where $S(x_p, x_a) = 1$ if $x_p > x_a$, 0.5 if $x_p = x_a$ and 0 if $x_p < x_a$

x_p = the predicted probability (or relative likelihood) for an evaluation site at which the species is recorded as present

x_a = the predicted probability (or relative likelihood) for an evaluation site at which the species is recorded as absent

n_p = the total number of evaluation sites at which the species is recorded as present

n_a = the total number of evaluation sites at which the species is recorded as absent

The statistic is measuring to what extent predicted probabilities (or relative likelihoods in the case of presence-only models) for presence sites are greater than those for absence sites. Values for the statistic can range from 0 to 1. Values greater than 0.5 indicate that predictions for presence sites are, on average, higher than those for absence sites (a value of 1 indicating complete discrimination). Values less than 0.5 indicate that predictions for presence sites are, on average, lower than those for absence sites (a poor result).

The discriminatory performance of a model, as measured by the Mann-Whitney statistic, can also be depicted graphically, either by plotting the frequency distributions of predictions for presence sites and absence sites, or by deriving and plotting an ROC curve (see examples in Figure 6.1).

6.4.2 Measure of overall accuracy: % Deviance explained

For presence/absence models predicting probability of occurrence, an overall measure of the accuracy with which a model predicts observed presences and absences at evaluation sites is the percentage of null deviance ‘explained’ by the model (Yee and Mitchell 1991), calculated as:

$$\frac{NullDev - ResDev}{NullDev} \times 100\%$$

where $NullDev$ = null deviance of the evaluation data

$ResDev$ = residual deviance of the evaluation data in relation to probabilities predicted by the model

Deviance is calculated as (Hosmer and Lemeshow 1989):

$$-2 \sum_i \{y_i \log(\mu_i) + (1 - y_i) \log(1 - \mu_i)\}$$

where y_i = observed presence (1) or absence(0) at evaluation site i

μ_i = predicted probability of presence at evaluation site i

To calculate null deviance, μ in this formula is set as the proportion of presences in the evaluation data and is therefore constant across all sites.

Percent deviance explained is a measure of overall goodness-of-fit of a model to actual observations. This goodness-of-fit can also be depicted graphically by grouping evaluation sites into classes of predicted probability and plotting observed proportions against predicted probabilities for these classes (see example in Figure 6.2).

6.5 Software

All software for calculating the measures of performance (Mann-Whitney U statistic and % Deviance explained) was implemented in the Microsoft Windows version of the S-PLUS statistical package (Statistical Sciences Inc. 1994).

Most of the software for effecting jackknifing of presence/absence models was also developed in S-PLUS supplemented by FORTRAN-77 programs, developed using the Lahey compiler (Lahey Computer Systems Inc. 1992), for transferring data between S-PLUS and various modelling software. Details of software used to implement specific modelling algorithms (e.g. GLM, BIOCLIM) are provided in Section 7.4.

6.6 Discussion

The techniques described in this chapter for evaluating performance of models have been applied to selected modelling techniques using data from north east NSW (see Section 7). The approach is sufficiently generic to be applied to other modelling techniques and/or other regions.

Many refinements could be made to the techniques including:

- A full ROC analysis of discrimination performance could be used in place of the approximate analysis based on the Mann-Whitney U statistic.
- The goodness-of-fit statistic used to measure the overall accuracy of probabilistic models could be decomposed to assess individual components of accuracy using the method proposed by Cox (1958) and Miller *et al.* (1991).
- Statistical significance testing could be applied to the accuracy measures using tests such as those proposed by Hanley and McNeil (1982) and Metz (1986) for ROC statistics and by Cox (1958) and Miller *et al.* (1991) for decomposed goodness-of-fit statistics.
- More emphasis could be placed on the use of independent field survey data for testing models as a more rigorous alternative to statistical resampling procedures.

