

# Fauna habitat modelling and mapping: A review and case study in the Lower Hunter Central Coast region of NSW

BRENDAN A. WINTLE,\* JANE ELITH AND JOANNE M. POTTS

*School of Botany, The University of Melbourne, Vic. 3010, Australia (Email: brendanw@unimelb.edu.au)*

**Abstract** Habitat models are now broadly used in conservation planning on public lands. If implemented correctly, habitat modelling is a transparent and repeatable technique for describing and mapping biodiversity values, and its application in peri-urban and agricultural landscape planning is likely to expand rapidly. Conservation planning in such landscapes must be robust to the scrutiny that arises when biodiversity constraints are placed on developers and private landholders. A standardized modelling and model evaluation method based on widely accepted techniques will improve the robustness of conservation plans. We review current habitat modelling and model evaluation methods and provide a habitat modelling case study in the New South Wales central coast region that we hope will serve as a methodological template for conservation planners. We make recommendations on modelling methods that are appropriate when presence-absence and presence-only survey data are available and provide methodological details and a website with data and training material for modellers. Our aim is to provide practical guidelines that preserve methodological rigour and result in defensible habitat models and maps. The case study was undertaken in a rapidly developing area with substantial biodiversity values under urbanization pressure. Habitat maps for seven priority fauna species were developed using logistic regression models of species-habitat relationships and a bootstrapping methodology was used to evaluate model predictions. The modelled species were the koala, tiger quoll, squirrel glider, yellow-bellied glider, masked owl, powerful owl and sooty owl. Models ranked sites adequately in terms of habitat suitability and provided predictions of sufficient reliability for the purpose of identifying preliminary conservation priority areas. However, they are subject to multiple uncertainties and should not be viewed as a completely accurate representation of the distribution of species habitat. We recommend the use of model prediction in an adaptive framework whereby models are iteratively updated and refined as new data become available.

**Key words:** bootstrapping, conservation planning, habitat modelling, logistic regression, model evaluation, ROC.

## INTRODUCTION

Governments at all levels place considerable emphasis on urban and regional planning, and have commitments to ensure that developments are socially and ecologically sustainable (Commonwealth of Australia 2003). Protected area planning exercises in Australia over the past 10 years have utilized statistical habitat modelling methods to define biodiversity attributes (National Parks & Wildlife Service 1998; Ferrier *et al.* 2002a). Many of the theoretical and technical advances in habitat modelling and evaluation methods have come about in response to the need for better information in public land planning in Australia (Austin & Meyers 1996; Ferrier & Watson 1997; Elith & Burgman 2002; Ferrier *et al.* 2002a,b; Pearce *et al.* 2001a).

However, there is currently little scientific input into the biodiversity aspects of the urban planning process and consideration of biodiversity values is characteristically *ad hoc*. Over 40% of Australia's nationally listed threatened ecological communities and more than 50% of threatened species occur in urban fringe areas (Yencken & Wilkinson 2000) and rapidly increasing urbanization rates are their primary threat. Although the areal extent of urban development is usually relatively small, the magnitude of the impacts is often large. Urbanization is second only to land clearing for agriculture as a threat to Australia's biodiversity (Burgman & Lindenmayer 1998) and there is an urgent need to improve conservation planning practices in urban fringe areas.

The efficacy of conservation planning relies critically on the quality of the underlying biodiversity information (Pressey *et al.* 1999; Wilson *et al.* 2005). Several authors have noted the impracticality of complete biological inventory and problems arising when incomplete biological survey data are used as a basis

\*Corresponding author.

Accepted for publication March 2005.

for reserve planning (Burgman & Lindenmayer 1998; Ferrier *et al.* 2002a). The role of habitat modelling methods in addressing this problem is well established (Burgman & Lindenmayer 1998; Ferrier *et al.* 2002a; Wilson *et al.* 2005). Reliable and defensible methods for defining and predicting the distribution of wildlife habitat are critical components of conservation planning.

Here we attempt to coalesce recent developments in wildlife habitat modelling into one modelling and evaluation framework and present them in a simple enough manner that they may be applied by planners with relatively little modelling experience.

## REVIEW OF WILDLIFE HABITAT MODELLING

At a simple level, a habitat model is a numerical representation of a species' habitat preferences. It may be used to make inferences about a species habitat requirements and likely response to environmental change, or it may be used to predict a species abundance, density, carrying capacity or probability of occupying a location based on its environmental attributes. The primary use of habitat modelling in conservation planning is in predicting the spatial distribution of suitable habitat for species of interest in a landscape. Many habitat modelling methods are available that may be more or less applicable depending on the type of biological and environmental data available, the species of interest and the end use of the model. There are numerous steps involved in fitting most types of habitat model, each requiring subjective judgements that are based on experience and statistical and biological insights. There are several detailed reviews and comparisons of wildlife habitat modelling methods in the literature (Franklin 1995; Manel *et al.* 1999a,b; Elith 2000; Guisan & Zimmermann 2000; Ferrier *et al.* 2002a; Zaniwski *et al.* 2002). Our goal is to briefly outline the available methods and present those that we believe are most appropriate for predicting the distribution of species habitat in a conservation planning context in which technical expertise is limited. We seek to provide enough detail to allow planners with little statistical experience to follow our recommendations. We have provided worked examples, along with code and data for fitting and evaluating statistical habitat models in the statistical freeware R (R Development Core Team 2004). These materials are available at <http://www.botany.unimelb.edu.au/envisci/brendan/model.html>. It is important to note that biological knowledge is a critical prerequisite to sound habitat modelling. Our recommendations are confined to addressing limitations in statistical expertise and offer no means of overcoming a lack of available ecological expertise.

## Choosing a modelling method appropriate for the available data

A primary consideration in deciding on which modelling method to apply in any given situation is the type of biological survey data that are available for model development. There are five main levels of data availability: (i) *Little or no data* are available for habitat modelling; (ii) *presence-only (or ad hoc) data* are available, where occupied locations are recorded but no attempt has been made to record locations that are unoccupied systematically; (iii) *presence-absence (or binary) data* are available, where locations that are occupied or unoccupied by a given species are recorded, usually in a systematic survey; (iv) *ordinal categorical data* are available, where the number of individuals at survey locations is recorded in coarse abundance categories; and (v) *counts*, where an attempt is made to count the actual number of individuals of a given species at survey locations. The latter two situations arise very rarely in conservation planning because of the prohibitive costs associated with capture of these data and are not dealt with in detail here. The following sections outline the modelling methods available when presence-absence, presence-only or no data are available. Though we do not tackle the no-data situation in our case study, we do provide references and a description of the basic approach. For details about modelling methods appropriate for count and ordinal categorical data, see Agresti (1996), Guisan and Harrell (2000), and Pearce & Ferrier (2001).

### *Little or no data*

The absence of biological survey data does not preclude the development of a habitat model. Habitat suitability indices (HSIs) were introduced by the United States Fish and Wildlife Service as a means of mapping species habitats for the purpose of impact assessment and conservation planning (Van Horne & Wiens 1991). A HSI model for a given species and area of land represents a conceptual model that relates each measured variable of the environment to the suitability of a site for the species, scaled from 0 (for unsuitable habitat) to 1 for optimum conditions (Burgman *et al.* 2001). HSIs are very flexible and have been widely applied in conservation management (Reading *et al.* 1996; Breininger *et al.* 1998). The weakness of HSIs is that their credibility depends wholly on the credibility of the expert(s) who constructs them. The lack of independent data in the process makes them impossible to evaluate statistically and therefore less robust to scrutiny.

*Presence-only data*

Presence-only data are the most common form of observation data, and are usually available from museums and herbaria (Graham *et al.* 2004). Presence-only data suffer from the problems that observations are unplanned and tend to be biased toward towns and roads, they are often of dubious reliability and unspecified spatial accuracy, and the variation in survey effort between different environments and geographical areas cannot be controlled or adjusted in model fitting (Ferrier *et al.* 2002a; Kadmon *et al.* 2003). Nonetheless, presence-only modelling methods are widely applied due to the prevalence of presence-only data. Presence-only data may be modelled using a variety of modelling packages that are based on different ecological assumptions. Presence-only methods fall into three main categories: (i) those that use the species data without reference to any environmental data; (ii) those that model a species–environment relationship in reference to the species presence data; and (iii) those that model a species–environment relationship by characterizing the ‘background’ environment across the region of interest, and modelling the species presence in comparison to this background. The first category includes hulls and kernels (Worton 1989); which can be thought of as geographical envelopes. They are primarily useful for estimation of ranges but not for more detailed maps of species distribution, because the envelopes will generally encompass many sites that are unsuitable habitat for the species. A limitation of all envelope methods is that they are particularly sensitive to missing data and spatial error. The second category includes BIOCLIM (Nix 1986) and DOMAIN (Carpenter *et al.* 1993). BIOCLIM is a climate envelope method that maps habitat that is climatically suitable for the species, based on the distribution of the known presence records across a suite of climate variables derived from long-term records of temperature, rainfall and radiation. It is useful over large extents for broadly defining climatically suitable regions, but because of its orthogonal geometries the envelope approach tends to include many sites that are in fact unsuitable for the species (Elith & Burgman 2003). DOMAIN takes an opposite approach by determining the similarity of each cell in a map to a known presence site. That is, it measures the environmental similarity between a target site and the most similar known record site, using the Gower metric (Legendre & Legendre 1998).

The third category includes most other presence-only methods, including ENFA (Hirzel 2001), the genetic algorithm GARP (Stockwell & Peters 1999), and presence–absence methods adapted to presence-only data. These have various strengths and weaknesses, and some are more thoroughly tested than others (Elith & Burgman 2003). Of these methods,

regression models (generalized linear models: GLMs and generalized additive models: GAMs) are ecologically realistic and have shown reasonable performance when used as logistic models that have been adapted to allow modelling of presence-only survey data (Ferrier & Watson 1997; Zaniewski *et al.* 2002). We use them in the case study, and present justifications for their use in the following sections.

A substantial draw-back to the use of presence-only data is the lack of available and broadly accepted methods for evaluating the predictive performance of fitted models. There are some methods currently under development, but which at the time of publication remain largely untested and not broadly accepted (Philips *et al.* in press). For this and other reasons mentioned earlier, we would always consider presence-only data, and therefore presence-only models to be inferior to presence-absence data and models. However, due to the prevalence of presence-only data we have provided a demonstration of their use in habitat modelling our case study.

*Presence–absence data*

A number of public planning exercises in Australia have utilized presence–absence data for deriving habitat models and maps (National Parks & Wildlife Service 1998, 2000; Ferrier *et al.* 2002a). Presence–absence data may suffer from the problem of uncertain zeros (MacKenzie *et al.* 2002; Tyre *et al.* 2003), but have the advantage that they are usually collected in a ‘systematic’ manner, involving some level of geographical and environmental stratification in the sampling design (Austin & Heyligers 1989). Consequently, such data are more likely to contain samples that span the environmental gradients of interest, making model fitting more reliable.

