This chapter accepted for publication in: A. Moilanen, K. A. Wilson, and H. P. Possingham, editors. Spatial conservation prioritization: quantitative methods and computational tools. Oxford University Press. Please do not circulate.

Chapter 6 - Conservation prioritisation using species distribution modelling

Jane Elith and John Leathwick

6.1. Introduction

6.1.1. Our aim

This chapter provides information on species distribution models (SDMs) and their use in conservation prioritisation. In it we aim to give readers an understanding of the breadth of methods and applications of SDM, and to equip them to identify and resolve key decisions in creating a robust model. Our primary objectives are to describe the main methodological steps required to fit and evaluate a SDM; and to summarise the most common problems with SDMs in the context of spatial prioritisation, pointing to methods for dealing with these problems or alternative approaches that might be more suitable. Note that in terms of scope, we ignore the wider use of SDMs to make inferences about ecological relationships only because this broader application is tangential to the particular focus of this volume.

To provide a context for these objectives we begin by outlining what SDMs are, their ecological bases, what they are commonly used for, and the benefits they offer for spatial prioritisation. In Section 6.3 we then outline the broad classes of models, and provide a commentary on their relative utility so that a newcomer can understand how to choose a method and set of analyses appropriate for their particular data and prioritisation problem. Details on evaluation and on methodological steps and problems in building SDMs are treated in Sections 6.4 and 6.5. We then provide an illustrative example of fitting and evaluating SDMs in Section 6.6. Finally, Section 6.7 includes brief comment on limitations and future directions.

6.1.2 Why use species distribution models?

Information on single species is often used in conservation planning, including for surrogate, iconic, focal and endangered species, or as inputs to summaries for broader biodiversity assessments (Rodrigues and Brooks 2007). The simplest way to incorporate this information is through direct use of point data – i.e. the locations where species of interest have been recorded (e.g. Eken *et al.* 2004). Such data might be derived from surveys, or from collection records in museums or herbaria. Rondinini *et al.* (2006) discuss the relative merits of point data compared with SDMs and clarify the trade-offs. Strengths of point location data are that they are easily accessible and can (depending on the taxon) be reasonably reliable. However point location data are usually sparse geographically, can include many errors of omission (Box 6.1), and are often biased in their sampling towards more easily accessible areas.

SDMs provide one of the most powerful ways to overcome sparseness typical of distributional data by relating them to a set of geographic and/or environmental predictors (Figure 6.1). Conceptually this can be seen as addressing problems caused both by inadequate sampling, and by noise (natural variability and errors) in the observations, and, if done carefully, can reduce substantially the impact of sampling biases. We do not address here the many uses or comparative benefits of single species versus multiple species (e.g. community or biodiversity) or land-type modelling (Chapter 7, and see Pressey 2004). We proceed on the premise that there will always be at least some place for single species information.

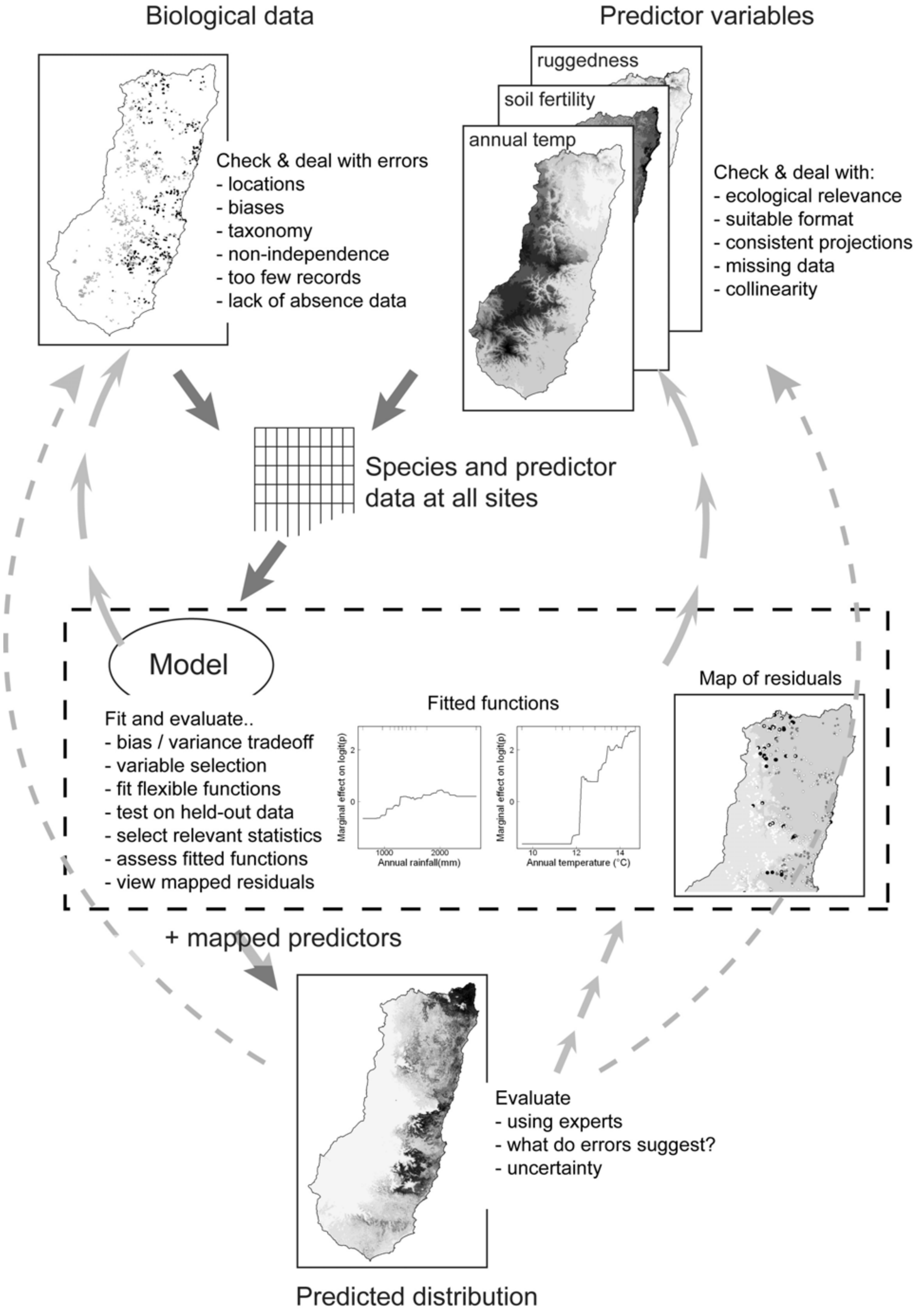


Figure 6.1: Diagram of the modelling and prediction process, highlighting potential for iterative model fitting, evaluation and improvement. Steps are discussed in detail throughout the text.

Box 6.1. Relevant words and definitions

Extent of occurrence (EOO): The region encompassing all localities where a species has been recorded (Rondinini *et al.* 2006)

Area of occupancy (AOO): A subset of the EOO, which excludes all areas within the EOO that are not occupied by the species, because they are unsuitable or presently not occupied (Rondinini *et al.* 2006)

Omission and commission errors: Falsely predicting or implying that a species is absent (omission) or present (commission).

Source and sink habitat: Habitats where birth rates exceed death rates (source) or vice versa (sink) (Pulliam 1988)

Fundamental niche: An *n*-dimensional hypervolume, every point in which corresponds to a state of the environment which would permit a species to exist indefinitely (Pulliam 2000)

Realised niche: Those portions of the fundamental niche where, even under competition and other biotic interactions, a species exists.

Presence-only: Observations of species presence (i.e. occurrence) that are part of a set where no records of absence exist – i.e., the only recorded information is of species presence.

Pseudo-absence: A location at which predictors are sampled, used in place of an absence record in a model that usually requires absences. Variously interpreted as a sample of the "background" or sampling universe or as an implied absence (Elith and Leathwick 2007).

Response (or dependent) variable: A variable whose value depends on the values of other variable(s) and constants in some relationship – e.g. in a SDM, the response might be the presence-or-absence of the species, or the number of individuals at a site.

