*Journal of Applied Ecology* 2006
**43**, 413–423

REVIEW

# Error and uncertainty in habitat models

SIMON BARRY* and JANE ELITH†

**Bureau of Rural Sciences, Australian Government Department of Agriculture Fisheries and Forestry, GPO Box 858, Canberra 2601, ACT, Australia; and †School of Botany, The University of Melbourne, Parkville 3010, Victoria, Australia*

**Summary**

**1.** Species distribution models (habitat models) relate the occurrence or abundance of a species to environmental and/or geographical predictors that then allow predictions to be mapped across an entire region. These models are used in a range of policy settings such as managing greenhouse gases, biosecurity threats and conservation planning. Prediction errors are almost ubiquitous in habitat models. An understanding of the source, magnitude and pattern of these errors is essential if the models are to be used transparently in decision making.

**2.** This study considered the sources of errors in habitat models. It divided them into two main classes, error resulting from data deficiencies and error introduced by the specification of the model. Common and important data errors included missing covariates, and samples of species' occurrences that were small, biased or lack absences. These affected the types of models that could be developed and the probable errors that would occur. Almost all models had missing covariates, and this introduced significant spatial correlation in the errors of the analysis.

**3.** A challenging aspect of modelling is that species' distributions are affected by processes operating in both environmental and geographical space. We differentiated between global (aspatial) and local (spatial) errors, and discussed how they arise and what can be done to alleviate their effects.

**4.** *Synthesis and applications.* This study brings together statistical and ecological thinking to consider the appropriate techniques for habitat modelling. Ecological theory suggests models capable of defining optima, while allowing for interactions between variables. Statistical considerations, including impacts of data errors, suggest models that deal with multimodality and discontinuity in response surfaces. Models are typically simple approximations of the true probability surface. We suggest the use of flexible regression techniques, and explain what makes such methods superior for ecological modelling. The most robust modelling approaches are likely to be those in which care is taken to match the model with knowledge of ecology, and in which each is allowed to inform the other.

*Key-words*: climate envelope, error, evaluation, model, regression, response surface, spatial autocorrelation, species distribution, uncertainty

*Journal of Applied Ecology* (2006) **43**, 413–423
doi: 10.1111/j.1365-2664.2006.01136.x

## Introduction

Predictions of species distributions are important for a range of land management activities. Examples include management of threatened species and communities, risk assessment of non-native species in new environments, and estimation of the likely magnitude of biological responses to environmental changes such as global warming (Ferrier *et al*. 2002).

However, despite the wide use of predictive models, many applications give insufficient consideration to model error and uncertainty. Models are an attempt to summarize complex distributional patterns with a reduced set of predictor variables, and will inevitably

Correspondence: Jane Elith, School of Botany, The University of Melbourne, Parkville 3010, Victoria, Australia (fax + 61 39347 5460; e-mail j.elith@unimelb.edu.au).

contain some degree of mismatch between their predictions and the actual distributional patterns they describe. Here we use 'error' and 'uncertainty' interchangeably, understanding that 'error' includes not only 'mistakes' and 'faults' but also the statistical concept of 'variation'.

While error analysis could be considered a purely statistical problem, this narrow view fails to address the question from an ecological perspective. Rather, the adequacy of models relies on the interplay of the ecological processes driving the true distribution and the process used to observe and model it. Understanding these issues therefore requires an understanding of the joint impact of a species' ecology and the measurement and prediction process.

The arguments made in this paper represent a new analysis of the species prediction problem. Previous attempts to identify sources of modelling error have followed three main approaches. The first class, of theoretically based papers (e.g. Austin 1976), focused on the response of species to environmental gradients; while the arguments were based on observations, their empirical basis was not formalized. For example, the relationship between response and environment was considered abstractly, and not related to productivity, growth rate or presence and absence. The second class of papers compared the performance of different models when used to analyse particular data sets (e.g. Manel, Dias & Ormerod 1999; Moisen & Frescino 2002). These studies were useful but limited because the adequacy of a particular model depends on the complexity of the response surface, making results difficult to generalize. Further, many such studies do not use independent data for evaluation, and therefore cannot address the impact of data error. The third class of papers has attempted to reconcile theory and observation (Austin 1980; Leathwick & Whitehead 2001). These papers are the precursors of this work but, while they focus on the ecological rigour of the modelling process, they provide minimal consideration of statistical rigour. This contribution provides the statistical context.

The ultimate impacts of errors on models typically depend on the context within which the model is being fitted and used. In this study we sought to define more accurately the nature of errors in species distribution models, review and discuss how these errors occur, and explore the key issues leading to the success and failure of common species prediction techniques. In this paper, we first describe the species prediction problem and outline modelling techniques and the nature of the responses that ecologists are trying to model. We then define two different types of error, global and local, and use them as a framework for understanding the errors that occur as the result of data errors and model misspecification. Finally, we make recommendations about techniques to control and diagnose errors.


*Journal of Applied Ecology*, **43**, 413–423

## The species prediction problem

A wide variety of approaches is currently used for the prediction of species' distributions (Guisan & Zimmerman 2000; Phillips, Anderson & Schapire 2006). In this paper, we restrict our attention to a subclass of methods in which a model is used to relate the presence or abundance of a species to some set of functionally relevant predictors. In most studies this involves use of survey data describing the distribution of a species, and associated environmental data describing factors that are either known to have a direct impact on the species or are correlated with variables that do. Typically, environmental data are known for the area in which the model is to be used to make predictions and defined over a grid or lattice of points.

Given that this correlative approach is the focus of most practical systems, we restrict our attention to statistical models. Here we use the word 'statistical' in a very broad sense to refer to those methods that consider the prediction problem by conditioning on the known environmental information for the site, and inferring the presence or absence of the species from relationships in the sampled data. More specifically, the approaches we consider are: envelope approaches, distance-based approaches and regression models. The following paragraphs describe these more fully.