There is a broad range of modelling methods that can utilize presence-absence data, and assessment of their relative performance has been the subject of considerable research effort (Ferrier & Watson 1997; Manel *et al.* 1999a,b; Elith 2000; Moisen & Frescino 2002). Multivariate association methods such as canonical correspondence analysis (ter Braak 1986), machine learning methods such as genetic algorithms (Stockwell & Peters 1999) and neural networks (Moisen & Frescino 2002), and tree-based methods such as classification and regression trees (Breiman *et al.* 1984) have all been proposed as potentially useful methods for modelling habitat preferences with presence–absence data.

Comparative studies have found the performance of logistic regression to be typically at least as good as other methods, if not better (Ferrier & Watson 1997; Elith 2000). Of the competing methods mentioned above, only logistic regression naturally

assumes data are derived from a binomial process, which is the correct distribution when data are binary and observations independent. Regression methods, primarily GLMs (McCullagh & Nelder 1989) and GAMs (Hastie & Tibshirani 1990) have been a commonly applied method for modelling and predicting habitat occupancy for planning purposes (e.g. National Parks & Wildlife Service 1998, 2000; Li *et al.* 1999; Loyn *et al.* 2001). One of the advantages of GLMs and GAMs is the availability of free statistical software (R Development Core Team 2004) and detailed documentation and guidance for fitting and interpreting models (Harrell 2001; Hastie *et al.* 2001; R Development Core Team 2004). For these and other reasons described below, we focus on the use of GLMs and GAMs in describing species-habitat relationships and predicting the spatial distribution of suitable habitats.

### GLMs and GAMs in habitat modelling

All GLMs are composed of a random component, described by the assumed distribution of the observation data (either binomial or Poisson for many wildlife observation data), a systematic component specifying a linear combination of explanatory (or independent) variables, and a 'link' between the random and systematic components of the model that specifies how the mean response (i.e. observation) relates to the explanatory variables in the linear predictor (Agresti 1996). When observation data are binary (presence-absence), the expected value may be modelled as  $\Pr(Y = 1)$  using the 'logit' transformation to link the random and systematic component. In this case, the regression model becomes (Agresti 1996):

$$\begin{aligned} \text{logit}(p_i) &= \log\left(\frac{p_i}{1-p_i}\right) \\ &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} \end{aligned} \quad (1)$$

where  $p_i$  is the probability that the species will be present at site  $i$ ,  $\beta_0$  is the intercept coefficient, the  $x_k$  are the habitat variables, and the  $\beta_k$  are habitat variable coefficients. Equation 1 defines the special case of the GLM known as the logistic regression model.

Generalized additive models are a non-parametric generalization of GLMs in which the relationships between the dependent and independent variables are defined by non-parametric smoothing functions (Hastie *et al.* 2001). In practice, this means that the

linear predictor ( $\beta_0 + \sum_{j=1}^k \beta_j x_j$ ) that defines the relationship between the dependent and explanatory variables in GLMs is replaced by smoothing functions

( $\beta_0 + \sum_{j=1}^k f_j(x_j)$ ). The  $f_j$  are estimated in a 'flexible'

manner, and there are a range of alternative smoothers available (Hastie *et al.* 2001). This flexibility confers an advantage to GAMs over GLMs in that they are able to 'fit' data more closely for a given number of degrees of freedom because they are not constrained to fit predefined parametric shapes (Bio *et al.* 1998). However, for the same reason, GAMs cannot be as easily interpreted as GLMs. Indeed, GAMs do not actually have a retrievable model formula in the classic sense, and interpretation generally requires a plot of the fitted response curves. Statistical packages such as R (R Development Core Team 2004) provide this facility for both GLMs and GAMs. GAMs can be fitted with the same 'link' functions as GLMs so are capable of fitting logistic regression models.

It is easy to build models that are ecologically unrealistic. Predictions from unrealistic models will have large errors and are likely to be less robust and generalizable than those from realistic models (Austin 2002). Model interpretability should therefore influence the choice of modelling method because a model can only be checked for its realism if it is interpretable. Two key features of ecological realism are the choice of explanatory variables included in the model and the shape of the response fitted for those variables (Austin 2002). Methods such as neural networks and genetic algorithms are difficult to interpret on both counts. Ennis *et al.* (1998) found that logistic regression provided as good, or better, performance than more complicated methods including multivariate adaptive regression splines (Friedman 1991) and back-propagated neural networks (Ripley 1995). Similar results have been found in other comparisons (Elith & Burgman 2002; Moisen & Frescino 2002). Even though GAMs tend to perform better than GLMs in such comparisons, the simplicity of GLMs, their broad availability in statistical packages, the ease with which they can be applied within a geographical information system (GIS) framework, and the ready availability of prediction intervals mean that they are still useful and frequently implemented. The construction of a mathematical formula that describes the relationship between the dependent variable and the environment provides a compact and communicative representation of a model. A logistic regression GLM formula that is published in a report or thesis can be used to predict species occurrence and compute prediction intervals without any direct reference to the data on which the model was fitted (providing that due care is taken not to extrapolate beyond the environmental scope of the original data). This degree of generality is not a feature of other modelling methods. A further strength of GLMs is the ease with which uncertainty about coefficients and predictions can be conveyed as standard errors and prediction intervals. Prediction intervals are not yet easily obtained from GAMs.

## A CASE STUDY: HABITAT MODELLING FOR CONSERVATION PLANNING IN CENTRAL NSW

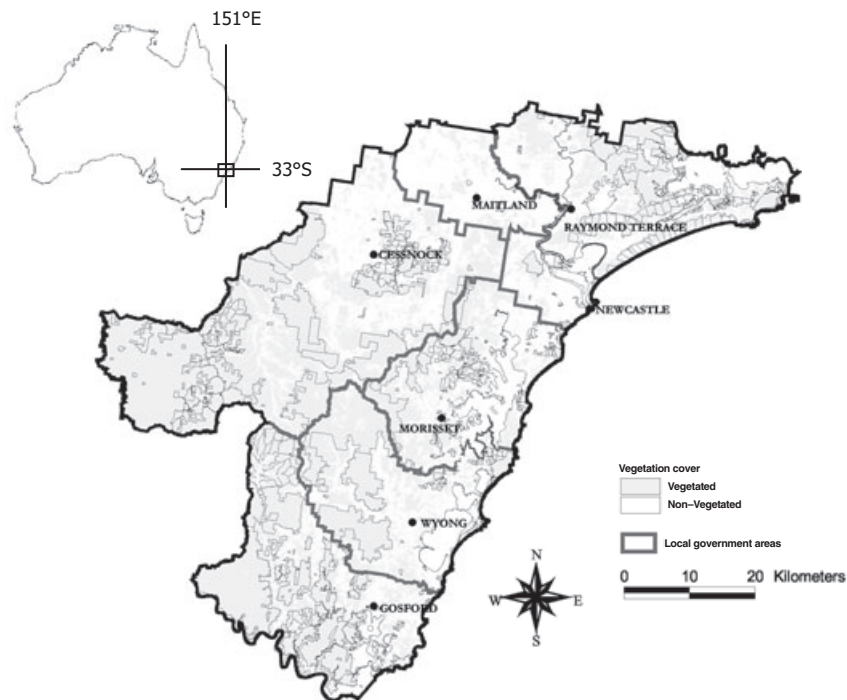
The Lower Hunter and Central Coast (LHCC) region of New South Wales (NSW) presents an opportunity to study urbanization pressures in a relatively pristine landscape with substantial biodiversity values. The region includes the seven local councils of Cessnock, Gosford, Lake Macquarie, Maitland, Newcastle, Port Stephens and Wyong (Fig. 1). It comprises large areas of native forest in both public and private tenures (National Parks & Wildlife Service 2000). The region is approximately 7200 square kilometres of which approximately 65% is under forest cover. The LHCC Regional Environmental Management Strategy (LHCCREMS) seeks to integrate biodiversity information into future land use planning and development in the region (LHCCREMS 2004). Biodiversity projects under the LHCCREMS include the development of vegetation mapping for the region, a gap analysis of biological survey data for the region and subsequent fauna surveys to augment existing biodiversity data (Ecotone Ecological Consultants 2001), the development of wildlife habitat models and habitat maps for priority species in the region (Wintle *et al.* 2004), and the development of preliminary conservation recommendations for each species (LHCCREMS 2004). This paper reports on the methods used to

develop and evaluate habitat models and maps for each of the priority species in the region.

### Priority species and their habitat requirements

Seven fauna species were selected for modelling on the basis of data availability and 'priority' status, with priority largely determined by threatened species status. By choosing species on the basis that they were deemed particularly sensitive to identifiable threats, the selection process loosely followed the rationale of the focal species approach (Lambeck 1997). The primary threatening process in the LHCC region is land clearing for development, though commercial forestry on public land may potentially pose a threat to some species (LHCCREMS 2004).

The koala (*Phascolarctos cinereus*), the yellow bellied glider (*Petaurus australis*) and the squirrel glider (*Petaurus norfolcensis*) are arboreal marsupials that are broadly distributed throughout forest and woodlands of eastern Australia. The suitability of habitat for arboreal marsupials is influenced by the size and species of trees present, soil nutrients, climate, rainfall, and the size and disturbance history of the habitat patches (Reed & Lunney 1990). The two gliders are specifically reliant on tree hollows for shelter. The squirrel glider requires food supplied by flowering acacias and banksias (Russel 1995) and the yellow bellied glider



**Fig. 1.** The Lower Hunter Central Coast (LHCC) region, situated north of Sydney in New South Wales. Seven local councils are located in the region. Light shading indicates areas under forest cover.

needs particular species of sap-trees and winter flowering eucalypts (Goldingay & Kavanagh 1993). Available habitat for these species is becoming increasingly fragmented due to habitat destruction and fragmentation caused by urban and infrastructure development, agriculture and mining (Reed & Lunney 1990).

The tiger quoll (*Dasyurus maculatus*) is a medium-sized, carnivorous marsupial that inhabits a wide range of habitat types from sclerophyll forest and woodlands, to coastal heathlands and rainforests (Edgar & Belcher 1995). The species requires suitable den sites (such as hollow logs, tree hollows, rock outcrops or caves) and a large area of intact vegetation for foraging (Edgar & Belcher 1995).