Most of these refinements have already been incorporated and applied in a separate consultancy conducted by NSW NPWS for ANCA (Pearce and Ferrier 1996).

The current consultancy has evaluated models in terms of their accuracy, i.e. how accurately a model can predict the probability or relative likelihood of a species occurring at sites not used to derive the model. Accuracy is only one of a number of criteria that can, or should, be used to assess the performance of predictive models of species distributions. Other criteria that deserve consideration, but were beyond the scope of this consultancy, include:

- interpretability or comprehensibility of derived models;
- computational efficiency of modelling techniques;
- skill level required to apply modelling techniques;

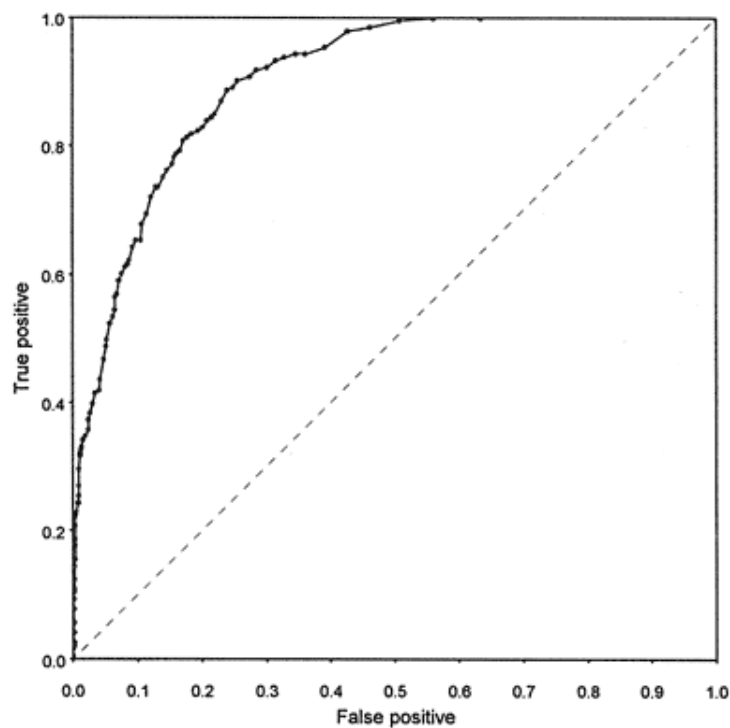
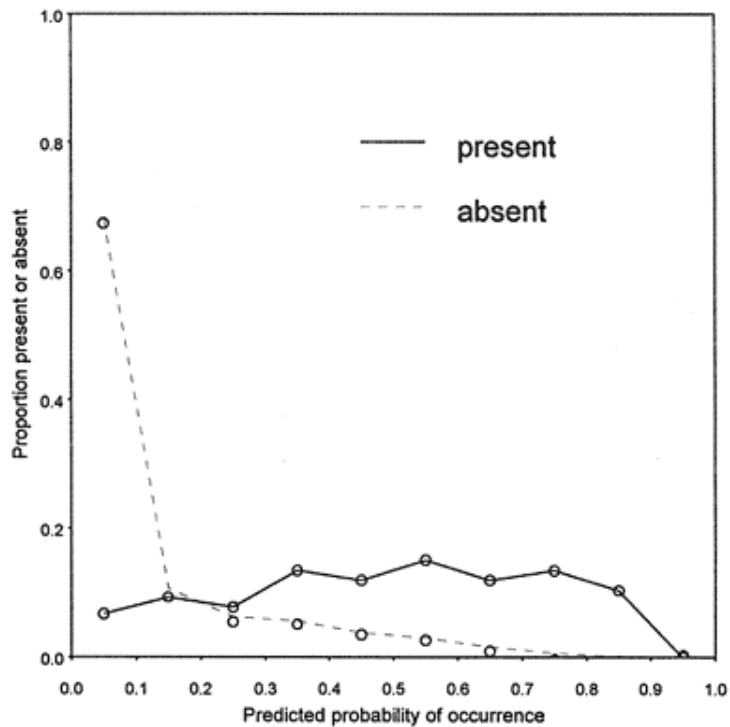