Envelope approaches (e.g. BIOCLIM; Busby 1991) use presence records and environmental data to form a profile for a species that summarizes how the known presences are distributed with respect to the environmental variables. With several environmental variables, the aggregated profile forms a multidimensional space (a hyper-rectangle or 'environmental envelope') that defines the environmental domain of the species. This envelope specifies the model in terms of upper and lower tolerances, and does not allow for regions of absence (i.e. 'holes') within the envelope. The concept is one of extremes and cores. A habitat map can be produced from the model by ranking each location according to its position in the species' environmental profile. Commonly these maps are grid-based and classify each cell into one of several ranked classes of environmental suitability for the species.

Distance-based measures (e.g. DOMAIN; Carpenter, Gillison & Winter 1993) estimate the environmental distance between a site of interest and the nearest presence record in environmental space. DOMAIN uses the Gower metric, a distance measure that standardizes each variable by its range over all presence sites to equalize the contribution of all variables. The scaling means that one unit in any direction in environmental space gives an equal change in similarity. A different distance metric (e.g. Euclidean distance) would give a different model of change. At a conceptual level, distance-based methods differ from envelope approaches because they focus on distances from adjacent sites of known occurrence rather than defining bounding envelopes that enclose all sites of known occurrence.

Regression models comprise a broad class of methods that include generalized linear models (GLM), generalized additive models (GAM), decision trees and multivariate adaptive regression splines (MARS) (Hastie,

Tibshirani & Friedman 2001). The unifying theme in all these methods is that, in a univariate setting, they fit a curve through a set of points using some goodness-of-fit criterion. There are numerous variations: the response can be calculated with respect to many variables; the response between a species and variable can be linear or non-linear; the models can be purely additive or include interactions between predictor variables. Depending on the method used, the response could be modelled as one with no optimum, or with single or multiple optima. For the purpose of exploring the characteristics of errors in models, we consider regression models as a broad class but highlight where particular regression methods have strengths or limitations.

## Model realism

While ecological theory and observation suggest that there should be non-random associations between environmental variables and species presence and absence or abundance, the ecological processes leading to the associations are generally complex. Both abiotic and biotic factors may influence the distribution and abundance of a species, and different processes may dominate in different parts of its range. For example, the limit of distribution of a plant at some point in environmental space may be determined by physiological responses to environment, but these responses are likely to be influenced, perhaps strongly, by interactions with other species (Leathwick & Austin 2001). It is therefore important not to be overly optimistic about the ability of purely environment-based predictions to recover the nature and pattern of abundance for most species. Further, the pattern of presence and absence of a species depends in part on the scale at which measurements are made and how these are related to the spatial scales over which variation occurs in the processes determining the distribution. For these reasons, most techniques used to model species' distributions rely on correlative approaches rather than mechanistic models. Even though the processes driving a species' relationship with an environmental gradient may vary, the correlative approach can still characterize, in a statistical sense, the complex response of a species to these different processes, as reflected in its distribution.

All species distribution models assume that the species is at equilibrium with the environment. Observed patterns are assumed to reflect the species' full biotic potential, implying that the species can potentially occur in all environmentally suitable locations, and its distribution has not been constrained locally by factors such as historic accidents. Whether this is plausible depends on the scale of the model, the dispersal ability of the organisms and the history and biology of the species (Pulliam 2000; Tyre, Possingham & Lindenmayer 2001). In practice, errors resulting from the equilibrium assumption are most acute when trying to predict distributions of species recently introduced to new locations. For clarity, we assume initially in this paper that some degree of equilibrium exists, but return to the issue of the mis-specification of a global equilibrium later.

## Global and local errors

Having fitted a model, the next logical step is to evaluate its fit to the modelling data, and/or its predictive ability. A range of measures of error can assess the discrepancy between data and model, but the end-use of the model dictates the most relevant measures and data sets for evaluation. If using a technique that produces predictions of presence or absence (rather than continuous estimates of probability), omission and commission errors or other statistics derived from a confusion matrix are appropriate (Fielding & Bell 1997). The confusion-matrix approach has two limitations in broader settings. First, an arbitrary choice of threshold is required in converting probabilistic predictions to binary ones. We argue that it is better to consider the discrepancy between the model inferences (such as predicted probability of occurrences) and the actual observations. This leads to statistics such as the deviance of the model (Hastie, Tibshirani & Friedman 2001), the area under the receiver-operating characteristic (ROC) curve (Hanley & McNeil 1982), Miller's calibration equations (Miller, Hui & Tierney 1991), and correlation tests (Zheng & Agresti 2000). McCarthy *et al.* (2001) give a clear account of the problems in ignoring the probabilistic nature of predictions and explore the performance of several statistics for testing the accuracy of population models. MacKenzie & Bailey (2004) present relevant tests of model fit for imperfectly detected species.

The second limitation of confusion matrices, which also applies to other measures of fit, is that they are only marginally relevant to the model and ignore the geographical pattern of the predictions. In particular, measures of fit do not identify where and how the errors occur, either spatially or environmentally. Others have recognized this problem and suggested evaluations that take into account the spatial context of the errors (Fielding & Bell 1997). An analyst may want the predictions from their model to be 'realistic' at each geographical location in the study area, or may simply be content with similar levels of realism at all geographical locations sharing some common set of environmental attributes. But where spatial processes such as dispersal or disturbance play an important role, spatial terms or other adjustments may be necessary in our models to make them predict accurately in geographical space. Legendre (1993) provides a clear description of the important issues concerning spatial autocorrelation and the partitioning of variation into its environmental and spatial components.

A more intuitive way of expressing this is to say that it is often desirable to have the outcomes of predictions interpretable 'locally' in a geographical sense. This thinking leads to a distinction between global (aspatial) vs. local (spatial) errors. Global error impacts at all environmentally similar locations in a similar way,

regardless of their location. For example, the model may consistently overpredict for a certain environment at all locations. Alternately, local error has different effects at environmentally similar sites, depending on their locations.

## Sources of modelling error

Species distribution models contain errors arising from (i) deficiencies in the data and (ii) deficiencies in their ecological realism. In practical terms, the overall error in prediction results from the combination and interaction of these two components. We consider each of these sources in turn.