The powerful owl (*Ninox strenua*), the sooty owl (*Tyto tenbricosa*) and the masked owl (*Tyto novaehollandiae*) are broadly distributed throughout forest, woodland and rainforest of eastern Australia. All owl species tolerate some degree of fragmentation and discontinuity in forest cover. However, they do rely on large tree hollows for nesting and denning and prey on forest dependent species such as marsupial gliders and small ground mammals, making them susceptible to declines in prey abundance resulting from high levels of forest fragmentation (Kavanagh 2002).

The choice of model variables was driven by what was known about the habitat requirements of each species. Of the species modelled, home range estimates vary from as low as 0.65 ha for the squirrel glider (Russel 1995) to 20 km<sup>2</sup> for the tiger quoll (Edgar & Belcher 1995). A modelling grid cell size of 1 ha was chosen on the basis that it approximates the home range size of the smallest ranging species.

### Data collation, filtering and handling

#### Biological data

Data for the seven target species were obtained from the biological systematic survey (BSS) module of the NSW NPWS Wildlife Atlas database and from surveys commissioned by the LHCCREMS specifically undertaken to fill gaps in the Atlas data for priority species (Ecotone Ecological Consultants 2001). Systematic survey data contained records of both presence and absence for each of the seven priority species. The availability of systematic survey data reduced the effort required to prepare data for modelling, as survey method and effort covariates were available for data filtering. Only data collected after 1990 and using survey methods appropriate to detect each species were used in modelling. Records were filtered to ensure a minimum geographical separation distance of at least 500 m in an attempt to increase the probability that observations were independent. Presence records were retained in preference to absence records when

**Table 1.** Number of presence and absence sites derived from the LHCCREMS and BSS databases used in model building and evaluation for each of the seven priority species

Species	Presence records	Absence records	Total
Koala	88	162	250
Tiger quoll	36	75	111
Squirrel glider	112	129	241
Yellow-bellied glider	92	152	244
Masked owl	55	149	204
Powerful owl	97	142	239
Sooty owl	56	156	212

BSS, biological systematic survey; LHCCREMS, The Lower Hunter and Central Coast Regional Environmental Management Strategy.

two (or more) records fell within 500 m of each other. The data that remained for model building and testing are summarized in Table 1.

Survey method or effort covariates may be incorporated in models when survey data are collected using methods of variable reliability (Pearce *et al.* 2001a). In our study, such covariates were not required because data used for modelling were relatively uniform with respect to method and effort.

#### Environmental data

One of the primary limitations of the predictive performance of wildlife habitat models is the availability of broadly mapped environmental variables that are closely related to environmental attributes that affect the distribution of wildlife. Austin (2002) delineated two types of independent variables for model building: (i) 'proximal' (direct) variables are those that represent resource, shelter or thermal gradients that have a direct influence on a species distribution (e.g. temperature and foliar-nutrient); and (ii) 'distal' (or indirect) variables that have no physiological effect on the species but are correlated with 'proximal' variables (e.g. altitude, latitude). Modelling with proximal variables will more often produce a model that makes transportable and robust predictions, whereas models based on distal predictors are likely to be more specific to the location in which they were constructed (Austin 2002). However, direct variables are not always available as GIS layers because they tend to be difficult to map (Guisan & Zimmerman 2000), so model building for the purpose of prediction is often undertaken using distal variables.

Experts should play a role in the identification of environmental variables that may be important predictors of a species habitat. They may be useful in developing new variables that are combinations or derivations of distal variables. For example, indices such

as a 'hollows index' or 'foliar nutrient index' have been developed from maps of forest age and floristic composition and used in habitat modelling in the past (National Parks & Wildlife Service 1998). Similarly, vegetation variables with many categories of vegetation are often difficult to use as habitat variables in their raw form. Experts may be used to aggregate vegetation classes into those most likely to be of relevance to the species being modelled or to develop interaction terms between, for example, vegetation type, topographic context and forest age. Derived variables may serve as useful predictors of habitat where the raw variables do not. Similarly, neighbourhood measures, such as 'the proportion of old forest within 2 km' have also been used successfully in habitat modelling (Ferrier *et al.* 2002a). Neighbourhood measures convey the local environmental context of a site, which may be as important as the attributes of the site itself, especially in the case where the home range of a species is larger than the cell size used in modelling, or where survey locations are imprecisely known. Compiling a set of candidate environmental variables should involve careful consideration of the biology of the species being modelled.

A set of climatic variables derived from ANUCLIM (Houlder *et al.* 1999) were available for use in model development. These may have a direct role in determining species distributions through metabolic constraints. However, the remainder of available predictor variables would be considered distal. Neighbourhood measures were derived and tested for all species at a range of neighbourhood distances. The absence of predictor variables relating to forest growth stage, a commonly used surrogate for tree-hollow and denning availability (National Parks & Wildlife Service 1998, 2000; Loyn *et al.* 2001) is likely to place substantial limitations on the predictive performance of final models. All environmental variables were stored as spatial layers (or grids) in a GIS with a grid cell resolution of 100 m, which was satisfactory with respect to the home range size of the target species.

### Testing the adequacy of survey data based on environmental strata

There were a sufficient number of species survey data points in the region for fitting statistical models for each of the seven priority species. However, having a sufficient number of data points for fitting a regression model does not guard against inherent bias in survey data. To address concerns about geographical and environmental biases, data were tested for their coverage of key environmental strata. This was undertaken to ensure that models developed would be relevant to the range of environmental conditions present in the region and that excessive extrapolation of fitted

responses would not be required. Environmental strata were defined by overlaying GIS raster maps (1 ha grid cell size) of four key variables; broad vegetation cover (seven classes), topographic position (three classes), mean annual temperature (four classes) and mean annual rainfall (seven classes), resulting in 384 possible strata, 256 of which were represented by at least 50 ha of forest in the region. By overlaying survey locations and the map of environmental strata it is possible to tabulate the proportion of sampled *versus* un-sampled strata for each species. Maps showing the regional coverage of sampled and un-sampled strata were created for each species to illustrate where model predictions may be less reliable and where future biological surveys could be targeted.

## Model development and evaluation

### Data preparation

Geographical information system layers representing environmental variables were sampled at each survey location using ArcInfo (ESRI 1997) to construct a modelling data frame for each species. Similar functions are available in most GIS software. Each row of data contained the survey observation (1 = species present, 0 = species absent) and values for each of the candidate predictor variables at the survey locations. This resulted in an  $n \times (k + 1)$  data matrix, where  $n$  is the number of survey locations and  $k$  is the number of potential predictor variables.

In order to construct presence-only models, a second data matrix was created for each species that contained records for which the species was present. Background samples (sometimes termed 'pseudo-absence' data) were generated for 10 000 random locations across the landscape according to methods described by Ferrier and Watson (1997) and using software developed by Landcare New Zealand (J. Overton, pers. comm. 2005). The rationale for using background samples when no 'real' or systematic absence data exist is that it can be used to create an environmental profile of the study area, which is then compared with the environmental profile of known 'presence' locations. An  $n \times (k + 1)$  modelling matrix was created as for the presence-absence data, where  $n$  is equal to the number of presence observations plus the 10 000 random 'observations' of absence.

### Data transformation

In some instances, the distribution of environmental variables at survey locations may be long-tailed. This may bring about problems with model fitting due to

points in the tails having high leverage, and tends to reduce the explanatory power of independent variables. Data may be transformed to avoid this problem. A common transformation is the log-transformation, though a range of other transformations exist. Inspection of our modelling data revealed no substantial problems with long-tailed predictor variables.

#### *Variable reduction*

Approximately 50 variables were available that described the environmental characteristics which may govern site occupancy for the seven target species. It is common to offer all possible variables as candidate model predictors (e.g. National Parks & Wildlife Service 1998) and utilize an automated variable selection routine to eliminate inappropriate predictors and to specify the final model. However, offering many candidate variables tends to result in models that include nonsense predictors and exclude variables that, in fact, influence the probability of occupancy (Derksen & Keselman 1992; Chatfield 1995; Steyerberg *et al.* 2001a). This is especially problematic when the number of species observations is low compared with the number of candidate predictors.

An alternative approach is to minimize the number of variables that are offered to the variable selection routine. Harrell *et al.* (1996) recommend a rule of thumb that less than  $m/10$  predictor degrees of freedom (PDF) should be offered as candidates to a variable selection routine such as backward selection, where  $m$  equals the number of observations of the least prevalent class in the modelling data set (often the number of presence observations). PDF indicates the number of possible parameters estimated in the largest model that could be constructed from the set of candidate predictors. The number of parameters depends on the number of predictors and the form in which the predictor variables are used in the model (see next section). For categorical variables or non-linear responses there are more than one PDF for each environmental variable. For example, categorical variables have a parameter for all but one category, sometimes leading to many PDF per categorical variable. Similarly, quadratic or cubic functional forms require two and three PDF respectively. In our situation,  $m$  ranges from 36 for the tiger quoll to 112 for the squirrel glider (Table 1). Therefore, the maximum number of *candidate* PDF offered to the variable selection routine should be no more than 4 or 11 respectively. In this case, the allowable PDF is much less than the total number available. The number of candidate PDF was firstly reduced by removing distal variables that were highly correlated with proximal variables ( $R > 0.6$ ) in the context of biological or metabolic requirements of the species.

Irrespective of variable reduction issues, fitting statistical models with collinear predictor variables may cause statistical problems and should be avoided (Belsley *et al.* 1980). Expert opinion and previous habitat models were consulted where further variable reduction was required. The number of candidate model variables (and PDF) offered to the variable selection routine varied for each species depending on the amount of biological survey data available. The final set of candidate predictors, presented in Table 2, includes variables that were offered as candidates for at least one species.

#### *Variable form*

One of the strengths of GLMs and GAMs is the possibility of fitting non-linear relationships between the dependent variable (probability of occurrence) and the independent environmental variables. However, it may be difficult to know *a priori* whether a particular relationship is likely to be linear, quadratic, cubic or other. One strategy that is commonly employed is to allow an automated variable selection routine full freedom to fit any relationships using a trade-off between complexity and variance explained. However, it has been shown that such an approach may result in the fitting of spurious relationships with no logical interpretation (Steyerberg *et al.* 2001a). Consequently, we recommend either choosing functional forms with no automated variable selection or limiting the range of functional forms available to the variable selection algorithm, to those that are supported by sound ecological intuition and preliminary data analysis. In this exercise we conduct preliminary analyses on response shapes by fitting univariate GAMs with five degrees of freedom and visually inspecting plots of response shapes (Austin & Meyers 1996). This approach allows the user to assess whether more complicated responses appear sensible and are justified by the data. Visual inspection of fitted response shapes is used again later in the process for evaluating the ecological realism of final selected models (see below).