Predictor (or independent) variable: A variable that is not dependent for change on other variables, can take any value, and is sometimes thought of as causal or influential – e.g. in a SDM, soil type and temperature might be predictor variables.

Training (test) data: Training data (also called modelling or calibration data) are those used to fit the model. Test data are not used in model fitting but to evaluate the predictive ability of the model.

6.1.3 What is a species distribution model?

Our working definition of the term SDM encompasses any method that creates a map showing geographic variation in site suitability for a single species. SDMs are usually based on records of species presence, presence-absence or abundance. Those that estimate site suitability based on statistical analyses of relationships between occurrence (or abundance) and mapped environmental predictors are also referred to as correlative or statistical models, habitat models or ecological niche models. Terminology used to describe SDMs is not consistent across the literature and some only include environment-based models, disregarding those that are geographic interpolations. A related term, "range maps", is also sometimes used to refer either to distributions predicted from correlative models, or to the broad geographic bounds of known occurrences of a species (Rondinini *et al.*

2006). Because all are attempting to predict species distributions, and methods vary in whether they exclusively use geographic or environmental predictors or both, here we will treat all that use some type of interpolation or model to make maps of species distributions as falling within the broad class of SDMs. This does not imply that all are equally useful for prioritisation, but simply that they share a common goal of producing geographically comprehensive predictions of occurrence from expert knowledge or scattered biological survey data.

Table 6.1: Applications of species distribution modelling in conservation priorisation (for definition of acronyms see Table 6.2)

Summary of application	Region	Reference
Used artificial neural networks to predict wolf occupancy and explored protected area selection	Italy	Bessa-Gomes and Petrucci-Fonesca 2003
Used GARP to predict distributions of birds and mammals, and assessed distributions in relation to biosphere reserves and areas identified as conservation priorities	NE Mexico	Huerta 2007
Used FloraMap (principal components analysis and distance measures in climatic space) to model distributions of wild peanut species and assess conservation status of the genus and prioritise conservation actions	World	Jarvis *et al* 2003
Used species data and GIS overlay methods to model distributions of endemic birds and identify hotspots of biodiversity	China	Lei *et al.* 2003
Used several methods (climate envelope, DOMAIN, logistic regression, GARP) to model 11 bird species and compared implications for reserve design based on complementarity algorithms in Worldmap.	Brazil	Loiselle *et al.* 2003
Used generalised additive models to model the distribution of 400 invertebrate taxa in Victorian rivers, and – combined with estimates of irreplacibility, condition and vulnerability – identified priority areas for reserves	Australia	Linke *et al.* 2007
Used a binomial model accounting for imperfect detection to identify the best habitat patches for Marbled Murrelets and prioritise land acquisition to protect them	USA	Stauffer *et al.* 2004
Used boosted regression trees to model the distribution of freshwater fishes, and – with directional measures of connectivity upstream and downstream – used Zonation to identify priority areas for conservation	New Zealand	Moilanen *et al.* 2008
Used logistic regression to model the distribution of four plant species, and Marxan for reserve planning. Tested the sensitivity of reserve design to threshold selection / use of probabilities.	Australia	Wilson *et al.* 2005
Used MAXENT to model the distributions of 131 amphibians and reptiles then – using ResNet software - evaluated the current conservation area network and identified potential new areas	NE India and Burma	Pawar *et al.* 2007

Some useful definitions relevant to SDMs in spatial prioritisation are provided in Box 6.1. Generally, prioritisation will be unreliable if based only on a species' geographic extent of occurrence (EOO), because it will include many locations unsuitable for the species. Use of environment-based SDMs enables more accurate identification of locations likely to support a species. However, regardless of how well they are done, predicted distributions will always include errors, which in the simplest case can be thought of as errors of omission and commission (Box 6.1, and see Section 6.4). These have different implications depending on the intended use. For instance, in reserve design errors of omission will tend toward inefficient networks. Alternatively, commission errors may lead to selection of unoccupied sites, affecting the representativeness and adequacy of selected areas (Rondinini *et al.* 2006). The costs or risks of omission and commission can be asymmetric. In biosecurity applications involving the potential spread of weeds or pathogens, managers will tend to require low omission errors if they view the costs of a false sense of security to be greater than alarmism. In contrast, an SDM supporting spatial prioritisation for translocation of a threatened species might aim for low commission errors. These contrasting emphases create an imperative to understand how to build the most appropriate model for the application, to identify likely errors and uncertainties, and to estimate their impacts on intended applications (Chapter 11).

6.1.4 Examples of SDMs in spatial conservation prioritisation

At the time of writing, interest is rapidly growing in the use of SDMs for prioritisation applications. Earlier work focused mostly on point data or simple maps estimated in geographic information systems (GIS), but attention is now shifting towards models that better capture important variation in the suitability of sites for species. Table 6.1 provides examples of applications using a range of SDM methods. As discussed in the following Sections, the quality of modelled predictions depends not only on the modelling method but on the whole approach to data and its interpolation.

6.2 Conceptual issues

6.2.1 The relative importance of space and environmental forcing

Before outlining how SDMs are constructed (section 6.5), it is useful to clarify some conceptual issues. A key consideration is the distinction between geographic and environmental space. Some modellers primarily think about species and their distribution in geographic space – sometimes because a modeller is used to thinking geographically and is expert in algorithms that deal in geographic space, or sometimes because the species is highly mobile and wide-ranging, i.e. its movement is largely determined by spatial relationships. For most species, however, aspects of the physical environment play a dominant role in determining their distributions, or at least in setting the bounds within which their movement occurs. This motivated the long history of interest in the sorting of species and communities along environmental gradients, starting with those physical geographers and ecologists who investigated responses to latitudinal and elevational gradients (e.g. Schimper 1903). Because of the consistency of these patterns, models that focus on relationships between species and their environment tend to predict distributions better than purely geographic ones, and are likely to be most useful in spatial prioritisation applications.

6.2.2 Predictor variables and their relevance to species ecology

From an ecological point of view a critical part of successful species modelling involves deriving a set of functionally relevant predictor variables (Box 6.1). The art is in considering the data and relationships from the perspective of the species. A number of studies show that predictors that are proximal – or directly and closely linked to the species requirements – have much greater predictive utility than distal, less directly relevant ones. Elevation provides one of the best known examples of this, having been used as a predictor in many studies (e.g.Whittaker 1956) largely because it is easy to measure. However, as Austin and Smith (1989) argue, few organisms respond to elevation *per se*, but rather are sorted along elevation gradients because of associated changes in climatic factors such as temperature, rainfall, solar radiation, and humidity. The difficulty in using elevation is that it is only effective as a predictor through its correlations with the more proximate variables. These correlations are imperfect and vary geographically. As a consequence, use of elevation as a predictor will result in models where key relationships are blurred, and predictive power in new regions with different correlations structures is reduced. In practice, the desirability for proximal predictors is often compromised by availability of data, although such difficulties can be overcome at least in part through use of appropriate techniques for interpolation of environmental factors (e.g. interpolated climate surfaces).

It is also important to match the spatial scale and neighbourhood context of predictors with the behaviour of the species being modelled. For example, sessile organisms will experience the environment at only one location, while mobile species may interact with their environments (feeding resources, den sites, and escape cover) over different spatial and temporal scales. Practically speaking, this can be accommodated by appropriate summaries of the predictors – e.g., summing the amount of food and shelter within the home range of a species' location (Wintle *et al.* 2005), or in freshwater ecosystems constructing predictors to reflect the network structure and upstream and downstream contributions to a given site (Moilanen *et al.* 2008). Predictors that successfully capture these functional and spatial relationships are most likely to produce robust models.

6.2.2 Disturbance, competition and equilibrium

Although environment-based SDMs may successfully reproduce geographic ranges of species, other factors can influence expected distributional patterns. These include climatic change, historical disturbances in the recent or distant past (e.g., volcanic eruptions, glaciations, fire, human activity), the influence of geographic barriers on landscape connectivity, pest outbreaks, interspecific competition and chance demographic groupings of individuals. These factors may cause species to be absent from some sites that are otherwise environmentally suitable or may displace individuals into hostile non-viable environments. The disturbances can be substantial, particularly because humans have modified large tracts of land throughout the world. It therefore cannot be taken for granted that the habitat a species would optimally choose still exists, or if it exists, that it is accessible within a fragmented landscape.