### DATA ERROR

In an ideal world modelling data would consist of a representative and accurate sample of the species' response, and a set of covariates such that a model could be specified that had quantifiable and bounded local and global error. In practice this is rarely achieved. We describe five of the most common sources of data error, exploring their particular characteristics. We analyse from a theoretical perspective how different statistical methods will respond to these errors, and from a more pragmatic perspective what refinements might provide a coherent solution to the problems that they cause.

#### *Missing covariates*

At least some predictor variables are missing from most models (i.e. limited covariates), reflecting lack of knowledge of which environmental factors constrain the distribution of a species throughout its range, and lack of spatial data sets describing attributes known to be important. A 'sufficient' set of covariates may be defined as that which allows a model to be specified that does not have significant spatial errors or global errors, with respect to a specific context and end use.

Predictor variables are incomplete in time and space, over the range of scales at which processes operate, and in relation to the suite of species' interactions and historical and current disturbances that affect a species' occurrence (Van Horne 1983; Levin 1992). Further, even for mechanisms that are relatively well understood, directly relevant quantitative data that can be used for modelling are usually unavailable. Rather than the proximal variable that directly affects the distribution of the species, the most common data quantify indirect (distal) variables that are correlated with the causal ones, to varying degrees (Austin 2002). For example, average temperature at a location is typically correlated with a set of more proximate temperature factors occurring at the site, including cumulative heating inputs (growing degree days), within-year or seasonal variation, and extreme maximum or minimum temperatures. These have lethal effects on some species, and may be important at decadal or even longer intervals. It is rare to have information on these causal variables and instead modelling is done with the indirect ones.

No studies claim to use a comprehensive suite of proximate predictors. Even with all covariates known prediction is not perfect because of demographic variation (Tyre, Possingham & Lindenmayer 2001). In practice, ecological thinking should be used to construct the most appropriate variables from the data at hand, but lack of relevant data can be beyond the control of the practitioner.

#### *Small sample size*

Small sample size is an error in the sense that the sample provides an insufficient basis for modelling. While statistical approaches may be able to characterize this error in simple cases, we focus here on problems caused by the sample being insufficient to specify all but the simplest models. For presence–absence data, sample size needs to be assessed in relation to the least frequent class rather than a simple count of total number of sites. A species may be genuinely rare, and a random sample of 500 sites may contain only five presence records. While this sample is not biased or incorrect and may be relatively large, it is nevertheless likely to be an inadequate sample for most modelling methods, because five presence records are too few to specify the model properly. This may seem obvious, but there are many ecological data sets with large numbers of sites but few presence records for many species (Ferrier 2002), and modellers need advice on the limitations of their methods where this occurs. The minimum number of records required for a method depends partly on the complexity of the pattern being modelled. In general, the broader the suite of explanatory variables and the more complex the responses (in terms of shape and interaction), the more data are required to construct a reliable model.

Models that include spatial autocorrelation terms need intense local sampling in at least part of the species range, and cannot be estimated accurately with small or unevenly distributed samples. Sparse data can lead to specifying simpler models that typically involve both global and local specification errors.

Several modelling methods have been suggested for small numbers of presence records. With very limited species records, one option is to create a habitat suitability index model (HSI; USFWS 1980). The method is based on the judgements of experts who identify critical variables that can be used to identify suitable habitat through a conceptual model of how the species responds to environment. It is difficult to test HSI models because usually there are no data for evaluation, but recent evaluations demonstrate varied outcomes, with some successes (Mitchell, Zimmerman & Powell 2002) and some evidence of poor predictive performance (Guay *et al.* 2003; Loukmas & Halbrook 2001). Nevertheless, HSI models are more useful than no model, with the advantage that they quantify expert opinion and provide a basis for ongoing discussion and refinement. Information

about uncertainty can be represented as intervals or fuzzy numbers, resulting in bounds (Burgman *et al*. 2001).

Another option is to use a statistical model, restricting the number of candidate variables to those that can be supported numerically. For example, rules of thumb for logistic regression suggest 10 records in the least prevalent class per predictor degree of freedom (Wintle, Elith & Potts 2005). So, for 20 presence records, coefficients for two linear or one quadratic response can be properly estimated. However, restricting a model to few predictor variables averages the response over all the omitted variables, and may result in a misleading model.

Community models are a third alternative. These can be used to model either the collective properties of the biota (Ferrier 2002) or to make predictions for individual species from a community model in which information for a wider set of species is used to construct a context in which individual species distributions are then described. Methods include generalized dissimilarity modelling (Ferrier *et al*. 2002), neural networks (Olden 2003), and multivariate adaptive regression splines (Leathwick *et al*. 2005).

### *Biased samples*

The ideal data for modelling are collected using a planned sampling regime, structured to sample the major environmental gradients likely to be important for the species, and covering the spatial extent of the region of interest (Austin & Heyligers 1989; Cawsey, Austin & Baker 2002). However, species data available for modelling are often not specifically collected for the purpose, and instead may consist of an *ad hoc* collection of existing data, biased in geographical and/or environmental space. For instance, sampling is often more frequent close to roads (Kadmon, Farber & Danin 2004) and population centres, focused on particular landscapes or vegetation types, and/or biased away from ecotones.

Biases in data mean that the modelled relationships are dominated by the patterns at sampled sites rather than the patterns across the entire study area, and this in turn is likely to lead to marked spatial variation in prediction uncertainty, i.e. to spatial error.

Such biases can be diagnosed by exploratory techniques. Examples include making plots of site locations in geographical space, analysing site density in environmental strata (Wintle, Elith & Potts 2005) and using statistics such as p-medians (Faith & Walker 1996) to measure distances between sets of sites in multivariate environmental space. Biased data sets can be supplemented with new data from targeted surveys that aim to increase representation from poorly sampled regions (Cawsey, Austin & Baker 2002). Implications of bias on model performance are discussed by Kadmon, Farber & Danin (2004).

Samples can also be biased in relation to the observations. Mobile and cryptic species are difficult to detect, and tend to be underestimated by common field survey techniques. Repeat surveys can be used to quantify detectability (MacKenzie & Royle 2005). If not adequately dealt with, false-negative observations are likely to affect both variable selection and coefficient estimation in models (MacKenzie *et al*. 2002; Tyre *et al*. 2003).