#### *Variable selection and model fitting*

Final GLMs and GAMs for all species were selected and fitted to both presence-absence and presence-background data using a backward stepwise variable selection algorithm (Venables & Ripley 2003) in R, resulting in four models for each species. The variable selection algorithm tests a series of nested models using Akaike's information criterion (AIC: Akaike 1973; Venables & Ripley 2003) to select between models. The choice of algorithm may influence the structure of the final model and alternative automated model selection algorithms, including forward selec-

**Table 2.** Abbreviated names and definitions of mapped environmental data used as candidate predictor variables for inclusion in habitat models. All environmental data were available in raster format with 100 m (side length) grid cell size

Candidate variable	Definition
Temp	Mean annual temperature derived from ANUCLIM
Cold	Mean temperature of the coldest period derived from ANUCLIM
Rain	Mean annual rainfall derived from ANUCLIM
Dry 2000	The percentage of cells in a 2000-m radius containing dry forest
Rf2000	The percentage of cells in a 2000-m radius containing rainforest
Solar	The solar radiation index of a cell derived from ANUCLIM
Elev	The elevation of a cell (in metres) above sea level
Rugg250 (500, 1000)	Topographic ruggedness (standard deviation in elevation) in a 250-m, 500-m and 1000-m radius,
Terr250 (500, 1000)	Relative terrain position in a 250-m, 500-m and 1000-m radius
Topo	The topographic position of a cell ranging from gully to ridge top (0–100)
Unmod500	The percentage of cells in a 500-m radius containing unmodified forest
Unmod2000	The percentage of cells in a 2000-m radius containing unmodified forest
Wet500	The percentage of cells in a 500-m radius containing wet forest
Wet2000	The percentage of cells in a 2000-m radius containing wet forest
Fert	Index of the soil nutrient content at a site based geochemical data (CSIRO)
Percnonfor2k	The percentage of cells in a 2000-m radius classified as cleared land
Ybglexp2000 <sup>†</sup>	The percentage of cells in a 2000-m radius containing suitable ybgl habitat
Sqglexp2000 <sup>†</sup>	The percentage of cells in a 2000-m radius containing suitable sqgl habitat
Sowlexp2000 <sup>†</sup>	The percentage of cells in a 2000-m radius containing suitable sowl habitat
Mowlexp2000 <sup>†</sup>	The percentage of cells in a 2000-m radius containing suitable mowl habitat
Powlexp2000 <sup>†</sup>	The percentage of cells in a 2000-m radius containing suitable powl habitat
Koalexp2000 <sup>†</sup>	The percentage of cells in a 2000-m radius containing suitable koal habitat

<sup>†</sup>denotes expert variable derived from vegetation classes (Michael Murray pers. comm. 2004); koal, koala; mowl, masked owl; powl, powerful owl; quol, tiger quoll; sowl, sooty owl; sqgl, squirrel glider; ybgl, yellow-bellied glider.

tion, are available. The backward selection algorithm is generally preferred to the other standard automated methods because it generally performs better in the presence of collinear candidate variables and because it requires consideration of the full model fit (Harrell 2001). There is a substantial body of literature dedicated to model selection issues and automated algorithms have been criticized (Chatfield 1995; Hoeting *et al.* 1999; Harrell 2001). However, the alternatives tend to be complicated. We have chosen to use a packaged model selection algorithm as a compromise between practicality and sophistication. We reiterate that expert reduction of predictor variables prior to final model selection is critical to the successful application of automated variable selection routines (Harrell 2001).

#### Model evaluation statistics

Presence–absence models were evaluated using two statistics: (i) the area under the receiver operating characteristic (ROC) curve (Hanley & McNeil 1982) (ROC area), closely related to the Mann–Whitney *U*-statistic; and (ii) Miller's calibration slope (Miller *et al.* 1991; Pearce & Ferrier 2000). The ROC area evaluates a model's ability to discriminate between presence and absence sites, and is therefore referred to as a measure of model (or predictive) 'discrimination' (Pearce &

Ferrier 2000). This provides an indication of the usefulness of the models for prioritizing areas in terms of their relative importance as habitat for the particular species. The ROC area ranges from 0 to 1, where a score of 1 implies perfect discrimination and a score of 0.5 implies predictive discrimination that is no better than a random guess (Bambar 1975). The actual value of the ROC area has a straightforward interpretation. It is the probability that, for a randomly selected pair of presence–absence observations derived from field surveys, the model prediction for presence will be greater than the prediction for absence.

Miller's calibration statistics (MCS) evaluate the ability of a model to correctly predict the proportion of sites with a given environmental profile that will be occupied. MCS are derived from a logistic regression of observations on the logit of predicted probabilities. The rationale is that the slope of the regression would be equal to one and the regression intercept would equal zero if the predictions from the model were perfectly calibrated (Harrell 2001; Pearce & Ferrier 2000). The model is known as the 'logistic calibration equation'. The two calibration statistics are literally the logistic regression slope ('calibration slope') and intercept ('calibration intercept'), though it is common just to report the calibration slope.

The presence-background models were evaluated qualitatively. The statistics available for evaluating presence-only models are limited, and a fair com-

parison with the presence–absence models was not possible here.

#### *The model evaluation method (bootstrapping)*

Presence–absence models for each species were evaluated using a ‘bootstrapping’ approach (Efron & Tibshirani 1997; Harrell 2001). Naively testing a model on the data that were used to fit it (also known as ‘in-sample validation’) is inappropriate and known to provide optimistic estimates of model performance (Fielding & Bell 1997; Steyerberg *et al.* 2001b), despite being a commonly used approach to model evaluation (National Parks & Wildlife Service 1998, 2000). Ideally, models should be tested on a completely independent data set, with data collection and stratification specifically designed for model evaluation. However, this is seldom a practical option, and may also be prone to high variance unless samples are large (J. Elith unpublished data 2005). Bootstrapping provides a realistic estimate of the predictive performance of a model, without incurring the expenses of collecting a completely new model-testing data set. Bootstrapping involves resampling the modelling data and conducting a series of model building and testing simulations that provide an estimate of the optimism arising from in-sample validation. The estimate of optimism is used to provide an adjusted estimate of the model evaluation statistics (ROC and MCS). The bootstrapping version implemented here is believed to provide the least biased estimate of predictive performance of any of the model evaluation methods that are based on re-sampling, including cross-validation (Hastie *et al.* 2001). Cross-validation (Efron & Tibshirani 1997) provides an alternative approach to model evaluation, and might be more feasible with methods or data sets that create large computational loads. However, its estimates of error rates with independent data can be less precise than those derived from bootstrapping (Steyerberg *et al.* 2001b), which can be thought of as a smoothed version of cross-validation (Efron & Tibshirani 1997). The bootstrapping method (the 0.632+ bootstrap) is detailed in Appendix I. R code for obtaining bootstrapped estimates of ROC and MCS is available from the URL given previously.

#### *Inspecting models for ecological realism*

Final models were inspected for their ecological realism using partial plots of univariate fitted functions (Venables & Ripley 2003), similar to those used to determine the complexity of candidate response shapes. Partial plots are available as an option in statistical software and provide a ready means to evaluate

the ecological realism of fitted responses. If an implausible fitted relationship between the probability of occupancy by the species and a particular environmental variable was observed in the partial plots, the variable was excluded from the model, or the model was re-fitted with fewer degrees of freedom for that particular environmental variable.

### **Case study results and interpretation**

#### *The adequacy of species survey data for model fitting*

Environmental strata were reasonably well sampled, even for the species with the least and most restricted survey data. For six of the seven species, sampled strata make up at least 70% of the region that is under forest cover, with the only exception being the quoll, for which only 56% of the region comprises sampled strata. Similarly, the geographical spread of sampled strata was reasonable for most species, with maps of sampled strata showing good coverage of the region for all species (Fig. 2). The maps of sampled and unsampled strata provide an indication of the areas in which habitat model predictions may be less reliable.

#### *Models and predictive performance*

The best models in terms of both predictive discrimination and calibration were those for the sooty owl, squirrel glider, yellow-bellied glider, koala and tiger quoll, all of which had predictive discrimination greater than 0.75 for their best model (Table 3). The worst models were the powerful owl and masked owl models, with bootstrapped ROC areas ranging from 0.61 to 0.69. In general, GAMs had similar variable structure and degrees of freedom to GLMs and comparable predictive discrimination. However, bootstrapped predictive calibration was always better for GAMs than for GLMs (Table 3). This is probably because the smooth functions in GAMs are better at fitting complex responses over the complete environmental range of the response, leading to more accurate predicted probabilities.

Presence-only models had similar model structures to the presence–absence models. The result suggests that the survey absence data used in the modelling exercise was not much more useful than a background random sample in defining unsuitable habitat for most species. This may result from the difficulties associated with obtaining reliable absence data for cryptic species. In addition, the quality of the presence-only data used in this modelling exercise is likely to be high compared with presence-only data found in most museums and herbaria because they were derived from systematic stratified surveys. Other studies have



**Fig. 2.** Coverage of sampled strata over the greater LHCC region for the best ((a): Koala) and worst ((b): Quoll) sampled species. Unsampled strata are represented by the light grey shading. Remaining areas within the region (not shaded) contain sampled strata or are non-forest and therefore not included in the analysis.

**Table 3.** Final presence–absence models and bootstrapped estimates of predictive discrimination (ROC) and calibration

Species	Preferred model	Preferred model type	ROC Area		Calibration	
			GAM	GLM	GAM	GLM
Koala	sp. ~ s(temp,2) + s(dry2000,3) + s(rugg500,2) + unmod2000 + percnonfor2k	GAM	0.76	0.76	0.84	0.75
Quoll	sp. ~ temp + rain + percnonfor2k + poly(dry2000,2)	GLM	0.71	0.75	0.73	0.59
Sqgl	sp. ~ s(rugg500,3) + sqglexp500 + s(unmod500,2)	GAM	0.78	0.77	0.76	0.68
Ybgl	sp. ~ s(temp,2) + s(rain,2) + s(rf2000,2) + unmod2000 + s(ybglexp2000,3) + percnonfor2k	GAM	0.78	0.77	0.73	0.63
Mowl	sp. ~ s(cold,2) + mowlexp2000 + rf2000 + s(unmod2000,2) + wet2000 + percnonfor2k	GAM	0.63	0.61	0.63	0.55
Powl	sp. ~ s(cold,3) + s(rain,2) + s(rugg500,2)	GAM	0.69	0.68	0.80	0.62
Sowl	sp. ~ s(rain,2) + rugg500 + s(sowlexp2000,3) + s(ter1000,2) + unmod2000	GAM	0.85	0.85	0.74	0.67

Only the preferred model, selected from the presence–absence models (either a generalized linear model or a generalized additive model) is presented for each species. Column ‘preferred model type’ indicates which of the methods produced the final preferred model for each species. ROC and calibration slope statistics are presented for the model derived under each modelling method. Variable abbreviations and details are given in Table 2. The expressions ‘poly(variable, *n*)’ and ‘s(variable, *n*)’ in the preferred model column indicate variables included as polynomial (‘poly’) or smoothed (‘s’) terms with ‘*n*’ degrees of freedom. ROC, receiver operating characteristic.

shown a decline in performance when fitting models with background samples compared with using presence–absence data (Ferrier & Watson 1997).