It is also important to recognise that disequilibrium patterns in one species may impact distributions of others. For instance, where the distribution of a strongly dominant species is made patchy by factors such as disturbance, disease or geographic barriers to dispersal, other species might exhibit competitive release in sites where their competitor is absent. In effect, the realised response of these other species is conditioned by the presence or absence of the strong competitor. While competitive interactions such as these can be accommodated in SDMs (e.g. Leathwick and Austin 2001), there are usually insufficient data to build them. It is, therefore, important to recognise that models developed with data from one location may fail to predict accurately in a different region with a different competitive context, i.e., where there are marked changes in the broader pool of competing species that are present. Conceptually, we view any SDM that is based on field observations of species distributions to be describing the realised environmental niche of the modelled species (Austin *et al.* 1990; Pulliam 2000). That is, the response to environment indicated by any SDM is not its fundamental response, but the realised response that is altered by disturbances and by interactions with other species.

Given these caveats, it is important to be aware of the impact of disequilibrium on modelled predictions, and to question to what extent species of interest are at equilibrium with their environment. Assessments of equilibrium need to be integrated over space and time in a way that is relevant to the species – for example, very mobile insects might be present in a region in pulses, and "equilibrium" if at all meaningful is only so in a broad sense over time. Unfortunately, no hard and fast rules can be provided for assessing the degree of equilibrium between a species and its environment – range expansions or contractions over time probably provide the only truly conclusive evidence of disequilibrium of a species (but see Leathwick *et al.* in press). More indirect evidence sometimes shows in the form of strong geographic clumping in model residuals (Longley *et al.* 2005; Figure 4.16), but equally this might indicate that an important predictor variable needs to be included in the model, or that some life-history or competition related factor is disrupting the otherwise dominant environmentally driven pattern (Legendre 1993). Species that are clearly not at equilibrium such as invasives can be problematic to model well, particularly if they have only dispersed into part of their potential environmental range.

In summary, the actual occurrence or abundance of a species is likely to be affected by many more processes than can be encapsulated in a static species model, resulting in varying degrees of misfit between the actual and predicted distribution. For conservation prioritisation a modeller needs to decide whether they want to predict the suitability of the environment for the species (for example, for restoration work) or the actual occurrence of the species (for reserve planning). For all applications, the best models will be fitted using predictors representing as many relevant factors as can be quantified, including disturbance events, nearby populations and other spatial processes. However, when the model is used to predict, choices can be made about whether to use all variables (for current distribution) or whether to "turn off" the spatial ones that constrain predictions geographically (e.g. Leathwick 1998).

6.3 Modelling methods

A difficulty facing modellers is the number of SDM methods that are available, with new ones appearing frequently. This Section gives an overview of the types of methods available, and provides suggestions about how to evaluate their suitability. We also outline the reasons underlying our recommendations, so that other methods not covered here can be evaluated using similar criteria.

We start by considering common applications of SDMs in conservation prioritisation. As most prioritisation occurs at a regional level, a method should be able to discriminate not only the broad area of occurrence, but also to identify within these areas that are more suitable than others. This broad constraint would generally favour the use of environment-based models over geographic-based ones, given that the latter err on the side of high commission errors. The following criteria provide further bases for choice.

6.3.1 Guidelines for selecting a modelling method

First, in deciding what method to use, a number of trade-offs must be considered: the time required to learn a method, the availability of an interface and/or tutorials for less experienced users, the quality of the underlying algorithm and particular implementation, flexibility for a range of data types and complexities, robustness to a range of settings (i.e. is the result dependent on expert tuning?), and computational demands. Table 6.2 summarises important attributes for a range of methods – it is informed both from experience and by published applications, and key references are included to allow further reading. From here on we use acronyms for methods defined in that table.

Second, modelling is generally best done by a person careful with data and thoughtful about interpretation. This applies broadly both to use of method but also to data preparation and model evaluation. Building expertise and allowing for a learning phase is a worthwhile investment. In our experience, results from a less-optimal method well applied by an experienced practitioner may be superior to those from a cutting-edge method implemented by an inexperienced analyst.

Third, if it is a one-off attempt or the modeller is not experienced in quantitative methods and has limited time, the safest option is a user-interface driven program that has a sound reputation suggestive of good programming and relevant settings for conservation applications. An added advantage would be straightforward linkage with a geographic information system (GIS). For presence-only data (Box 6.1), of those known to the authors, MAXENT is currently a good choice, combining ease of use with proven predictive ability. For presence-absence or abundance data, we do not know of a free interface-driven algorithm with good settings, but new software is emerging all the time. Because ecologists often use regression methods (e.g. GLMs, GAMs) to fit SDMs, these are most likely to be implemented with relevant options for conservation. GRASP (Lehmann *et al.* 2002) is useful and available as a free library for Splus (Insightful Corporation 2007), but its R (free software under General Public License; R 2007) implementation is currently less advanced. We mention more about regression later; here the focus is on implementations with interfaces. Machine learning methods such as MARS or RF are also potentially useful and available in some commercial implementations with graphic user interfaces, but at present are only rarely used in ecological applications, and can be expensive to purchase. Open access (non-interface) implementations of these methods are available in R (see later). A new user should explore what is available and look for features that are emphasised in this chapter, particularly those described in Section 6.5.

Last, if a user has a continuing need for modelling there will be considerable advantage in selecting a method and software that is flexible enough to be appropriate for a range of applications – our preference is for regression methods (including GLM, GAM, MARS, BRT) implemented in free software (e.g. R) that is serviced by a large user group, including leading statisticians and analysts. Our reason for favouring regression methods is that they are generally capable of modelling the complex relationships typical of ecological data; their assumptions are clear and mostly appropriate; they can be fitted for a range of response types; they come from statistical and computational disciplines that tend to rigorously test model functioning; and they provide access to a broad range of more specialised models that deal with a range of more complex tasks. Ecological examples of regression applications can usually be found, and a range of books and tutorials are available that provide a comprehensive introduction to their use (see Table 6.2 and Section 6.5; and Guisan and Thuiller 2005). Several tutorials have been specifically written with species modelling in mind, and code provided (Wintle *et al.* 2005, Elith and Leathwick 2007, Elith *et al.* 2008). For classification problems ensembles of trees (such as RF and bagged trees) are also likely to be useful (Prasad *et al.* 2006),

Table 6.2: Modelling methods, key features and references.

Name	Category (for algorithm)[1]	Type(s) of species data[2]	Complexity of fitted functions[3]	Uncertainty estimates[4]	Predictive performance[4]	Graphing[5] of fitted functions	Type of Prediction[6]	Comment	Useful references
Habitat suitability Index	Expert	expert	L-H	Y	L	Y	C	Needs expert knowledge and cannot be automated.	Burgman *et al.* (2001)
Kernels	Kernel	P	L	N	L	N	C	Usually in geographic space.	Worton (1989)
Hulls (convex / alpha)	Hull / envelope	P	L	N	L	N	B	Usually in geographic space.	Burgman and Fox (2003)
Kriging	Interpolation		L	N	L	N	C	Usually in geographic space.	Longley *et al.* (2005)
BIOCLIM	Envelope	P	L	N	L	N	R	The original implementation just for climate variables and meso-scale analyses; now many versions. Restricted to equal weights on variables.	Busby (1991)
DOMAIN	Similarity	P	L-M	N	M	N	C	Takes the opposite approach to an envelope model, defining distances to similar points in environmental space. Restricted to equal weights on variables.	Carpenter et al. (1993)
ENFA (Ecological Niche Factor Analysis)	Factor analysis	P	L-M	N	M	N	C	Also known as "biomapper"; interface driven in a single implementation.	Hirzel *et al.* (2002)
GARP (Genetic Algorithm for Ruleset Production)	Genetic algorithm	P	L-H	N	L-M	N	R	Interface driven; few programs all using related source code.	Peterson *et al.* (2007)
MAXENT	Maximum entropy / regression	P	L-H	N	H	Y	C	Interface driven; a single program.	Phillips *et al.* (2006)
Resource Selection Function	Regression	P*, PA, count	L-H	Y	M-H	Y	C	Often implemented as logistic regression; the broad framework developed for mobile species.	Manly *et al.* (2002)
GLM (Generalised Linear Model)	Regression	P*, PA, count	L-H	Y	M-H	Y	C	The basis for many extensions e.g. mixture models, mixed models.	McCullagh and Nelder (1989)
GAM	Regression	P*, PA,	L-H	Y	M-H	Y	C	Like GLMs but allow smoothed	Hastie and Tibshirani