### *Lack of absence records*

Lack of absence records is a data error in the sense that it limits the creation of models that accurately discriminate between suitable and unsuitable habitats. In particular, lack of absence data leads to inaccurate identification of the attributes of unsuitable sites. Data sets without absence records (presence-only data) are common, particularly in natural history collections (NHC), *ad hoc* collections of field observations (Graham *et al*. 2004) and distribution data used in biosecurity risk analysis (Panetta & Mitchell 1991). Because NHC data cover many species, have large numbers of records and are accessible through searchable World Wide Web databases, they are used frequently in distributional modelling (Anderson 2003). These types of presence-only samples often have other errors associated with them, such as small sample sizes and bias (Graham *et al*. 2004). An alternative source of presence-only data is radio-telemetry records (Keating & Cherry 2004; Rushton, Ormerod & Kerby 2004) and these are less error-prone because they tend to be structured samples with precise geo-locations.

Presence-only data are commonly modelled with methods that only use the presence data (e.g. climate envelopes and DOMAIN) or methods that characterize the background, either as a space over which to model (Manly *et al*. 2002; Phillips, Anderson & Schapire 2006) or using a sample in place of absence records (e.g. regression methods). The latter are sometimes treated as a surrogate for absences and called pseudo-absences (Zaniewski, Lehmann & Overton 2002). Analyses of predictive performance have shown that regression models developed with presence-only data may have reasonable predictive performance, although usually poorer than equivalent models fitted with true presence–absence data (Ferrier & Watson 1997).

The absence of true zeros in these models results in different response shapes than those fitted with presence–absence data, resulting in both global and local errors. In many cases, the fitted values, rather than reflecting the true prevalence of the species, will simply reflect the balance between the presences and the constructed absences. For example, increasing the number of absences will reduce the average of the fitted probabilities. The most statistically accurate way to model with pseudo-absence data is through the use of specialized models such as case control (Keating & Cherry 2004; Pearce and Boyce 2006) and resource selection functions (Manly *et al*. 2002).

### *Errors in variables*

Predictor variables also have errors, and these errors can be random, biased or spatially aggregated. For example,

some variables are interpolated from point data and will have errors consistent with those of the interpolation method and the quality of the original point data. Errors in digital elevation models can be globally small and unbiased but locally large and spatially autocorrelated (Holmes, Chadwick & Kyriakidis 2000). In variables describing vegetation classes or soil types mapped as polygons, the location of the polygon boundary is uncertain, as is the width of the ecotone (Fortin & Edwards 2001). Biases are often also linked to the spatial scale. For example, as the grain at which data are recorded becomes coarser, units that exist at a finer grain may be subsumed into more prevalent ones, leading to a bias against unusual classes (e.g. rare vegetation classes).

The problem of such errors in predictor variables can become overwhelming, and a common reaction is to ignore them, an approach that can have some justification in a statistical sense. In cases where the prediction sites have the same errors as those used for model building, the model will already reflect the errors and the predictions will be consistent with the data. For regression models, random error in the predictor variables will be reflected in the width of the confidence intervals. However, a number of methods are available for improving coefficient and error estimates (measurement error models; Reeves *et al*. 1998)

But even where such errors are ignored, difficulties can arise. For example, strong spatial patterns in predictor variable errors will inevitably produce local prediction errors. Sensitivity analyses can help to explore the extent, location and impact on final predictions of such problems (Van Niel, Laffan & Lees 2004). More importantly, if the data used for prediction have an error structure differing from that inherent in the modelling data, then serious prediction errors may result. For example, the relationship between proximal and distal variables may change significantly with location (Austin 2002). Subtle errors related to this phenomenon are particularly likely when predictions are made from a model describing a natural distribution in one location to some new region, as in pest risk assessment. Typically, the relationship between the measured variables and the true underlying processes will vary significantly over large distances and errors will result.

## MODEL SPECIFICATION ERRORS

With a well-specified model and an adequately large random sample from the population, the statistical component of error is easily quantified and incorporated into predictions. But in most practical settings, the data sample is rarely random and the model not fully adequate, and these factors combine to induce error in the final predictions. We consider these sources of error in the following sections.

### *Does our model approach the true model?*

To consider the impacts of specification error on the model it is useful to review briefly the statistical theory describing the process of fitting models to data. Most studies conclude that, provided the 'true' model is nested within the model specification, the model estimated using maximum likelihood or other consistent techniques will converge to the true model as the sample size increases (Welsh 1996). As an example, if the relationship between the conditional mean of $y$ given $x$ is linear, regression using a linear model will be arbitrarily close to the population values if the sample is sufficiently large.

If the 'true' relationship is not contained in the model, then over- and underestimation will typically result in different parts of the covariate space. For example, if the true relationship is quadratic and the fitted model is linear then errors are inevitable, will not be corrected by an increased sample size and will lead to errors in inference and prediction.

### *How do we estimate the response surface?*

With the assumption of equilibrium described earlier, we consider the possible response surface shapes in environmental space as maps showing how the probability of presence of a species varies with the environmental variables. As an example, consider the response surface shown in the top panel of Fig. 1. This plots a hypothetical example where probability of occurrence depends on rainfall and temperature, and there is a significant interaction between them. Note that there are still marginal relationships between the variables and the response (lower panels; Fig. 1). In this simple case we can try to model the joint distribution in a number of ways. Heuristically, we can attempt to estimate the full response surface using some technique. Alternatively, it can be approximated by using functions of the marginal relationships.

Practical situations are more complicated. In fitting a model, we are trying to estimate the response surface that gives the best spatial predictions. In its most complex form, this is a $k$-dimensional surface where $k$ is the number of predictor variables. In practical terms, however, estimation of such surfaces for values of $k > 2$ (i.e. three-way and above interactions) is difficult given the sizes of data sets and species prevalence typically available in many studies (100–1000 observations), because the number of parameters is prohibitive given the sparseness of the data.