Preferred models were primarily chosen on the basis of predictive discrimination and calibration. Plausibility of response shapes was used to discern between models that showed similar predictive performance. Only model structures for the preferred model for

each species are presented in Table 3. There was some variation between the four modelling methods in terms of the variables retained and their response shapes for each species. For example, the yellow-bellied glider GLM contained the term ‘dry 2000’, while the GAM for the same species did not include this term. Similarly, the sooty owl GLM contained rainfall as a linear term, while the GAM for that species

included rainfall as a smoothed spline with two degrees of freedom (a non-linear term). However, the structural differences between GLM and GAM models were minor and the similarity in ROC areas of the two model types across all species (Table 3) implies that the differences had a minor impact on predictive discrimination.

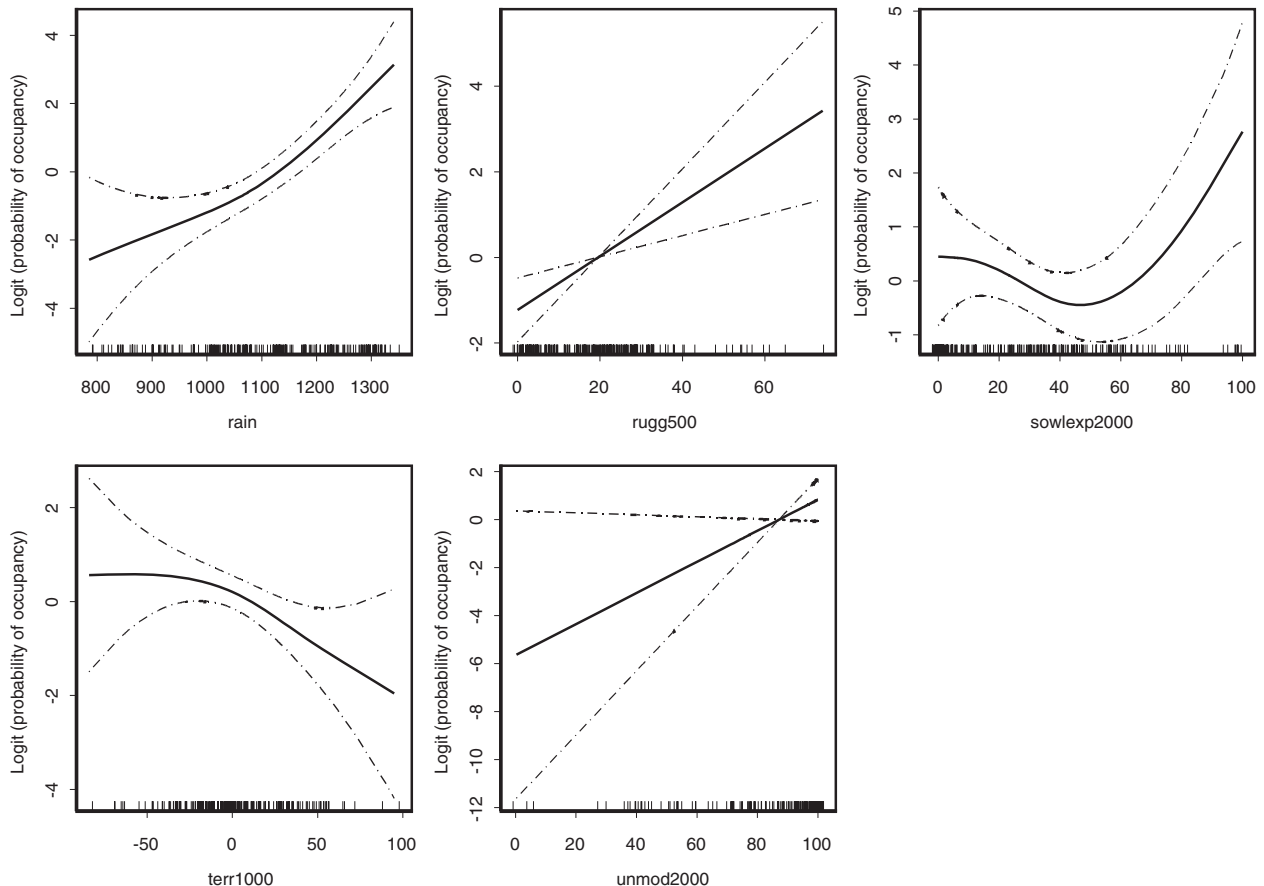
#### Plots of fitted functions

Plotting fitted relationships provided a useful means to visualize and manipulate the behaviour of fitted models to ensure that they were ecologically meaningful. In general, fitted relationships were found to be ecologically plausible. Plots of sooty owl response shapes (Fig. 3) show how the probability of sooty owl site occupancy varies with each of the predictor variables contained in the final GAM model for that

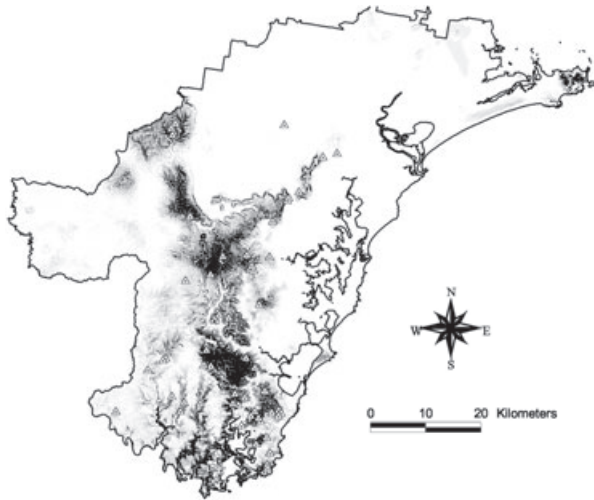
species. Fitted functions for individual variables in a multivariate model may be different to univariate fitted functions because of interactions between predictor variables. Consequently, it is important to observe response shapes for the final model to ensure that they remain ecologically reasonable in the multivariate context.

#### Habitat maps

A habitat map is provided for the sooty owl as an example of the type of output that can be used in conservation planning (Fig. 4). The map shows the predicted probability of sooty owl presence throughout the region. Probabilities have been classified into four categories (0–0.2, 0.21–0.5, 0.51–0.7, 0.71–1.0) for presentation. These maps also display the presence records for the species that were used in model



**Fig. 3.** Partial plots of the relationship between the probability of occupancy and environmental variables included in the final sooty owl model. The X-axis represents the range of values for each environmental variable. Probabilities on the Y-axis are plotted in transformed 'logit' space, so that they can be interpreted in the same way as linear regressions. Response shapes in each plot represent the relationship between each variable and the probability of sooty owl occupancy in the multivariate model context, independent of the other variables included in the model. Dashed lines represent 95% confidence intervals around the fitted response shape. Rugg500, topographic ruggedness (standard deviation in elevation) in a 500-m radius; Sowlexp 2000, the proportion of suitable vegetation (as defined by experts) in a 2000-m radius; terr1000, relative terrain position in a 1000-m radius; unmod2000, the percentage of cells in a 2000-m radius containing unmodified forest.



**Fig. 4.** The predicted probability of sooty owl occupancy across the LHCC region. Probabilities are represented as a grey-scale (white = 0–0.2, light grey = 0.21–0.5, medium grey = 0.51–0.8, dark grey = 0.81–1.0). The open triangles show point survey locations where the sooty owl is recorded as present. Habitat maps were constructed by creating a vector of habitat model predictions (one for every 1 ha grid cell in the landscape) in R using the ‘predict.gam’ command. The prediction vector is then written to an ascii text file in the correct dimensions for the landscape and imported to ArcInfo (or ArcView) for display.

fitting, allowing a visual assessment of predictive performance. Mapped predictions and uncertainties (based on upper and lower 95% prediction intervals) for the remaining six species are available from the authors.

## DISCUSSION

### The importance of ecological realism in modelling and the role of experts

The choice of modelling methods and approach to model building and evaluation presented in this paper reflects the importance we place on maintaining ecological realism throughout the model building process. While maintaining statistical rigour is central to good model building, there is an expansive literature on the statistical nuances of model fitting, interpretation and prediction (Guisan & Zimmerman 2000). Conversely, there is much less literature available concerning the importance of maintaining ecological realism in model building (Austin 1991, 2002). The process of plotting and evaluating fitted functions throughout model building and interpretation is fundamental to ensuring the ecological realism and general credibility of the final models.

Expert opinion forms an integral part of sound model building and evaluation, irrespective of the modelling platform and theoretical approach used (Steyerberg *et al.* 2001b). Some modellers view expert opinion as a source of undesirable subjectivity, and prefer to let the data dictate the model. However this ignores the limitations commonly encountered with ecological data (small to medium-sized species data sets that may include biases, and many candidate predictor variables) and the related challenges in developing robust and sensible models. Expert opinion can (and should) be used to assist in the identification of candidate model variables, interactions between variables, and likely response shapes (Pearce *et al.* 2001b; Ferrier *et al.* 2002a), for corroboration of a model’s ecological realism (Austin 2002), for *ad hoc* evaluation of model predictions (National Parks & Wildlife Service 1998), and for preparation of predictive maps for use in decision making (Ferrier *et al.* 2002a). Experts may also be used effectively in the creation of indices that may themselves be used as candidate variables for statistical modelling (National Parks & Wildlife Service 1998). Consequently, the role of experts should be thought of as complementary to other, more data-driven methods, rather than as a competing alternative. In our case study we emphasize the use of experts in variable reduction prior to automated variable selection. We use Harrell’s (2001) rule of thumb to identify the maximum PDF that should be offered to the variable selection routine. We contend that if expert knowledge is so lacking that reducing the candidate set to  $m/10$  is impossible, resulting models should be treated with scepticism.

The lack of an accepted model evaluation statistic for presence-only models detracts from their utility in conservation planning and increases the importance of expert evaluation. Development of robust evaluations of presence-only models is ongoing.

### Interpretation of habitat maps

#### Threshold predictions

In some instances, users of habitat maps may wish to identify a level of predicted probability of occupancy below which an area would be considered ‘unsuitable’ habitat for a given species. In general, such an interpretation of habitat maps is ill-advised unless the model that underpins the map is particularly well calibrated. However, in such instances where it is required, thresholds could be assigned in a risk weighted manner. For example, if it is very important not to classify an area as ‘non-habitat’ when it may be used by a species, a low probability threshold (e.g. probability of occurrence = 0.1) should be specified when delineating non-habitat. By choosing a threshold

of 0.1 for a well-calibrated model, it is implied that a 10% chance of a site (a grid cell of 100 m) being occupied by the species is a tolerable risk of failing to include areas that are, in fact, habitat for the species. This decision must be made on a case-by-case basis by evaluating the management costs of the two types of error (a false-negative and false-positive prediction), rather than by relying on an arbitrary threshold set by the modeller.