Name	Category (for algorithm) [1]	Type(s) of species data[2]	Complexity of fitted functions[3]	Uncertainty estimates[4]	Predictive performance[4]	Graphing[5] of fitted functions	Type of Prediction[6]	Comment	Useful references
(Generalised Additive Model)		count						functions.	(1990); Wintle *et al.* (2005)
Decision tree	ML /Tree	P*, PA, count	L	N	L-M	N	C/B	Classification and regression trees; single trees are useful visualisations of data structure but tend to be inaccurate.	Brieman *et al. (*1984); De'Ath & Fabricius (2000)
MARS (Multivariate Adaptive Regression Splines)	ML / regression	P*, PA, count	L-H	N	M-H	(Y)	C	A fast regression method with recursive fitting procedures similar to those in trees. Multi-response models possible and useful.	Friedman (1991)
SVM (Support Vector Machines)	ML / kernel	P *	L-H	N	M-H	N	C	Not yet used much in ecology.	Bishop (2006), Drake *et al. (*2006)
Neural nets	ML / regression	P*, PA	L-H	N	M-H	(Y)	C	Need to be well tuned for optimal performance.	Pearson *et al. (*2002); Bishop (2006)
BRT (Boosted Regression Trees)	ML / tree / regression	P*, PA, count	M-H	N	H	(Y)	C	Need to be well tuned for optimal performance.	Friedman *et al.*(2000)
RF (Random Forests)	ML / tree / ensemble	P*, PA, count	M-H	N	H	(Y)	C	Uses bootstrap aggregation to average trees.	Bishop (2006), Prasad *et al.* 2006)
Bayesian methods	Bayesian	P*, PA, count	L-H	Y	M-H	(Y)	C	Flexible and allow for inclusion of prior knowledge including process information; the simpler versions (eg bayesian regression) can be fitted in WinBUGS.	McCarthy (2007)

[1] ML = machine learning

[2] P* - for these methods, pseudo-absences need to be provided to model presence-only data (P = presence only; PA = presence-absence)

[3] L=low, M=medium, H=high

[4] Since uncertainty can be calculated from multiple runs of any method (e.g. through bootstrapping), this is for uncertainty estimates obtained in other generally more direct ways

[5] Y=yes, N=no, (Y)= in some implementations

[6] C=continuous, R=rank, B=binary

6.3.2 Overview of regression

Key references in Table 6.2 provide details of the statistical formulation of several types of regression. We only give a broad overview here, because our aim for this chapter is to explain how to think about modelling and what to consider, rather than to provide details about particular methods. Section 6.5 provides further details about model fitting, prediction, and evaluation.

In general, regression methods model the relationship between a response (the "dependent variable", i.e. the species data, as presence / pseudo-absence, presence / absence, ordinal or count data) and one or more predictor ("explanatory" or "independent") variables. An equation that relates the response (y) to the predictors (**X**) is fitted: y ~ f(**X**), and values for the coefficients (parameters) of the model and the error term (the unexplained variation) are estimated. The f(X) is called the linear predictor and includes all selected variables and their coefficients. Non-linear functions are allowed: for example, through polynomial terms (e.g. as $x_1 + x_1^2$), splines, or smooth functions, variously in GLMs, GAMs or MARS. Categorical variables can be included, and interactions between predictor variables are also possible. The model can then be used to predict to new data. See Wintle *et al.* (2005) for a practical introduction to GLMs and GAMs, Leathwick *et al.* (2006) and Elith and Leathwick (2007) for MARS, or Hastie *et al.* (2001) for a comprehensive statistical treatment of all regression methods. Generalised regression models deal with different types of species data by specifying an error distribution and a link function that enable mathematically sensible relationships between the response and the linear predictor. We include BRT in the broad category of regression-based models because it has the same general structure – a linear predictor, error distribution and link function. It is, in fact, a statistical interpretation of a machine learning method, in which the linear predictor is formed by summing the contributions of a large set of simple regression tree terms that are fitted in a forwards stagewise fashion. This newer method is proving particularly useful and flexible for species modelling. Details and ecological applications can be found in De'ath 2007, Elith *et al.* (in press) and Leathwick *et al.* (in press).

Our preference is to implement these models within R (R 2007), although this does require some effort in becoming adept with the modelling software and manually interfacing with a GIS. However, increasingly links to other software such as free GIS programs are being written as add-on packages to R.

6.3.3 Methods for fitting presence-only data and for modelling in data-poor situations.

While regression-based methods were developed mostly for presence-absence or abundance data, they can also be used successfully with presence-only data (Elith *et al.* 2006), replacing absences with pseudo-absences and assuming a binomial error distribution (see Section 6.4.2).

There is also a range of non-regression methods for presence-only data. Here we mention some of the currently popular methods without trying to cover all, aiming to present ideas on how to think about the problem. Because species' responses to environment can be complex, models that rely on very simple descriptions of relationships and equal weighting of variables tend to perform relatively poorly for modelling current distribution (for example, BIOCLIM and DOMAIN; Elith *et al.* 2006). ENFA may be useful but it has an underlying factor analysis that will combine untransformed variables in a linear fashion (not ideal), and it has not been widely tested against the range of other methods now available. The approach of GARP – a genetic algorithm selecting from a range of rules – suggests that it might identify and fit complex relationships, but current implementations tend to predict relatively poorly when tested against independent presence-absence data (e.g. Elith *et al.* 2006). In contrast, MAXENT is a method with the requisite features for presence-only models; it has been developed with species modelling in mind so it produces maps and facilitates inspection of the fitted functions. MAXENT will fit complex functions where these are required, but the process can be simplified for datasets with fewer observations. We describe MAXENT briefly in Box 6.2. Our recommendation is to use it or a regression-based method.

Finally, two additional methods also provide useful options, even though they fall outside the realm of the species – environment models described above and are likely to be inferior to them in most situations. If there were only a few species records but some expert knowledge, or if environmental predictors were unavailable, these might be useful.. Alpha hulls (Table 6.2) are an extension of convex hulls, and are usually applied in geographic space. They can define polygons that encompass all of the species records, but with enough complexity so the enclosed regions do not incorporate vast tracts of uninhabited (or unsampled) space, reducing errors of commission compared with convex hulls. In spatial prioritisation such a model might be useful if suitable environmental predictors were lacking, or if the distribution was severely perturbed say by disturbance or human activity. Their biggest drawback is their simplistic dichotomous categorisation of sites as "suitable" or "unsuitable". Other options that would allow definition of more gradational boundaries include regression with geographic predictors (e.g. latitude and or longitude as a smoothed surface; Ferrier *et al.* 2002) or kernel smoothers.

Box 6.2. Description of MAXENT

MAXENT (Phillips *et al.* 2006, Phillips and Dudik 2008) predicts species distributions from presence records and predictor variables presented as grids. The model expresses the suitability of each grid cell as a function of the environmental variables at that grid cell. A high value of the function at a particular grid cell indicates that the grid cell is predicted to have suitable conditions for that species. Maxent is often referred to as a density estimation method because the computed model is a probability distribution over all the grid cells. In fitting the model, some constraints are set that enable the prediction to reflect patterns in the sample, and then the model is selected that maximises entropy (i.e. it chooses the most uniform or spread out distribution) given that those constraints are met (either exactly or approximately). Specifically, the constraints are that the model must have the same expectation for each feature (derived from the environmental layers) as the average over sample locations. Statistically, the approach can also be thought of as modelling the probability of the covariates (the predictor variables) conditional on species presence. This is a different perspective to that of regression, which models the probability of species presence conditional on the covariates.