Figure 6.1 Examples of two graphical techniques for evaluating the discriminatory performance of predictive models. In the top graph two probability distributions are plotted, one for evaluation sites where the species was absent. Each point indicates the proportion of sites in either the present or absent group that fall within an interval of predicted probability (in this case ten 0.1 intervals). The bottom graph is an ROC curve plotting the relationship between true positive and false positive fraction for a continuous range of predicted probabilities (see Swets (1998) for details). A model with good discrimination should display clear separation of the two distributions in the top graph and a steep ROC curve in the bottom graph (a curve following the 45° line is no better than random).

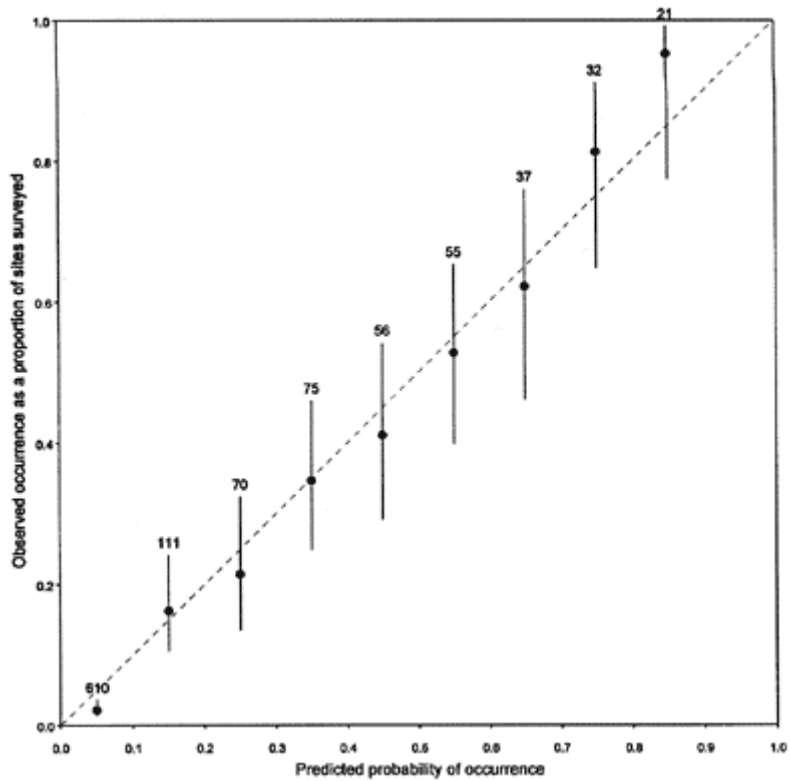


Figure 6.2 Example of a graphical technique for evaluating the overall accuracy (goodness-of-fit) of a presence/absence model. Evaluation sites are divided into classes according to predicted probability of occurrence (in this case ten 0.1 interval classes). The proportion of sites within each class at which the species was recorded as present is then plotted as a point. Ideally these points should lie along the bottom line. The vertical bars indicate the number of sites within each probability class.

- capacity for modelling techniques to estimate uncertainty in predictions and to communicate this uncertainty using confidence limits or similar tools;
- ecological rationality of modelled responses of species to environmental gradients in terms of current ecological theory; and
- ability of models to predict distributions beyond the geographical and/or environmental space of the region under consideration.

Most of these criteria have been considered in a recent consultancy conducted by CSIRO Division of Wildlife and Ecology for ANCA and ERIN (Austin and Meyers 1995; Austin *et al.* 1995). The CSIRO work also differed from the current consultancy in focusing more on using artificial data generated from theoretical ecological models, rather than real survey data, to evaluate performance of predictive modelling techniques.