#### *Representing uncertainty*

Many forms of uncertainty impact on maps of predicted probability of species occurrence (Elith *et al.* 2002). The prediction intervals available from regression models describe the part of this uncertainty associated with estimating model parameters, and this stems from inadequate data, errors in measurement, and natural variation (Elith *et al.* 2002). There are additional uncertainties, such as model selection uncertainty and errors in explanatory variables that are, however, not addressed by standard prediction intervals. While some methods have been developed for explicitly incorporating these forms of uncertainty in prediction intervals (Burnham & Anderson 2002; Wintle *et al.* 2003), there is currently very little guidance on how prediction intervals should be represented and interpreted in applications of habitat modelling and conservation planning.

#### **Interpreting model evaluation statistics**

The two statistics used for model evaluation in this study address different model applications. Specifically, model discrimination, measured by the ROC area, measures the degree to which the model successfully ranks presence sites higher than absence sites across the region in terms of predicted probability of presence. Models with satisfactory ROC areas will provide reliable ranking of areas in terms of habitat value. This mode of evaluation is relevant when the goal of a model is to rank or prioritize areas of interest in terms of their relative value as habitat for a species. This goal is different to that of accurately predicting the proportion of sites that are expected to be occupied at a given predicted probability of occurrence. For instance, at sites with a predicted probability of occurrence of 0.5, it is reasonable to expect that approximately 50% of such sites would contain the species. The degree to which this is true is described as model calibration (Miller *et al.* 1991) and is assessed using the logistic calibration equation. It is possible that a model with good discrimination can have poor calibration. It is

also possible to improve a model's predictive ability by adjusting the model parameters with calibration statistics (Harrell 2001) or other shrinkage methods (Hastie *et al.* 2001). A model with poor calibration may still be useful for prioritizing or ranking sites in terms of habitat value, but should not be used to predict the raw probability of finding a species at a given site. The choice of whether to focus on ROC areas or calibration statistics depends on the intended application of the models.

#### **The issue of occupancy *versus* persistence**

Models of the sort discussed in this paper make the implicit assumption that occupancy of a location implies suitability of the habitat. Van Horne (1983) identified problems associated with this assumption, and Pulliam's (1988) 'source-sink' model of population dynamics formalizes its conceptual deficiencies. Pulliam's model differentiates 'sink' areas where mortality exceeds population growth from 'source' areas that maintain overall population size through emigration. Similarly, if the total population is in a state of decline, it is possible that 'remnant' individuals may exist in unsuitable habitat that will not be occupied in the next generation. When a population is expanding its range, observations of absence of the species may falsely imply that the suitability of the habitat is low. These anomalies arise due to what is known as the 'equilibrium' assumption, which underpins the current static approach to wildlife habitat modelling (Austin 2002). Inclusion of disturbance history variables, dispersal barriers, competition and successional dynamics may assist in modelling such situations, though examples of such approaches are rare.

#### **Survey design and sampling issues**

This paper has focused on the choice and execution of particular modelling methods given a set of data. Issues of data quality and sampling have not been addressed. Because modelling results are only as good as the available data, the issues of data quality and sampling demand some discussion here.

#### *Sample size*

From a statistical perspective, the choice of sample size is a function of the desired precision of results. In studies where the goal is to make inference about a particular effect (e.g. the influence of time since logging on the probability of greater glider occupancy), power analysis may be used to determine the sample size required to achieve a statistically significant result

for a given effect size, type I error rate, and sampling variance (Burgman & Lindenmayer 1998). The relationship between statistical significance, power and model predictive performance is unclear, and there is no guarantee that a model with statistically significant terms will give good predictive performance. Unfortunately, there is no simple way to determine, *a priori*, the predictive performance of a habitat model or to calculate a sample size that will ensure a predictive performance that is statistically significantly better than that of a null or alternative model. We could find only one attempt to address this problem in the statistical literature (Obuchowski 1994), making this an important area of future research. A simulation study that provided rules of thumb about sample sizes required to achieve a given predictive performance under range of likely modelling scenarios would be an important contribution to the habitat modelling literature.

#### *Data quality*

A more subtle consideration is the geographical spread and environmental stratification of observations compared to the extent and environmental variability of the region for which models are to be built. We briefly touch on this issue in our case study when testing the adequacy of survey data using environmental strata. Little guidance is provided in the literature as to the minimum geographical or environmental coverage of data for statistical modelling, though it is common sense to expect that model predictions are unlikely to be reliable in environmental domains for which there are no survey data (Austin & Meyers 1996). Stratified sampling designs such as 'GRADSECT' (Austin & Heyligers 1989), that ensure a geographical and bioclimatic coverage of sampling locations, are therefore appealing. GRADSECT theory, or newer developments such as Generalized Dissimilarity Modelling (GDM: Ferrier *et al.* 2002b) or a p-median criterion (Faith & Walker 1996), may also be applied to existing data sets to identify 'gaps' in the geographical and environmental coverage of samples.

Most statistical analyses, including the regression methods used in our case study, assume that observations are independent of each other. Survey locations that are close together are much less likely to be independent, especially when surveys target animals with large home ranges. Survey design should aim to ensure a minimum separation distance that is at least as great as the home range radius of the widest ranging species in the study. Random sampling within geographical and environmental strata assists in minimizing unwanted dependencies in observation data. Guidance on sampling theory should be sought from one

of the many texts that deal with the topic in detail (e.g. Sutherland 1996; Thompson 2002).

### **Technical issues and promising advances in modelling methods**

#### *Species versus community approaches*

Models of the distribution of single species are not the only way to approach conservation planning questions. Single species models are important for a subset of species such as threatened, focal or flagship species. However, landscape planning aims to conserve the biodiversity of a region, and there are too many species and too little data to achieve this through single-species modelling. Therefore, there is a role for community-level modelling (Ferrier *et al.* 2002b). One example of community-level modelling is GDM, already mentioned in the context of sampling; this models compositional dissimilarity across the landscape. This could then contribute to a conservation plan that aimed to span the range of species patterns in the landscape. A further role for methods based on data from many species is to use them for modelling rare species or species in data sets with relatively few records per species. In such situations there may be insufficient data for developing a robust single-species model. GDMs, canonical correspondence analysis (CCA) or multivariate adaptive regression splines (MARS) may be useful for this purpose.

#### *Spatial autocorrelation*

Spatial autocorrelation in wildlife observation data arises when environmental processes and patterns that influence the spatial distribution of wildlife are themselves spatially structured, and/or because the species is subject to demographic processes, territoriality or dispersal limitations, causing spatial dependence (contagion or dispersion effects). Demographic and environmental processes underlying spatial patterns in wildlife distributions are usually poorly understood, and therefore difficult to incorporate in model fitting. Consequently, model residuals are often spatially correlated (i.e. not independent), violating one of the basic assumptions of regression modelling. In practice, non-independence usually results in underestimation of standard errors, and overestimation of the importance of habitat variables (Legendre & Fortin 1989). Methods such as autologistic regression (Augustin *et al.* 1996), generalized estimating equations (Albert & McShane 1995) and geographically weighted regression (Fotheringham *et al.* 2002) may be used to incorporate spatial autocorrelation in habitat analyses,

but, they are still not incorporated in standard statistical software and are technically demanding to implement.

#### *Model selection and model uncertainty*

The standard approach to representing prediction uncertainty involves the calculation of prediction (or confidence) intervals that incorporate uncertainty about parameter estimates. However, this approach to prediction implicitly assumes that the model chosen is the best available representation of the truth, and effectively ignores model selection uncertainty, resulting in overconfident predictions. Automated variable selection algorithms exacerbate this problem by providing a method for dredging through many candidate predictors in search of an explanatory model (Chatfield 1995; Hoeting *et al.* 1999). One solution is to completely avoid automated model selection and to analyse only a small set of 'a priori' models (Burnham & Anderson 2002). This approach may be practical when strong prior knowledge exists and where the number of plausible models is small. However, if prediction is the primary goal and many possible models exist, some level of automated variable selection is often desirable. Model averaging approaches, including Bayesian model averaging have been promoted in a range of disciplines as a means of incorporating model selection uncertainty into statistical inference and prediction (Hoeting *et al.* 1999) and has been applied in some ecological examples (Burnham & Anderson 2002; Wintle *et al.* 2003).

#### *Estimating and incorporating detectability in habitat models*

Unless the probability of detecting a species when it is present is equal to 1, false negative observation errors will occur in species surveys. The probability of detecting the presence of the case study species in any single standard survey based on spot-lighting and call elicitation has been found to be very low ( $\text{Pr}[\text{detection/presence}] \sim 0.12\text{--}0.45$ ; Wintle *et al.* in press), making the reliability of absence data a potentially serious form of uncertainty in our case study. Recent studies have demonstrated the negative impact that false-negative observation error may have on species-habitat analyses (Tyre *et al.* 2003), meta-population models (Moilanen 2002) and monitoring studies (MacKenzie *et al.* 2002). Recently developed techniques for incorporating detectability in model estimation (MacKenzie *et al.* 2002; Tyre *et al.* 2003) reduce bias in model estimation brought about by false absences, though little effort has been invested in testing the relative predictive performance of such models.

#### *Incorporating biotic interactions*

Despite the prevalence of interspecific competition as a key concept in community ecology (Diamond 1975), very little has been attempted by way of incorporating interspecific competition in distribution models (for one exception see: Leathwick & Austin 2001). Regression models such as those described by Leathwick and Austin (2001) could be applied to fauna species. Future development of 'loop analysis' (Levins 1975) or systems of simultaneous regressions (Guisan & Zimmerman 2000) may facilitate spatial models of wildlife distributions that incorporate competitive interactions.

The GLMs and GAMs presented here are relatively simple, though suited to the conservation problem and available data. There is scope for greater sophistication in the modelling method where a high level of modelling expertise exists, though we believe our recommended approach represents a reasonable trade-off between practicality and rigour.

#### **Model improvement and adaptive management**

Due to environmental change, natural and anthropogenic disturbance and population stochasticity, all predictive models will become redundant if they are not updated through time. We recommend that models be applied in an adaptive framework (Walters & Holling 1990), allowing immediate application of models in management, but ensuring a commitment to continual improvement of models through future data collections and refinements to modelling methods.