Different techniques approach the problem of trying to estimate a complex surface with limited data in different ways, but all try to approximate the surface by simple components. In climate envelope approaches, the shape of the response surface is mostly ignored. The use of percentiles implies a belief in core and non-core habitat, and presupposes unimodality of response pattern to a gradient. Envelope approaches seek to delineate the non-zero components of the response along each gradient. The BIOCLIM model assumes that this region is rectilinear and orientated with the environmental axes. Distance-based approaches use

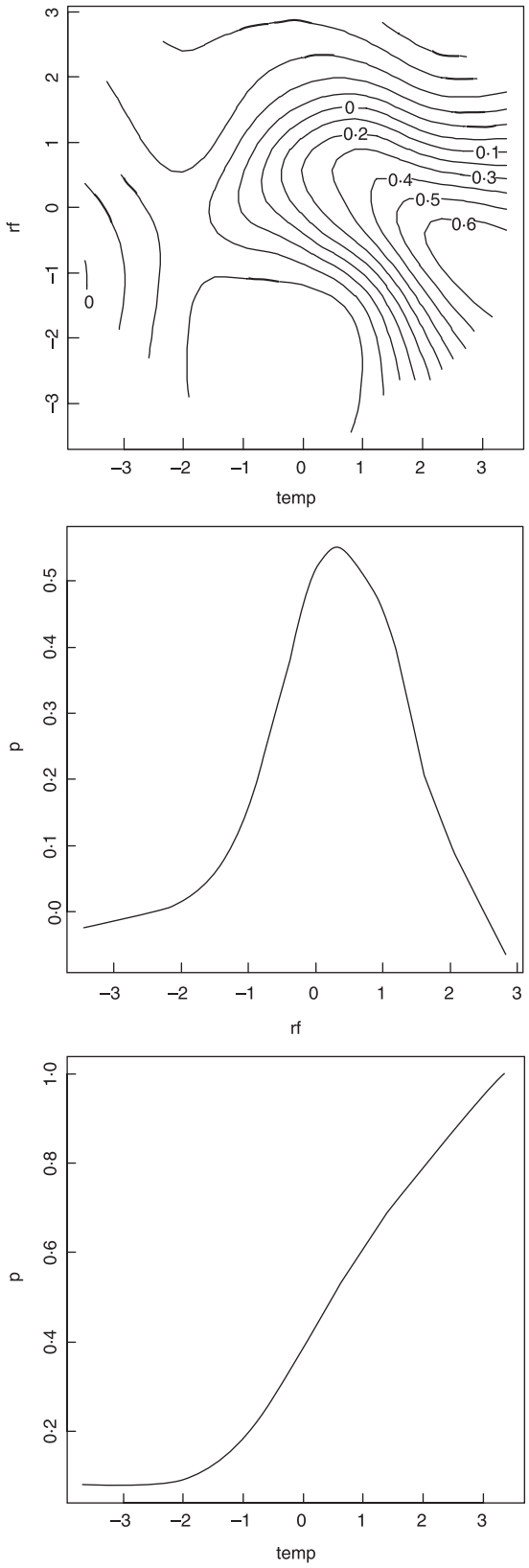


**Fig. 1.** A response surface (top) showing the response of the species to rainfall (rf) and temperature (temp). The lower two panels show its marginal relationships (y axis = probability of presence, p).

observations close to the point that is being predicted to assess suitability. Thus they attempt to model the response surface as a local smoothing. This non-parametric element of their construction is one of their attractions.

Regression approaches model the response surface parametrically in GLM, via smooth terms in GAM and via step functions with regression trees. With limited data, the number of parameters that can be estimated is limited, so the problem becomes one of finding a parameterization that produces a reasonable approximation to reality. This can be a complicated undertaking. Ecological theory supports several beliefs about the nature of species' response to gradients. The idea that species may have optima along environmental gradients is well accepted (Austin, Nicholls & Margules 1990; Austin 2002). It is plausible that in many circumstances a species' response to a continuous covariate is typically smooth (i.e. a small change in the covariate produces a small change in the response). Often, significant interactions will exist between variables for physiological reasons.

Unfortunately, this simple analysis of the nature of response curves is not particularly useful in practice. Errors in data complicate the picture. Missing covariates can produce discontinuities in the apparent surface. They can also produce multimodalities and other unexpected effects. To understand this, consider a missing covariate that has a strong impact on a species' distribution. For example, the missing covariate may be a soil classification, and the species may not survive if the soil is a particular type. If this unsuitable soil type is correlated with a certain range of the available covariates, the species may exhibit a multimodal response. In the most extreme case, when the species cannot grow within a restricted range of the covariate space, a discontinuity may result. The result is that the response surface that we have to model with the available covariates is more complex than the simple surfaces implied by abstract vegetation theory.

These observations give an interpretation for why flexible regression-based techniques that can fit complex surfaces, for example GAM and boosted regression trees (Friedman, Hastie & Tibshirani 2000), generally give improved model performance. A range of information–theoretic (Burnham & Anderson 2002) and resampling (Hastie, Tibshirani & Friedman 2001) approaches can be used to control overfitting and to balance effectively the trade-off between model complexity and predictive power. Even so, errors in model specification are essentially ubiquitous and the majority of models are typically simpler than the real-world complexities they seek to describe. The impact of this simplification will of course depend on the extent and magnitude of the discrepancies. If a species is in equilibrium with the environment and all 'significant' variables are included in the analysis, then the failure of the specification leads to over- and underestimation of the response at different points in the covariate space. This

is the same as fitting a line to a quadratic response. Assessment of the magnitude of these errors can be achieved by graphical display of the distribution of model residuals in both environmental and geographical space.

Techniques vary widely in their ability to model non-linearities and interactions (Moisen & Frescino 2002; Leathwick *et al*. 2006) but use of flexible techniques does not guarantee robust model fitting. Unfortunately this also requires a level of judgement that is generally only obtained through familiarity with a technique and with the ecology of the species (Austin, Nicholls & Margules 1990). Burgman, Lindenmayer & Elith (2005) call this modelling frame uncertainty.