Maxent is able to model complex relationships between the species and environment, including interactions between predictor variables. It can use continuous and categorical predictor variables, supplied as grids or as point data. It is set up so that models are kept simple if only a few species records are available, but can become more complex when more data are available. This is achieved through appropriate choice of *features* (the predictors made ready for modelling) and through carefully controlled *regularisation* of the model (i.e. balancing specificity and generality). Compared with other methods of similar complexity it is relatively fast and easy to use. The default settings have been established through tests on large data sets, and tend to perform well (Elith *et al.* 2006). Maxent is available free (www.cs.**princeton**.edu/~schapire/**maxent**/), with help files and tutorials.

The second method, habitat suitability indices or HSI (Table 6.2), enables model fitting without species data. The method is based on the judgements of experts who select critical variables that can be used to identify suitable habitat through a conceptual model of how the species responds to its environment. HSI models are difficult to test because usually there are no data for evaluation; recent testing (e.g. Mitchell *et al.* 2002; Guay *et al.* 2003) demonstrates varied outcomes. Nevertheless, HSI models are more useful than no model, with the advantage that they quantify and formalise expert opinion and provide a basis for ongoing discussion and refinement. They allow more complex fitted functions than simple overlays of GIS data, and methods are available for characterising the uncertainties associated with expert predictions (Burgman *et al.* 2001).

6.4 Evaluation

It is important to keep in mind that a model is only a mathematical simplification of the data, and will never be entirely correct. While it is tempting to place confidence in the accuracy of a map once it is produced, a healthy scepticism about any predicted distribution is advisable. Often there is as much insight to be gained from model misfit as from model fit. Misfit can point to critical and missing variables, errors in species data, and/or ecological processes or disturbances that should also be considered (Figure 6.1). Acknowledging that structured probing of a model's veracity is a key component, we first present information on evaluation before describing model building. An iterative process of model fitting, evaluation, and refinement can improve prediction quality and contribute substantially to understanding.

SDMs are used in spatial prioritisation because data are typically much sparser than required, and predictions are required in unsampled locations. The aim of evaluation is to assess how well the model predicts to new locations. By contrast, fit to the data used to develop the model (training data) is only relevant in as far as it might reveal problems in the model structure or in the data. For assessing predictive performance the two main areas requiring decisions are: (i) what data or knowledge will be used to assess the modelled predictions, and (ii) what statistics or other measures are most relevant? Both of these should be informed by the way in which the predictions will be used (Pielke 2003).

Box 6.3. Subsetting data for training and testing

Cross-validation is a straightforward and useful method for resampling data for training and testing models. In k-fold cross validation (Kohavi 1995; Hastie *et al.*; 2001) the data are divided into a small number (k, usually five or ten) of mutually exclusive subsets. Model performance is assessed by successively removing each subset, re-estimating the model on the retained data, and predicting to the omitted data. The average error when predicting to new (withheld) sites can then be calculated by averaging the predictive performance across each of the subsets. Cross-validation can be adapted to special data situations – e.g., stratified to give equal prevalence in each fold or to represent different geographic regions, or applied repeatedly to allow calculation of an average if the result has high variance.

An alternative, bootstrapping can act like a smoothed form of cross-validation. For evaluation, choices need to be made about whether predictive performance is assessed only on omitted sites (ones not in the bootstrap sample), or on weighted or unweighted combinations of omitted and modelled sites (Steyerberg *et al.* 2001). Cross-validation is more straightforward, faster, and usually adequate, but in situations with high variance bootstrapping might be more reliable.

The ideal approach to model assessment is to have an independent set of data to use for evaluation – one that covers the region of interest that reliably measures the feature of interest (e.g. presence / absence or abundance of the species), is large enough to provide confidence in the analysis, and is current. However, data resources are generally limited to the extent that this is rarely possible, so data are usually partitioned into training and testing sets and either kept separate or used in some form of resampling during which the model is iteratively fitted (trained) and tested on subsets of the data. Box 6.3 outlines useful approaches. Allocation of data to training and test sets can be structured to suit the data and application – either randomly, or focusing, for instance, on prediction to different geographic regions or extremes of the environmental space. If a resampling method such as cross-validation is used, the overall performance is instructive, as is variability across subsets of the data.

Evaluation on independent or resampled data usually focuses on summary measures that assess performance of the model in predicting to data not used in model fitting. Relevant statistics are summarised in Box 6.4. It is worth using a range of statistics to build up a picture of different aspects of predictive performance.

Other less formal knowledge can also be used for evaluation. Experts can be asked to evaluate mapped predictions and identify areas that are contrary to their expectations. This may be best done in workshops or group sessions with structured approaches for presenting the data and asking questions, and formal methods for eliciting opinions. Experts are prone to overconfidence in their own knowledge and its generality, but there is a growing literature on how to use expert opinion robustly (Ayyub 2001; Pearce *et al.* 2001; Soll and Klayman 2004; Burgman 2005). Modellers or species experts and ecologists, physiologists and GIS experts can also be involved in assessing features of model fit that affect performance. For example, mapping the spatial distribution of the residuals from the model (to identify any trends in these) and evaluation of the fitted functions (Section 6.4.3 and Figure 6.1) are both useful tools for evaluating model robustness. Errors identified through such explorations can be related to factors including a lack of predictivity, data problems (in species or predictor data or lack thereof), or model specification problems. Several experts might suggest a range of possibilities that could then be explored and fixed, where possible.

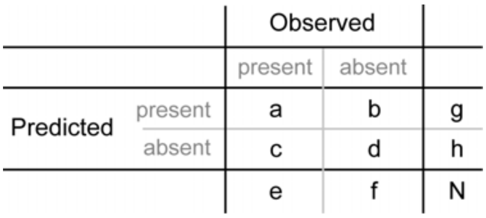
Box 6.4. Evaluation metrics

A thorough treatment of this topic is not possible here, so we outline a useful range of statistics and point to other sources of information. The appropriate statistics depend on the response being modelled and the intended use of the predictions. For instance, predictions from presence-absence data can either be assessed as continuous predictions or thresholded to produce presence-absence predictions, and evaluated as such.

For presence-only data: Limited measures are available if only presence records are available for evaluation. The most useful are the correlation-based method of Boyce *et al.* (2002) and AUC with pseudo-absences (Phillips *et al.* 2006).

For thresholding predictions: Several papers evaluate different methods for setting thresholds (e.g. Liu *et al.* 2005), some of which use costs to weight errors (e.g. Fielding and Bell 1997; Wilson *et al.* 2005).

For presence-absence data: Many statistics are based on the confusion matrix (Figure 6.2), which first requires that non-binary predictions are thresholded to convert them to presence/absence predictions.



		Observed		
		present	absent	
Predicted	present	a	b	g
	absent	c	d	h
		e	f	N

Figure 6.2: A 2 x 2 confusion matrix, where a to d are counts, and sum to N. The marginal totals are e to h.

Box 6.4 cont'd……………………………………………….

Some common measures:

- Prevalence = (a + c) / N
- Percent correctly classified (PCC) = 100.(a+ d)/N = Accuracy
- Kappa: K = {(a + d) - [(e.g + f.h)]/N}/{N – [(e.g + f.h)]/N}
- Sensitivity = a/e = true positive rate (TPR)
- Specificity = d/f = true negative rate (TNR)
- False positive rate (FPR) = b/f = (1 Specificity)
- False negative rate (FNR) = c/e

Of these, several are severely affected by prevalence (e.g. very high values of PCC can be achieved simply by predicting zero everywhere for a rare species) and those adjusted for chance agreement are more useful (e.g. Kappa). Kappa can be used for two or more classes, and weighted to reflect the seriousness of disagreement.

For ordered, probabilistic or continuous predictions:

Receiver Operating Characteristic (ROC) curve and the area under the curve (AUC) – for continuous or ranked predictions and binary observations. AUC reports the probability that, for a randomly selected pair of presence-absence observations, the prediction for the presence will be greater than the prediction for the absence. It is useful for finding whether the predictions are well ordered, but assesses the ranking of predictions rather than the magnitude of any errors. Values range from 0 to 1, with 0.5 indicating that a model performs no better than random predictions while a value of 1.0 indicates perfect discrimination of presences from absences. Values over 0.75 are likely to indicate useful models (see Pearce and Ferrier 2000). Full ROC curves (sensitivity vs. 1-specificity) are also useful for visualising performance across a range of thresholds. Related statistics that are valuable for giving additional insight are partial ROCs (Dodd and Pepe 2003) and precision-recall curves (Davis and Goadrich 2006). For all measures see the ROCR package in R.