Future data collection will be central to rigorous model testing and model refinements. They should be targeted to fill gaps in current data sets (Fig. 3). Particular attention should be paid to the collection of high quality observation data based on multiple site visits aimed at minimizing false absences (Wintle *et al.* 2004). Some important environmental variables, including those based on forest growth stage data were not available for model fitting in the LHCC region. It is likely that models would have better predictive accuracy if forest growth stage information became available and was incorporated in models.

#### **CONCLUSION**

Habitat models are now a widely used tool in public land conservation planning. Systematic conservation planning in the urban fringe will require such tools to ensure transparency and repeatability of planning outcomes. However, access to statistical and ecological expertise for urban conservation planning is likely to be limited when compared with the public land

planning processes that have taken place in Australia. It is therefore important that we continue to develop and refine planning tools, including habitat modelling methods, with an emphasis on statistical and ecological rigour and simplicity.

## ACKNOWLEDGEMENTS

This work was instigated by the LHCCREMS organization (now Hunter Councils). Steven House (EcoLogical), Michael Murray (Forest Fauna Surveys P/L), Philip Gleeson (NPWS), Daniel Connolly (NPWS), Jill Smith (NPWS), Peter Bone (NPWS), Peter Ewin (NPWS), Mick Andren (NPWS), and Rod Kavanagh (SFNSW) provided data, expert guidance on model construction, and valuable insights. Mark Burgman and Kuniko Yamada (University of Melbourne) helped throughout the project. Comments from Terry Walshe (University of Western Australia), David Tierney and Robbie Economos improved the manuscript substantially. Discussions with Trevor Hastie and John Leathwick helped clarify bootstrapping issues. Two anonymous reviewers provided constructive feedback. Sophie Powrie, Kirsty Winter and Meredith Liang (LHCCREMS) provided excellent support and useful feedback throughout the project.

## REFERENCES

- Agresti A. (1996) *An Introduction to Categorical Data Analysis*. John Wiley and Sons, New York.
- Akaike H. (1973) Information theory and an extension of the maximum likelihood. In: *Proceedings of the 2nd International Symposium on Information Theory* (eds B. N. Petrov & F. Cs'aki) pp. 267–81. Akademia Kaido, Budapest.
- Albert P. S. & McShane L. M. (1995) A generalized estimating equations approach for spatially correlated data: applications to the analysis of neuroimaging data. *Biometrics* **51**, 627–38.
- Augustin N. H., Muggleston M. A. & Buckland S. T. (1996) An autologistic model for the spatial distribution of wildlife. *J. Appl. Ecol.* **33**, 339–47.
- Austin M. P. (1991) Vegetation: data collection and analysis. In: *Nature Conservation: Cost Effective Biological Surveys and Data Analysis* (eds C. R. Margules & M. P. Austin) pp. 37–41. CSIRO, Canberra.
- Austin M. P. (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecol. Model.* **157**, 101–18.
- Austin M. P. & Heyligers P. C. (1989) Vegetation survey design for conservation: GRADSECT sampling of forests in north-eastern NSW. *Biol. Conserv.* **50**, 13–32.
- Austin M. P. & Meyers J. A. (1996) Current approaches to modelling the environmental niche of *Eucalypts*: implications for management of forest biodiversity. *For. Ecol. Manage.* **85**, 95–106.
- Bambar D. (1975) The area above the ordinal dominance graph and the area below the receiver operating characteristics graph. *J. Math. Psych.* **12**, 387–415.
- Belsley D. A., Kuh E. & Welsch R. E. (1980) *Regression Diagnostics*, John Wiley and Sons, New York.
- Bio A. M. F., Alkemande R. & Barendregt A. (1998) Determining alternative models for vegetation response analysis – a non-parametric approach. *J. Veg. Sci.* **9**, 5–16.
- Breiman L., Friedman J. H., Olshen R. A. & Stone C. J. (1984) *Classification and Regression Trees*. Wadsworth International Group, Belmont.
- Breiner D. R., Larson V. L., Duncan B. W. & Smith R. B. (1998) Linking habitat suitability to demographic success in Florida scrub-jays. *Wild. Bull.* **26**, 118–28.
- Burgman M. A. & Lindenmayer D. B. (1998) *Conservation Biology for the Australian Environment*. Surrey Beatty and Sons, Chipping Norton.
- Burnham K. P. & Anderson D. R. (2002) *Model Selection and Multimodel Inference: a Practical Information-Theoretic Approach*. Springer, New York.
- Burgman M. A., Breiner D. R., Duncan B. W. & Ferson S. (2001) Setting reliability bounds on Habitat Suitability Indices. *Ecol. Appl.* **11**, 70–8.
- Carpenter G., Gillison A. N. & Winter J. (1993) DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodivers. Conserv.* **2**, 667–80.
- Chatfield C. (1995) Model uncertainty, data mining and statistical inference. *J. Roy. Stat. Soc. Ser. A. Stat.* **158**, 419–66.
- Commonwealth of Australia (2003) *Terms of Reference*. Standing Committee on Environment and Heritage Inquiry into Sustainable Cities, Parliament of Australia, House of Representatives, Canberra.
- Derksen S. & Keselman H. J. (1992) Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *Br. J. Math. Stat. Psych.* **45**, 265–82.
- Diamond J. M. (1975) Assembly of species communities. In: *Ecology and Evolution of Communities* (eds M. L. Cody & J. M. Diamond) pp. 342–444. Harvard University Press, Cambridge.
- Ecotone Ecological Consultants (2001) Lower Hunter Central Coast Regional Biodiversity Strategy, Fauna survey and mapping project – Module 1: Fauna Surveys, Ecotone Ecological Consultants for LHCCREMS, Waratah, pp. 43.
- Edgar R. & Belcher C. (1995) Spotted-tailed Quoll. In: *The Mammals of Australia* (ed. R. Strahan) pp. 67–8. Reed Books, Chatswood.
- Efron B. & Tibshirani R. (1997) Improvements on cross-validation: the 632+ bootstrap method. *J. Am. Stat. Ass.* **92**, 548–60.
- Elith J. (2000) Quantitative methods for modelling species habitat: comparative performance and an application to Australian plants. In: *Quantitative Methods for Conservation Biology* (eds S. Ferson & M. A. Burgman) pp. 39–58. Springer-Verlag, New York.
- Elith J. & Burgman M. A. (2002) Predictions and their validation: rare plants in the Central Highlands, Victoria, Australia. In: *Predicting Species Occurrences: Issues of Accuracy and Scale* (eds J. M. Scott, P. J. Heglund, M. L. Morrison, M. G. Raphael, W. A. Wall & F. B. Samson) pp. 303–14. Island Press, Covelo.
- Elith J. & Burgman M. A. (2003) Habitat models for PVA. In: *Population Viability in Plants* (eds C. A. Brigham & M. W. Schwartz) pp. 203–35. Springer-Verlag, New York.
- Elith J., Burgman M. A. & Regan H. M. (2002) Mapping epistemic uncertainties and vague concepts in predictions of species distribution. *Ecol. Model.* **157**, 313–29.

- Ennis M., Hinton G., Naylor D., Revow M. & Tibshirani R. (1998) A comparison of statistical learning methods on the GUSTO database. *Stat. Med.* **17**, 2501–8.
- ESRI (1997) *Arcinfo 7.2*. Environmental Systems Research Institute Inc, Redlands.
- 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. *Biodivers. Conserv.* **5**, 399–415.
- Ferrier S. & Watson G. (1997) *An Evaluation of the Effectiveness of Environmental Surrogates and Modelling Techniques in Predicting the Distribution of Biological Diversity*. Department of Environment Sports and Territories, Commonwealth of Australia and NSW National Parks and Wildlife Service, Canberra.
- Ferrier S., Watson G., Pearce J. & Drielsma M. (2002a) Extended statistical approaches to modelling spatial pattern in biodiversity in north-east New South Wales. I. Species level modelling. *Biodivers. Conserv.* **11**, 2275–307.
- Ferrier S., Drielsma M., Manion G. & Watson G. (2002b) Extended statistical approaches to modelling spatial pattern in biodiversity: the north-east New South Wales experience. II. Community-level modelling. *Biodivers. Conserv.* **11**, 2309–38.
- Fielding A. H. & Bell J. F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environ. Conserv.* **24**, 38–49.
- Fotheringham A. S., Brunsdon C. & Charlton M. (2002) *Geographically Weighted Regression: the Analysis of Spatially Varying Relationships*. John Wiley and Sons, London.
- Franklin J. (1995) Predictive vegetation mapping: geographic modeling of biospatial patterns in relation to environmental gradients. *Prog. Phys. Geog.* **19**, 494–519.
- Friedman J. H. (1991) Multivariate adaptive regression splines (with discussion). *Ann. Stat.* **19**, 1–141.
- Goldingay R. L. & Kavanagh R. P. (1993) Home-range estimates and habitat of the yellow-bellied glider (*Petaurus australis*) at Waratah Creek, New-South-Wales. *Wildl. Res.* **20**, 387–404.
- 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 Ecol. Evol.* **19**, 497–503.
- Guisan A. & Harrell F. E. (2000) Ordinal response regression models in ecology. *J. Veg. Sci.* **11**, 617–26.
- Guisan A. & Zimmermann N. E. (2000) Predictive habitat distribution models in ecology. *Ecol. Model.* **135**, 147–86.
- 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.
- Harrell F. E. (2001) *Regression Modelling Strategies: with Application to Linear Models, Logistic Regression, and Survival Analysis*. Springer, New York.
- Harrell F. E., Lee K. L. & Mark D. B. (1996) Multivariate prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15**, 361–87.
- Hastie T. & Tibshirani R. (1990) *Generalized additive models*. Monographs on statistics and applied probability (eds D. R. Cox, D. V. Hinkley, D. Rubin & B. W. Silverman). Chapman & Hall, London.
- Hastie T., Tibshirani R. & Friedman J. H. (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York.
- Hirzel A. H. (2001) When GIS come to life. *Linking Landscape and Population Ecology for Large Population Management Modelling: the Case of Ibex (Capra Ibex) in Switzerland*. The University of Lausanne, Lausanne.
- Hoeting J. A., Madigan D., Raftery A. E. & Volinsky C. T. (1999) Bayesian model averaging: a tutorial. *Stat. Sci.* **14**, 382–401.
- Houlder D. J., Hutchinson M. F., Nix H. A. & McMahon J. P. (1999) *ANUCLIM User Guide, Version 5.0*. Centre for Resource and Environmental Studies, Australian National University, Canberra.
- Kadmon R., Farber O. & Danin A. (2003) A systematic analysis of factors affecting the performance of climatic envelope models. *Ecol. Appl.* **13**, 853–67.
- Kavanagh R. P. (2002) Conservation and management of large forest owls in southeastern Australia. In: *The Ecology and Conservation of Owls* (eds I. Newton, R. Kavanagh, J. Olson & I. Taylor) pp. 201–19. CSIRO, Melbourne.
- Lambeck R. J. (1997) Focal species: a multi-species umbrella for nature conservation. *Conserv. Biol.* **11**, 849–56.
- Leathwick J. R. & Austin M. P. (2001) Competitive interactions between tree species in New Zealand's old-growth indigenous forests. *Ecology* **82**, 2560–73.
- Legendre P. & Fortin M. J. (1989) Spatial pattern and ecological analysis. *Végétatio* **80**, 107–38.
- Legendre L. & Legendre P. (1998) *Numerical Ecology*. Elsevier, New York.
- Levins R. (1975) Evolution of communities near equilibrium. In: *Ecology and Evolution of Communities* (eds M. L. Cody & J. M. Diamond) pp. 16–50. Harvard University Press, Cambridge.
- LHCCREMS (2004) *Lower Hunter Central Coast Regional Biodiversity Strategy 2004 (4 Volume)*. Hunter Councils on behalf of the Lower Hunter Central Coast Regional Environmental Strategy, Thornton.
- Li W., Wang Z., Ma Z. & Tang H. (1999) Designing the core zone in a biosphere reserve based on suitable habitats: Yancheng Biosphere Reserve and the red crowned crane (*Grus japonensis*). *Biol. Conserv.* **90**, 167–73.
- Loyn R. H., McNabb E. G., Volodina L. & Willig R. (2001) Modelling landscape distributions of large forest owls as applied to managing forests in north-east Victoria, Australia. *Biol. Conserv.* **97**, 361–76.
- McCullagh P. & Nelder J. A. (1989) *Generalized Linear Models*. Chapman & Hall, London.
- 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**, 2248–55.
- Manel S., Dias J. M., Buckton S. T. & Ormerod S. J. (1999a) Alternative methods for predicting species distribution: an illustration with Himalayan river birds. *J. Appl. Ecol.* **36**, 734–47.
- Manel S., Dias J. M. & Ormerod S. J. (1999b) Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecol. Model.* **120**, 337–47.
- Miller M. E., Hui S. L. & Tierney W. M. (1991) Validation techniques for logistic regression models. *Stat. Med.* **10**, 1213–26.
- Moilanen A. (2002) Implications of empirical data quality to metapopulation model parameter estimation and application. *Oikos* **96**, 516–30.
- Moisen G. G. & Frescino T. S. (2002) Comparing five modeling techniques for predicting forest characteristics. *Ecol. Model.* **157**, 209–25.