## INTERACTIONS BETWEEN DATA ERRORS AND MODEL MIS-SPECIFICATION

Interactions between data and model errors can be illustrated through the impacts of missing covariates on model robustness. In this context a missing covariate is a variable that would provide additional predictive power if it was known and observed, for example soil attributes, which can have a major impact on vegetation but which are often poorly described. More subtly, the covariates in most models are typically coarse correlates with the more proximate factors that control distributions (Austin 2002). The use of such a correlated variable assumes that the correlation structure between the fitted predictor and its more proximate components is stable throughout the sampling domain. Departures from this will result in spatially correlated errors and, given the difficulties in assessing errors with binary point data, there is a danger that small global errors may lead the analyst to overlook the magnitude of local errors. Significant spatial patterning in the residuals, i.e. large local errors, are also likely if missing covariates have a non-random spatial distribution, a feature commonly observed, for example, in the distributions of soil attributes. A simple indicative example is shown in Fig. 2 and we summarize the impacts in Table 1.

This spatial patterning effect has been noted by numerous authors and is typically described in terms of spatial autocorrelation in the regression residuals. Missing covariates are only one source of spatially autocorrelated residuals (Legendre 1993). Typical solutions to the presence of spatially autocorrelated residuals include the use of more complicated models, such as autologistic regression (Augustin, Mugglestone & Buckland 1996) with geographical space as a covariate, or use of techniques such as geographically weighted regression

**Table 1.** Summary of types of model error and impact of missing covariates

| Model | Model form error | Impact of missing covariates | Modelling recommendation |
|---|---|---|---|
| BIOCLIM | High. Assumes independent rectilinear bounds and that all variables are known Will cause overprediction with few variables and underprediction with many variables | Increase area predicted introducing spurious predictions | |
| Distance-based | Medium. Estimates model non-parametrically but difficulties arise with data density and the definition of distance | Algorithm will not choose the appropriate data points as being 'close' and biases will result | Work needs to be performed to assess best distance measure. Cross validation? |
| Regression | High–low. Flexible techniques exists such as GAM and boosting. Simple model may suffer from considerable specification bias | Spatial correlation in residuals | Use flexible models unless clear theoretical reasons to ignore. Consider spatial patterns of errors to diagnose models problems. Truncate response range |

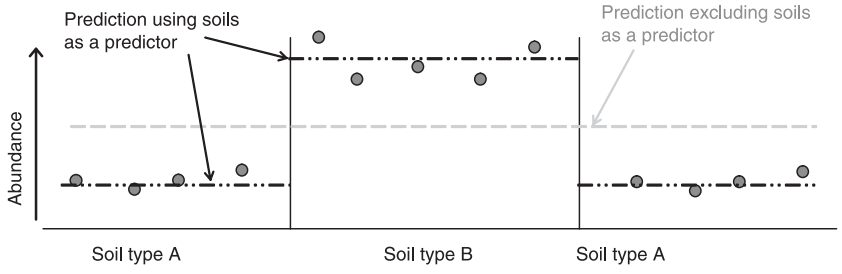


**Fig. 2.** An illustration of the impact of a missing covariate on modelled predictions of species abundance. The *x* axis is in geographical space and circles represent the observations. When the covariate is missing, predictions are averaged across both soil types.

(Fotheringham, Brunsdon & Charlton 2002). While these approaches are designed to detect spatial variation in the relationship between response and predictor variables, the danger is that such variation may be fitted to compensate for a missing predictor with strong spatial pattern. Use of these techniques therefore requires a clear understanding of the ecological processes that are being modelled, i.e. the effectiveness of these techniques must be assessed in the light of the ecological plausibility of the model that they are used to fit. In particular, while the incorporation of geographical space in the regression may be justified, it may also represent an *ad hoc* solution that lacks theoretical justification, which may therefore introduce more model error than it removes. Finally, the lack of common elements in the effects of missing variables makes the prescription of any general solutions difficult.

## Discussion

We have outlined common sources of error. Modelling data have frequent limitations: comprehensive, purposive sampling is rare; conservation often has to deal with data sets compiled from sources collected for other purposes; it is difficult to get good data for rare species; key environmental variables may be undescribed or even unknown. Further, there are inherent limitations to our ability to model species' distributions because they are so complex. Environment is clearly important in many settings, but species' responses to environment depend on the competitive context, and this in turn varies given the dynamic nature of species' distributions, the effects of natural and human disturbance, and the complicating effects of variation in the speed with which different species re-occupy sites from which they have been displaced.

The combined effect is that we are usually trying to model a complex response surface. This helps us understand why regression-based techniques such as GAM, MARS, boosted regression trees (Hastie, Tibshirani & Friedman 2001) and maximum entropy modelling (Phillips, Anderson & Schapire 2006) have, on average, improved performance compared with simpler methods. These techniques are more flexible and can better approximate the complex patterns seen in practice. For example, GAM can much more robustly describe the non-linear and/or skewed ecological responses that are typical of species distributions, compared with GLM (Yee & Mitchell 1991). MARS also has the flexibility to fit complex response shapes but does so efficiently with respect to the number of parameters used to fit the model, and is also capable of fitting local interactions (Leathwick *et al*. 2006). Boosted trees and other emerging methods are able to fit complex response surfaces and have improved ability to fit interactions in an efficient manner (Friedman & Meulman 2003). The simpler methods, such as climate envelopes, do not have the capacity to fit complex response surfaces and are not suitable for this task.

Perhaps the most challenging aspect is that distributions are affected by processes operating both in environmental and geographical space. Consistent responses can be observed in particular environments and these may occur repeatedly in many geographical locations. In contrast, other processes linked to disturbance and dispersal operate relatively independently of environment and may be strongly clustered geographically. This leads to our distinction between global error and local error. It emphasizes two issues. First, model evaluation needs to explore the spatial pattern of errors, and research on efficient methods to achieve this would be particularly useful. Some could be simple, for example mapping regression residuals or calculating ROC areas separately for different geographical areas within a study region. Secondly, the major modelling challenge is that apparent geographical patchiness may reflect either the effect of disturbance-related processes that are independent of environment, or missing environmental effects that are geographically patchy.