Calibration statistics and reliability diagrams assess whether the fitted or predicted probabilities accurately reflect the probabilities occurring in the data. See Pearce and Ferrier (2000) and Wilks (1995)

Deviance measures the magnitude of the deviations of observations from fitted or predicted values, and has various forms for binary, count or gaussian responses. It is often used by comparing the null deviance with residual deviance as a percentage to reduce the influence of sample size and prevalence, although it is still sensitive to prevalence, especially at the extremes. For binary responses ecological models commonly fall into the range 20-50% deviance explained (e.g. Leathwick *et al*. in press)

Correlation: although generally used for continuous (Gaussian) data, correlation statistics can also be used for predictions fitted to binary data. (see. e.g. Elith *et al*. 2006)

There are a range of other relevant measures that test different aspects of performance – e.g. Moisen and Frescino (2002). The machine learning, weather forecasting and medical literatures also describe relevant research.

Choice of a test statistic should be based on the needs of the application. For example, if sites need to be ranked for conservation prioritisation the ordering rather than the exact value of the predictions will be important, and AUC will test that. However, situations where different errors have different costs will either require predictions that have been turned into binary values and tested with e.g. kappa, or continuous predictions that are well calibrated (e.g. calibration statistics, deviance)

6.5 Steps in making a model using species and environmental data

Keeping in mind these principles of predictive performance and its evaluation, how should a model be developed? Here we aim to communicate not only what to do, but why, and to provide practical information on how to deal with common problems encountered with data. Our commentary is largely informed by our experience using regression methods, but many of the steps and elements described here are also relevant to non-regression alternatives.

6.5.1 Gather species records

A model quantifies the relationship between the species (the response) and its environment (the predictors), and its quality will therefore depend on the quality of the available data. The main features of species data that contribute to a good model are: a sufficient number of records to model the main trends, and observations that are independent of each other, reliable, accurately recorded, and sampling the main gradients of environmental and geographic space at relevant times (e.g. recently, if current distribution is required) and without bias. We discuss these below, keeping the commentary broad enough to be relevant to either the collation of existing records or new data collection (but see also Cawsey *et al.* 2002).

Sample size: Despite some claims in the literature, no algorithm can model extremely sparse species data successfully – with few records the best that can be done is identification of the strongest environmental gradient(s) or trend(s). This means that the true response to less important variables will not be explicitly modelled, but rather averaged across them (Barry and Elith 2006). The number of records required depends on the complexity of the relationship between the species and its environment. In general, more records are better, because large data sets allow more subtle aspects of the response to environment to be modelled (Elith *et al.* in press). In the past, statisticians have used as a rule of thumb for regression models (Harrell 2001) a requirement for at least 10 records to allow good estimation of each fitted parameter (coefficient) - e.g. to fit a model with three coefficients would require more than 30 data points. Fewer data tend towards spurious overfitted models that can be highly imprecise and unreliable. Most models require estimation of many parameters, because one parameter is needed for each variable, and more than one is required when fitting non-linear responses. Further, for presence-absence species records, the "records" in the "10 records per parameter" does not refer to the number of sites in the data set, but to the number of presences or absences recorded (whichever is smaller). This suggests that with, say, 50 presence records from 200 sites, one might only expect to fit a model using three predictors, assuming that some responses will be non-linear and require estimation of multiple parameters.

Clearly, this is not good news for modelling rare species. However, recent advances in model fitting methods provide some assistance in more reliably selecting models. Examples include the lasso and model averaging (Burnham and Anderson 2002; Reineking and Schröder 2006). However, even with a well-behaved algorithm, with few species records the model is likely to provide only a general indication rather than a realistic and detailed prediction. As a guideline, only attempt to model species with more than 20, and preferably 30 occurrences. Below this, the model is unlikely to be useful for conservation prioritisation. Alternatively Kremen *et al.* (2008, supplement) give useful examples of testing models and excluding those with poor predictive power. In the absence of sufficient records, the best options are to use an expert model without species data, or to provide ancillary information by modelling the taxon of interest either with other species in a similar functional group or guild in a multi-response model (Elith and Leathwick 2007), or as a broader component of biodiversity (Chapter 7).

Independence and coverage: Records need to be independent in the sense that each brings new information to the model. Non-independence leads to overconfidence in the precision and reliability of parameter estimates. Multiple observations at one location should usually be reduced to one record (e.g. by averaging counts, assigning presence if presence was recorded at least once, or by down-weighting each replicate site in the model), unless they contribute important seasonal or long term data or provide information on detectability. Clusters of samples may be spatially so close that from a biological perspective one would not consider them as independent (e.g., very strong similarities in environment and within easy dispersal distance). These could be reduced to one sample, unless they are to be used to explore spatial autocorrelation (Wintle and Bardos 2006; and see Box 6.5) or they represent a nested survey that could be modelled in a mixed model (Box 6.5). Samples should be spread across the whole region of interest, primarily in environmental space but also geographically. The degree to which samples cover the space can be assessed in simple summaries (e.g. by binning the space and assessing record spread; see Cawsey *et al.* 2002) or with statistics such as p-medians that assess the environmental coverage of a set of sites in relation to another set such as the whole landscape (Faith and Walker 1996).

Reliability, accuracy and bias: Typical errors in species data include those of mis-identification, failure to detect a species that is present, errors in describing the location, and temporal issues. It is usually worth committing time and resources to cleaning species records, either using expert opinion or careful cross-checking of records and investigation of outliers in geographic and environmental space. A clear set of criteria should be set in advance for accepting or rejecting data, reflecting the required end use of the model.

If a species has low detectability it may be represented by few records, and in presence-absence data sets there will be a bias towards false negative records, affecting the modelled relationships and the overall estimate of the species prevalence. We point to specialised methods for dealing with imperfect detection in Box 6.5.

Box 6.5. Dealing with imperfect detection, spatial autocorrelation and structured data.

Detection: This is an issue in presence-absence and abundance data, where imperfect detection leads to false absences. The two broad approaches to dealing with it are (i) using distance sampling and related analyses (Buckland 2004), or (ii) repeat visit surveys followed by modelling that allows simultaneous estimation of detection and occurrence probabilities, and can include co-variates e.g. for survey effort (Tyre *et al.* 2003).

Spatial autocorrelation (SA): SA is defined as the tendency for geographically proximate sites to have values that are more similar (positive SA) or more different (negative SA) than for random associations (Legendre 1993). SA results from either spatial patterning in variables describing the environment or disturbance, or from community processes. If predictors with spatial pattern are missing or there are strong social interactions that aren't modelled, a fitted model will have spatial structure in the residuals. SA in residuals should be checked (e.g. using variograms – see Bio *et al.* 2002, or by plotting in geographic space) and, if present, might point to the need for specialised models (Leathwick 1998; Wintle and Bardos 2006; Dormann *et al.* 2007)

Structured data: Ecological data can have inherent structure (i.e., grouping) – e.g. measures at quadrats nested within large sites, or in longitudinal data, where for example several research boats could do repeated trawls over time. It is usually most efficient and statistically correct to model the structure, because such models require fewer parameters, and correlation structure in the model is properly dealt with. This is achieved in regression with mixed models – usually generalised linear mixed models (GLMMs; Venables and Dichmont 2004) or their non-parametric counterparts (GAMMs; Wood 2006). Alternatively, the survey structure may be modelled by variables in the model, in which case standard models could be used. Wood(2006) presents relevant tests that check for structure in model residuals.