- National Parks & Wildlife Service (1998) *Eden Fauna Modelling*. New South Wales National Parks & Wildlife Service, New South Wales Regional Forest Agreement Steering Committee, Canberra.
- National Parks and Wildlife Service (2000) *Modelling Areas of Habitat Significance for Fauna and Flora in the Southern CRA*. NPWS NSW, Canberra.
- Nix H. (1986) A biogeographic analysis of Australian elapid snakes. In: *Atlas of Elapid Snakes of Australia* (ed. R. Longmore) pp. 4–15. Australian Government Publishing Service, Canberra.
- Obuchowski N. A. (1994) Computing sample size for receiver operating characteristic studies. *Invest. Radiol.* **29**, 238–43.
- Pearce J. & Ferrier S. (2000) Evaluating the predictive performance of habitat models developed using logistic regression. *Ecol. Model.* **133**, 225–45.
- Pearce J. & Ferrier S. (2001) The practical value of modelling relative abundance of species for regional conservation planning: a case study. *Biol. Conserv.* **98**, 33–43.
- Pearce J., Ferrier S. & Scotts D. (2001a) An evaluation of the predictive performance of distributional models for flora and fauna in north-east New South Wales. *J. Environ. Manage.* **62**, 171–84.
- Pearce J. L., Cherry K., Drielsma M., Ferrier S. & Whish G. (2001b) Incorporating expert knowledge and fine-scale vegetation mapping into statistical modelling of faunal distribution. *J. Appl. Ecol.* **38**, 412–24.
- Phillips S. J., Anderson R. P. & Schapire R. E. (in press) Maximum entropy modeling of species geographic distributions. *Ecol. Model.*
- Pressey R. L., Possingham H. P., Logan V. S., Day J. R. & Williams P. H. (1999) Effects of data characteristics on the results of reserve selection algorithms. *J. Biogeogr.* **26**, 179–91.
- Pulliam H. R. (1988) Sources, sinks and population regulation. *Am. Nat.* **132**, 652–61.
- R Development Core Team. (2004) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: <http://www.R-project.org>.
- Reading R. P., Clark T. A., Seebeck J. H. & Pearce J. (1996) Habitat Suitability Index model for the eastern barred bandicoot, *Perameles gunnii*. *Wildl. Res.* **23**, 221–35.
- Reed P. C. & Lunney D. (1990) Habitat loss: the key problem for the long-term survival of koalas in New South Wales. In: *Koala Summit: Managing Koalas in New South Wales* (eds D. Lunney, C. A. Urquhart & P. C. Reed) pp. 9–31. University of Sydney, Sydney.
- Ripley B. D. (1995) *Pattern Recognition and Neural Networks – a Statistical Approach*. Cambridge University Press, Cambridge.
- Russel R. (1995) Yellow-bellied glider. In: *The Mammals of Australia* (ed. R. Strahan) pp. 226–8. Reed Books, Chatswood.
- Steyerberg E. W., Eijkemans M. J. C., Harrell F. E. & Habbema J. D. F. (2001a) Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. *Med. Decis. Making.* **21**, 45–56.
- Steyerberg E. W., Harrell F. E., Borsboom G. J. J. M., Eijkemans M. J. C., Vergouwe Y. & Habbema J. D. F. (2001b) Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J. Clin. Epidemiol.* **54**, 774–81.
- Stockwell D. & Peters D. (1999) The GARP modelling system: problems and solutions to automated spatial prediction. *Int. J. Geogr. Inf. Sci.* **13**, 143–58.
- Sutherland W. J. (1996) *Ecological Census Techniques: a Handbook*, 1st edn. Cambridge University Press, Cambridge.
- ter Braak C. J. F. (1986) Canonical Correspondence Analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology* **67**, 1167–79.
- Thompson S. K. (2002) *Sampling*, 2nd edn. Wiley, New York.
- 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. *Ecol. Appl.* **13**, 1790–801.
- Van Horne B. (1983) Density as a misleading indicator of habitat quality. *J. Wildl. Man.* **47**, 893–901.
- Van Horne B. & Wiens J. A. (1991) *Forest Bird Habitat Suitability Models and the Development of General Habitat Models*. United States Department of the Interior Fish and Wildlife Service, Washington DC.
- Venables W. N. & Ripley B. D. (2003) *Modern Applied Statistics with S*, 4th edn. Springer, New York.
- Walters C. & Holling C. S. (1990) Large-scale management experiments and learning by doing. *Ecology* **71**, 2060–8.
- Wintle B. A., McCarthy M. A., Volinsky C. T. & Kavanagh R. P. (2003) The use of Bayesian Model Averaging to better represent the uncertainty in ecological models. *Conserv. Biol.* **17**, 1579–90.
- Wintle B. A., Elith R. J., Yamada K. & Burgman M. A. (2004) LHCCREMS fauna survey and mapping project. Module 2: Habitat modelling and conservation requirements. Lower Hunter & Central Coast Regional Environmental Management Strategy, Callaghan.
- Wintle B. A., McCarthy M. A., Parris K. M. & Burgman M. A. (2004) Precision and bias of methods for estimating point survey detection probabilities. *Ecol. Appl.* **14**, 703–12.
- Wilson K. A., Westphal M. I., Possingham H. P. & Elith J. (2005) Sensitivity of conservation planning to different approaches to using species distribution data. *Biol. Conserv.* **122**, 99–112.
- Wintle B. A., Burgman M. A., McCarthy M. A. & Kavanagh R. P. (in press) The magnitude and management consequences of false negative observation error in surveys of arboreal marsupials and large forest owls. *J. Wildl. Man.*
- Worton B. J. (1989) Kernel methods for estimating the utilization distribution in home-range studies. *Ecology* **70**, 164–8.
- Yencken D. & Wilkinson D. (2000) *Resetting the Compass: Australia's Journey Towards Sustainability*. CSIRO Publishing, Melbourne.
- Zaniewski A. E., Lehmann A. & Overton J. M. (2002) Predicting species distribution using presence-only data: a case study of native New Zealand ferns. *Ecol. Model.* **157**, 261–80.

## APPENDIX 1

Method for 0.632+ bootstrap evaluation, after Harrell *et al.* (1996) and Efron and Tibshirani (1997).

- 1 Develop model on all  $n$  observations
- 2 Calculate the statistic(s) of choice for evaluation on the same data (i.e. the training data) – call this  $\text{Stat}_{\text{app}}$  because it is the apparent value of the statistic
- 3 Take a bootstrap sample, i.e. a sample of size  $n$  with replacement, of rows of the data matrix.

- Keep track of which sites are in the bootstrap sample, and which are excluded
- 4 Fit the model on the bootstrap sample (using the same methods as used on the full set)
  - 5 Compute the statistic on the bootstrap data set (observations *vs.* fitted values) and call it  $\text{Stat}_{\text{boot}}$
  - 6 Also compute the statistic on a version of the bootstrap data where the observations are randomized ( $\text{Stat}_{\text{permute}}$ )
  - 7 Use the bootstrap model to predict to the excluded data set calculate the statistic on these predictions:  $\text{Stat}_{\text{excl}}$
  - 8 Use  $\text{Stat}_{\text{boot}}$ ,  $\text{Stat}_{\text{permute}}$  and  $\text{Stat}_{\text{excl}}$  to calculate the amount of overfitting, the relative overfitting rate and weights that are then used to make a best estimate of predictive performance,  $\text{Stat}_{\text{best\_est}}$ . This statistic puts most emphasis on predictions to the excluded data, particularly when the model is overfitted (i.e. when  $\text{Stat}_{\text{boot}} - \text{Stat}_{\text{excl}}$  is large). For details of this step see Steyerberg *et al.* (2001b).
  - 9 Measure how optimistic the fit on the bootstrap sample was:  $O = \text{Stat}_{\text{boot}} - \text{Stat}_{\text{best\_est}}$
  - 10 Repeat steps 3–7, 100–200 times
  - 11 Calculate an average optimism,  $\bar{O}$
  - 12 Use to correct  $\text{Stat}_{\text{app}}$  for its optimism:  $\text{Stat}_{\text{app}} - \bar{O}$ . This is a near unbiased estimate of the expected value of the external predictive performance of the process that generated  $\text{Stat}_{\text{app}}$ .