Given these complexities, the most robust modelling approaches are likely to be those in which care is taken to match the model with knowledge of ecology, and in which each is allowed to inform the other, i.e. models should be constrained to be congruent with ecological knowledge, with successive improvement in model specification that is driven by increasing knowledge of the ecology of the system (Leathwick & Whitehead 2001). Ecological understanding must also be able to be informed by the modelling outcome, recognizing the ability of models to raise new questions about ecology by elucidating subtle and/or complex relationships. In this latter respect, model misfit may be as informative as model fit.

When species distribution models are used in conservation and planning, an understanding of error can inform two broad paths of action (Edwards & Fortin 2001). The first views uncertainty as an obstacle that needs to be reduced or removed. This leads to actions directed at improving the data (removing errors, collecting more samples, refining the variable set) or changing the model structure (seeking more powerful modelling techniques). The second views uncertainty as a fact of life, a phenomenon that needs to be understood, characterized and sensibly factored into decision-making. Typical outcomes of this view include explorations of error, sensitivity analyses and decision strategies that aim to be robust to likely errors (Burgman, Lindenmayer & Elith 2005). Both perspectives are valid and not necessarily mutually exclusive, and in this paper we have given examples from each of them.

The practical implications of this paper are clear. If limited data are available, only simple models can be considered, but while these models avoid problems associated with overspecification they will almost always involve significant spatial error. If sufficient data are available, flexible regression-based techniques capable of fitting non-linear relationships should be used, as these allow for the complicated response surfaces that are frequently observed in distributional data. Careful

evaluation of the geographical and environmental deviations of a model from validation data will contribute substantially to the evaluation of ecological ideas and respecification of models.

## Acknowledgements

We discussed error and uncertainty at a workshop in Switzerland in 2004: 'GLM/GAM spatial modelling of species distribution'. Interactions with participants helped develop a number of these ideas. Thanks to John Leathwick, referees and editors, whose detailed comments improved the manuscript substantially.

## References

Anderson, R.P. (2003) Real vs artefactual absences in species distributions: tests for *Oryzomys albigularis* (Rodentia: Muridae) in Venezuala. *Journal of Biogeography*, **30**, 591–605.

Augustin, N.H., Mugglestone, M.A. & Buckland, S.T. (1996) An autologistic model for the spatial distribution of wildlife. *Journal of Applied Ecology*, **33**, 339–347.

Austin, M.P. (1976) On non-linear response models in ordination. *Vegetatio*, **33**, 33–41.

Austin, M.P. (1980) Searching for a model for use in vegetation analysis. *Vegetatio*, **42**, 11–21.

Austin, M.P. (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, **157**, 101–118.

Austin, M.P. & Heyligers, P.C. (1989) Vegetation survey design for conservation: gradsect sampling of forests in north-eastern NSW. *Biological Conservation*, **50**, 13–32.

Austin, M.P., Nicholls, A.O. & Margules, C.R. (1990) Measurement of the realized qualitative niche: environmental niches of five eucalypt species. *Ecological Monographs*, **60**, 161–177.

Burgman, M.A., Breininger, D.R., Duncan, B.W. & Ferson, S. (2001) Setting reliability bounds on habitat suitability indices. *Ecological Applications*, **11**, 70–78.

Burgman, M., Lindenmayer, D.B. & Elith, J. (2005) Managing landscapes for conservation under uncertainty. *Ecology*, **86**, 2007–2017.

Burnham, K.P. & Anderson, D.R. (2002) *Model Selection and Inference: A Practical Information–Theoretic Approach*, 2nd edn. Springer-Verlag, New York, NY.

Busby, J.R. (1991) BIOCLIM: a bioclimate analysis and prediction system. *Nature Conservation: Cost Effective Biological Surveys and Data Analysis* (eds C.R. Margules & M.P. Austin), pp. 64–68. CSIRO, Canberra, Australia.

Carpenter, G., Gillison, A.N. & Winter, J. (1993) DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation*, **2**, 667–680.

Cawsey, E.M., Austin, M.P. & Baker, B.L. (2002) Regional vegetation mapping in Australia: a case study in the practical use of statistical modelling. *Biodiversity and Conservation*, **11**, 2239–2274.

Edwards, G. & Fortin, M.J. (2001) A cognitive view of spatial uncertainty. *Spatial Uncertainty in Ecology: Implications for Remote Sensing and GIS Applications* (eds C.T. Hunsaker, M.F. Goodchild, M.A. Friedl & T.J. Case). Springer-Verlag, New York, NY.

Faith, D.P. & Walker, P.A. (1996) Environmental diversity: on the best possible use of surrogate data for assessing the relative biodiversity of sets of areas. *Biodiversity and Conservation*, **5**, 399–415.

Ferrier, S. (2002) Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Systematic Biology*, **51**, 331–363.

Ferrier, S. & Watson, G. (1997) *An Evaluation of the Effectiveness of Environmental Surrogates and Modelling Techniques in Predicting the Distribution of Biological Diversity*. Consultancy Report. NSW National Parks and Wildlife Service for Department of Environment, Sport and Territories, Environment Australia, Canberra, Australia.

Ferrier, S., Drielsma, M., Manion, G. & Watson, G. (2002) Extended statistical approaches to modelling spatial pattern in biodiversity: the north-east New South Wales experience. II. Community-level modelling. *Biodiversity and Conservation*, **11**, 2309–2338.

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

Fortin, M.J. & Edwards, G. (2001) Delineation and analysis of vegetation boundaries. *Spatial Uncertainty in Ecology* (ed. C.T. Hunsaker, M.F. Goodchild, M.A. Friedl, T.J. Case), pp. 158–174. Springer-Verlag, New York, NY.

Fotheringham, A.S., Brunsdon, C. & Charlton, M. (2002) *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. John Wiley and Sons, London, UK.

Friedman, J.H. & Meulman, J.J. (2003) Multiple additive regression trees with application in epidemiology. *Statistics in Medicine*, **22**, 1365–1381.