Species records that are gathered from existing collections such as those in museums or herbaria are also commonly biased both spatially and temporally, e.g., with more intense collection effort close to towns and roads, and old records in habitats that no longer exist. In terms of the spatial bias, if predictors in the model are mostly environmental, geographic biases in the sample data may not be problematic, particularly if they aren't strongly correlated with environment. However, this should not be assumed – roads are often topographically biased and cities tend to be in warmer and/or coastal environments, or associated with fertile soils. Unless there are resources for supplementing the data with new targeted sampling, the best way to deal with bias is either to weight the presence records to give greater weight to those in sparsely sampled areas, or to incorporate the same bias into the pseudo-absences (S. Phillips *et al.* in press). Similarly, temporal issues need to be addressed sensibly, either by deleting records that are no longer relevant, or by preparing predictor variables that reflect conditions at the time of sampling. There will also be sampling biases that favour some taxa over others. It is well worth attempting to characterise the biases in the data and to deal with it in the modelling (e.g. Dennis and Hardy 1999; Reese *et al.* 2005; Wieczorek *et al.* 2005; Schulman *et al.* 2007; Phillips *et al.* in press), so that the models are truly indicating response to environment and are not primarily models of bias in survey effort.

Pseudo-absences: Often only presence data are available, so the choice is either to model just the presence data (e.g., by defining an environmental envelope), or to compare their distribution with that of the environments in the region of interest. Methods that compare the distribution of presence records against the "background" environment tend to predict better than those that only use the presence records (Elith *et al.* 2006). Pseudo-absences can be used to characterise the background or region, and are often compiled by taking a random sample of the whole region. This is the most naïve sampling approach, since in many instances the whole region includes places that are clearly unsuitable for the species. For instance, if the aim is to model possible natural locations and most of a landscape is urban, urban areas should not be sampled for pseudo-absences. An underlying idea is that the pseudo-absences should be placed in areas that could reasonably have been sampled for the presence or absence of the species. Stated differently, placement of pseudo-absences should take into account bias in survey effort in the presence records, so that the model is not dominated by the survey effort. (Note that this is conceptually different to the idea reported by some, that absences should be placed in the most unlikely areas; we do not see that as consistent with the model theory). The scale, extent and intended application of the model are all pertinent in deciding how to sample pseudo-absences, and the impact of pseudo-absence selection is substantial, affecting both overall predictive performance and spatial trends in the predictions (S. Phillips *et al.* in press). Unfortunately at the time of writing the literature exploring practical alternatives is sparse.

6.5.2 Collate and pre-process predictors

Earlier in Section 6.2 we explored issues related to the importance of building ecologically sensible models. It is hard to overstate the importance of a functionally relevant predictor set because this can have substantial impacts both on model fit and the ability of a model to predict robustly in unsampled areas. Useful sets of predictors for terrestrial, freshwater and marine environments are outlined in Table 6.3. These focus on predictors relevant at a regional level; across larger extents (e.g. global models) the focus tends to be on climatic variables (Mackey and Lindenmayer 2001). Both modelled environmental variables and remotely sensed data are being used increasingly in fitting SDMs, adding to the suite of available information from which to construct relevant predictors (Turner *et al.* 2003). To some extent, the difficult and subjective task of collating suitable predictors is constrained by the amount of species distribution data that is available, which imposes a limit on the number of predictors that can reasonably be handled. It motivates effort in presenting predictors efficiently and in as relevant a form for the species as possible, as described in the following Sections.

Formatting considerations: At a practical level, production of a digital map predicting a species distribution from an SDM requires that all the model predictors are available across the entire region of interest. These are usually stored in a geographic information system (GIS) either in grid (raster) or vector (polygon or line) format. For terrestrial and marine applications environmental predictors are usually stored as grids, because these generally provide the most straight forward route for calculating predictions. Rivers are a special case because of the directional nature of their flows, and the manner in which attributes aggregate additively along the river network downstream from its headwaters (Leathwick *et al.* in press; Moilanen *et al.* 2008).

When using any GIS data, it is important to ensure first that any inconsistencies in projections, datums, grid cell sizes and alignment are corrected before use. Sometimes coarser data (i.e. data with low spatial resolution) will have to be interpolated to a smaller grid cell to match that of other grid layers, but the lack of true finer level detail in such layers should be clearly acknowledged. Some researchers prefer to use the coarsest available grid size as the cell size for analysis, because this clearly represents the lack of detail in some predictors; however, it discards all finer detail in other predictors. The spatial accuracy of the species records should also be considered in choice of cell size. Whatever decision is made, explicit reference to it and representation of the likely errors resulting from it are essential. Some modelling methods can predict to areas where there are missing data in some predictors; if this is not the case for the method being used, missing data will have to be masked out or values imputed. If the spatial prioritisation is to involve measures of distance (e.g. for characterising connectivity or dispersal) the data should be projected so that both grid cell sizes and geographic distances are consistent across the region. Geographic coordinates based on latitude and longitude are not suitable because grid cell sizes and calculated distances become compressed with increasing latitude.

Table 6.3: Examples of functionally-based predictor variables useful at regional and national extents

Realm	Predictors	Examples

Terrestrial	Climate summarised relevant to species (e.g. minimum temperature of coldest period; driest month, potential evapotranspiration; humidity or vapour pressure deficit; solar radiation adjusted for topography; snow cover; frost-free days); wetness; inundation duration; distance to freshwater; topographic: slope position, roughness (variation), slope, cosine of aspect ; soil: type, depth, fertility, drainage; vegetation: provision of food and shelter resources, shading.	Leathwick 2001;
		Ferrier *et al.* 2002;
		Elith *et al.* 2006;
		Drake *et al.* 2007

Freshwater	Variables characterising upstream (catchment), local and downstream conditions, including flow, flow variability, frequency of flood and drought events, velocities, temperature, water quality (N, P), barriers, streambed conditions (bedrock type, surface complexity, particle size); habitat complexity (snags etc).	Linke *et al.* 2006; Leathwick *et al.* in press
Marine	Depth, seabed topographic features, temperature, salinity, dissolved oxygen, currents, productivity (e.g. chlorophyll-*a*, spatial temperature gradients)	Leathwick *et al.* in review

Pre-processing data: Species distributions will be best modelled where the predictors are appropriately cleaned and pre-processed. Neighbourhood measures (e.g. focal means) may help to summarise variables from a species perspective (Section 6.2), particularly for those that are mobile, and circular predictors such as aspect or month of the year (where the start and end are identical) should be converted to linear measures, e.g., using sine and cosine transformations (Flury and Levri 1999). Categorical data may have many classes (e.g., many soil types), and in regression models each class requires a modelled parameter, making use of many classes a data-hungry process. Classes should be collapsed to the smallest relevant subset, using expert knowledge or quantitative methods (Dormann *et al.* in prep).

Dealing with collinearity: Correlations between predictors will create problems both in interpreting the models and in predicting to regions where correlation structure differs, so correlations need to be identified and dealt with. The broad steps in dealing with collinear data are to assess whether collinearity is present and then to either reduce it, or to use modelling methods that algorithmically are not compromised by it (Box 6.6). However, none of these avoid the fundamental reality that statistical models identify correlation, not causation, and there is no straightforward way to decide which variables to focus on and which to ignore. The implications of this are that: (i) the relative importance of the correlated variables cannot be untangled; and (ii) predictions to novel environments (in space or time) are likely to be compromised by them having different correlation structures that present unique combinations of environments not accounted for in the model. If variable selection methods have excluded some predictors that are correlated with others, the risk is that the non-causal variable might have been selected, and changes in it will be less relevant to the species than changes in the causal variable. These problems are amplified if indirect or distal predictors are used, because then the correlations between the available predictors and the more proximate true drivers of distribution are also likely to vary (Austin and Smith 1989). There are limited methods for dealing with these problems, but careful thought and exploration is essential. For example, it could be worth testing the sensitivity of the predictions to choice of one variable over another, perhaps using a geographically or environmentally structured cross-validation.

Box 6.6. Dealing with collinearity (ideas developed in a working group led by Carsten Dormann, 2007. To be published as a review)

Detection: Metrics can be used to assess pairwise correlations (e.g. Pearson correlation coefficients, graphing pairwise variation) or multi-collinearity (e.g. Variance Inflation Factors (VIF) , Booth *et al.* 1994; Condition Indices based on Eigenvalues, Belsley *et al*. 1980).