Friedman, J.H., Hastie, T. & Tibshirani, R. (2000) Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, **28**, 337–407.

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A.T. (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution*, **19**, 497–503.

Guay, J.C., Boisclair, D., Leclerc, M. & Lapointe, M. (2003) Assessment of the transferability of biological habitat models for Atlantic salmon parr (*Salmo salar*). *Canadian Journal of Fisheries and Aquatic Sciences*, **60**, 1398–1431.

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

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.

Hastie, T., Tibshirani, R. & Friedman, J.H. (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag, New York, NY.

Holmes, K.W., Chadwick, O.A. & Kyriakidis, P.C. (2000) Error in a USGS 30-meter digital elevation model and its impact on terrain modeling. *Journal of Hydrology*, **233**, 154–173.

Kadmon, R., Farber, O. & Danin, A. (2004) Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, **14**, 401–413.

Keating, K.A. & Cherry, S. (2004) Use and interpretation of logistic regression in habitat selection studies. *Journal of Wildlife Management*, **68**, 774–789.

Leathwick, J.R. & Austin, M.P. (2001) Competitive interactions between tree species in New Zealand's old-growth indigenous forests. *Ecology*, **82**, 2560–2573.

Leathwick, J.R. & Whitehead, D. (2001) Soil and atmospheric water deficits and the distributions of New Zealand's indigenous tree species. *Functional Ecology*, **15**, 233–242.

Leathwick, J.R., Rowe, D., Richardson, J., Elith, J. & Hastie, T. (2005) Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. *Freshwater Biology*, **50**, 2034–2052.

Legendre, P. (1993) Spatial autocorrelation: trouble or new paradigm? *Ecology*, **74**, 1659–1673.

Levin, S.A. (1992) The problem of pattern and scale in ecology. *Ecology*, **73**, 1943–1967.

Loukmas, J.J. & Halbrook, R.S. (2001) A test of the mink habitat suitability index model for riverine systems. *Wildlife Society Bulletin*, **29**, 821–826.

McCarthy, M.A., Possingham, H.P., Day, J.R. & Tyre, A.J. (2001) Testing the accuracy of population viability analysis. *Conservation Biology*, **15**, 1030–1038.

MacKenzie, D.I. & Bailey, L.L. (2004) Assessing the fit of site-occupancy models. *Journal of Agricultural, Biological and Environmental Statistics*, **9**, 300–318.

MacKenzie, D.I. & Royle, J.A. (2005) Designing efficient occupancy studies: general advice and allocating survey effort. *Journal of Applied Ecology*, **42**, 1105–1114.

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, J.A. & Langtimm, C.A. (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, **83**, 2248–2255.

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

Manly, B.F.J., McDonald, L.L., Thomas, D.L., McDonald, T.L. & Erickson, W.P. (2002) *Resource Selection by Animals: Statistical Design and Analysis for Field Studies*, 2nd edn. Kluwer Academic, Dordrecht, the Netherlands.

Miller, M.E., Hui, S.L. & Tierney, W.M. (1991) Validation techniques for logistic regression models. *Statistics in Medicine*, **10**, 1213–1226.

Mitchell, M.S., Zimmerman, J.W. & Powell, R.A. (2002) Test of a habitat suitability index for black bears in the southern Appalachians. *Wildlife Society Bulletin*, **30**, 794–808.

Moisen, G.G. & Frescino, T.S. (2002) Comparing five modeling techniques for predicting forest characteristics. *Ecological Modelling*, **157**, 209–225.

Olden, J.D. (2003) A species-specific approach to modeling biological communities and its potential for conservation. *Conservation Biology*, **17**, 854–863.

Panetta, F.D. & Mitchell, N.D. (1991) Homoclime analysis and the prediction of weediness. *Weed Research*, **31**, 273–284.

Pearce, J.L. & Boyce, M.S. (2006) Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology*, **43**.

Philips, S.J., Anderson, R.J. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Pulliam, H.R. (2000) On the relationship between niche and distribution. *Ecology Letters*, **3**, 349–361.

Reeves, G.K., Cox, D.R., Darby, S.C. & Whitley, E. (1998) Some aspects of measurement error in explanatory variables for continuous and binary regression models. *Statistics in Medicine*, **17**, 2157–2177.

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

Tyre, A.J., Possingham, H.P. & Lindenmayer, D.B. (2001) Matching observed pattern with ecological process: can territory occupancy provide information about life history parameters? *Ecological Applications*, **11**, 1722–1738.

Tyre, A.J., Tenhumberg, B., Field, S.A., Possingham, H.P., Niejalke, D. & Parris, K. (2003) Improving precision and reducing bias in biological surveys by estimating false negative error rates in presence–absence data. *Ecological Applications*, **13**, 1790–1801.

USFWS (1980) *Habitat Evaluation Procedures*. Report No. ESM 102 Release. United States Fish and Wildlife Service, Department of the Interior, Washington, DC.

Van Horne, B. (1983) Density as a misleading indicator of habitat quality. *Journal of Wildlife Management*, **47**, 893–901.

Van Niel, K.P., Laffan, S.W. & Lees, B.G. (2004) Error and uncertainty in environmental variables for predictive vegetation modelling. *Journal of Vegetation Science*, **15**, 747–756.

Welsh, A.H. (1996) *Aspects of Statistical Inference*. John Wiley & Sons, New York, USA.

Wintle, B.A., Elith, J. & Potts, J. (2005) Fauna habitat modelling and mapping in an urbanising environment: a case study in the Lower Hunter Central Coast region of NSW. *Austral Ecology*, **30**, 729–748.

Yee, T.W. & Mitchell, N.D. (1991) Generalized additive models in plant ecology. *Journal of Vegetation Science*, **2**, 587–602.

Zaniewski, A.E., Lehmann, A. & Overton, J.M. (2002) Predicting species distribution using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, **157**, 261–280.

Zheng, B. & Agresti, A. (2000) Summarizing the predictive power of a generalized linear model. *Statistics in Medicine*, **19**, 1771–1781.

*Received 14 May 2005; final copy received 29 October 2005*
*Editor: Phil Stephens*