Reducing it: Approaches include:

Identify a set (a "cluster" or "proxy set") of correlated variables, using e.g. cluster analysis iterative VIF, or principal components analysis (PCA),

Deal with the cluster: select one variable, or use PCA axes, or transform one variable in relation to the other (e.g. the normalisation methods in Leathwick *et al.* 2005)

Use a model that deals with correlations within the method e.g. sequential regression, latent root regression

Using modelling methods robustly: The problem with fitting a traditional regression model with collinear variables is that the true coefficients cannot be estimated for the collinear set, so interpretation of the relative importance of variables is confounded. If that is accepted, and if the sites at which predictions are to be made have the same correlation structure as those for model training, there is no problem. Unfortunately, changing correlation structures are not easy to quantify. It is possible (though apparently untested) that ensembles of trees (e.g. BRT, RF) might be more robust than traditional methods because the model doesn't rely on estimation of coefficients, but simpler splitting rules.

6.5.3 Fitting the model

With species data and candidate predictor variables ready, the first modelling task is to sample the predictor variables at all sites at which there are species records (and pseudo-absences, where relevant). The idea is to produce a matrix of data in which columns contain species and predictor data, and each row represents a site (Figure 6.1). For a step-by-step tutorial to fitting a model see, for example, Wintle *et al.* (2005). Here we address more general issues, especially those relevant to regression models.

Selecting variables from the candidate set: Variable selection is an important part of model fitting in many regression techniques, and a number of fundamentally different approaches are used. Options include: use of all collated candidate predictors (from Section 6.5.2) assuming that the model can sort out the relevant variables by assigning coefficients reliably (unlikely to work well unless there is a very large amount of species data); use of a variable selection algorithm such as stepwise selection (shown to be prone to errors so needs to be used with care – see Whittingham *et al.* 2006); creation of many models with differing subsets of predictors, followed by selection of the best few using information-based criteria, perhaps averaging their results (Wintle et al. 2003); or, use of all predictors followed by shrinkage of the coefficients with either global or local shrinkage methods (also called regularisation; Reineking and Schröder 2006). Subsets / model averaging methods have become popular in ecology. Newer regularisation methods are likely to gain more attention because they provide a coherent link between variable selection and coefficient estimation. Some of these are well established methods in statistics (e.g. ridge regression – Hastie *et al.* 2001) but their use is most prevalent in the data mining / machine learning community, where prediction is a key aim of model building. Examples of effective use of regularisation include the lasso for regression (Reineking and Schröder 2006), boosted regression trees formed with gradient boosting (Friedman *et al.* 2000) and MAXENT (Phillips *et al.* 2006).

Model complexity and structure: Whatever method is used, it is important to address the tradeoff between the maximization of model fit and the minimizing of prediction error, in particular avoiding the problem of over-fitting, i.e., producing a model that is adapted to the patterns in the training data to such an extent that it no longer predicts (generalises) well at new sites. The problem is partially avoided by restricting the number of candidate predictors and the complexity of modelled relationships according to the sample size of the species data. Nevertheless, there will still be some uncertainty about the relevant model structure. The situation can be described in terms of a tradeoff between bias and variance (Hastie *et al.* 2001). The main objective in fitting a predictive model is to control the process so that major trends in the data are captured but sample-specific noise is ignored. This can be achieved in various ways, but the most common is to use a stopping rule such as an information criterion (e.g. Akaike's Information Criterion, AIC) in stepwise selection routines or in subset selection, selecting the best subset of models from many, based on different combinations of predictors. Alternatively, regularisation methods reduce model complexity by reducing the influence of either all or some parameters (i.e. coefficients) in the model, usually by testing predictive performance on independent data (e.g. Friedman 2001).

Many ecological phenomena are complex, and non-linear functions are usually necessary to model them realistically. Models with only linear fitted functions are unlikely to be justifiable from an ecological viewpoint. Because of this, it is important to check that the intended software can model non-linear relationships. If the data have particular properties such as observations with imperfect detection or survey design with nested sites, specialised model structures may be required (e.g. see Box 6.5).

Use of weights: Many regression methods allow the use of weights that control the relative contributions of the presence and absence observations. This is particularly useful if the species data include pseudo-absences: a common strategy is to apply a weight of one to each presence record, and then to weight pseudo-absences so that the sum of the pseudo-absence weights equals the sum of the presence weights. This centres predictions so their mean value is 0.5. Weights can also be used more broadly to reflect relative certainties about records or to upweight records in under-sampled regions.

Model Checking: Model checking should be considered an indispensable part of the process of fitting an SDM, and is linked to the evaluation principles described earlier. Some SDM methods are easier to assess than others, with regression-based methods generally providing a range of diagnostics. There is a large literature on how to check model fit, including methods for the assessment of the relative importance of predictors, the leverage of particular observations, the shapes of the fitted functions, the and detection of any regular patterns remaining in the residuals (Belsley *et al.* 1980; Borcard and Legendre 1994). Remaining spatial autocorrelation might require specialised models (see Box 6.5). Some SDM methods are more "black-box" in nature, but it is still worth trying to interpret the model, and understanding how it is producing its predictions. Formal techniques for doing this are still evolving.

6.5.4 Predictions

Model fitting and generation of predictions for new sites are separate tasks even though they may be processed simultaneously in some programs (e.g., MAXENT). Prediction generally involves applying the model to a dataset containing values of the predictor variables for new sites (Figure 6.1). With some methods (e.g., GLM and MARS) this can be implemented relatively simply in a geographic information system (GIS). However, with more complex methods it is often best achieved within the modelling software, with the predictions then transferred to the GIS. In these cases, gridded data containing the predictors are first transferred from the GIS to the modelling environment, ensuring first that extent and cell size are identical for all predictors. Elith *et al.* (in press) and Wintle *et al.* (2005) provide a tutorial and code for boosted regression trees and GLM and GAMs, respectively, in R, including examples of this style of prediction.

Numerical interpretation: What the predicted values actually represent varies, depending on the data used to create the model. All predictions are linked to the survey methods used to collect the training data. For example, predictions from a presence-absence model indicate the probability of encountering the species in new sampling using the same method. If the data record that a species is either present or absent in a 1000m^2 quadrat with two hours of survey effort, predictions indicate the probability that the species will be observed in one two-hour session in a quadrat of that size. In reality this link to survey method is usually ignored (perhaps not always advisedly), but it is important to be aware of it. Some data sets include observations from a range of survey methods, and in these a categorical variable describing the method for each observation can be used to correct for resulting differences (e.g., Leathwick *et al.* in press).

Despite their appearance, predictions from models fitted to presence-only data do not indicate probabilities of occurrence, because these types of species records lack information on species prevalence. Presence-only models can only provide predictions of the *relative* likelihood of species presence at a site. More broadly, for any type of species data, if biases in survey effort have not been accounted for during model fitting, it is likely that predictions will provide more of an expression of survey effort than of actual species distribution. Clearly this is undesirable, and underlines the importance of accounting for bias.

Applying thresholds: Conservation planning is often done with binary maps that indicate sites as providing either suitable or unsuitable habitat, i.e., as a 0/1 variable, rather than using the more continuous predictions of probability or relative likelihood available from regression models. The reasons for discarding much of the information contained in continuous predictions are rarely convincing, but the practice still persists in many planning implementations, despite good arguments for using continuous information (Wilson *et al.* 2005) and the availability of software that allows its analysis (e.g. Moilanen *et al.* 2006a). If continuous predictions must be reduced to a binary level, a threshold needs to be chosen to achieve this conversion, and the best method will partly depend on the relative costs of different errors. For instance, if the map is to predict possible locations of a threatened species to guide survey effort prior to development, errors of omission have a high cost to conservation. Some thresholding methods can be used with explicit costs included, though examples are rare (but see Wilson *et al.* 2005). Other methods that have proved useful for some datasets include using the prevalence or average fitted values in the training data, or balancing sensitivity against specificity (Liu *et al.* 2006). Care needs to be taken when applying several of these methods to predictions from presence-only models, given that the values are not true probabilities.

Uncertainty estimates: Predictions will always be uncertain, both because the intrinsic mechanism of many models provide predictions of mean estimates at any given site, and because there are many sources of error that are unavoidable (Elith *et al.* 2002; Burgman *et al.* 2005; Ray and Burgman 2007). SDMs that ignore uncertainty oblige decision-makers to be risk-neutral (Burgman 2005), but few applications of spatial prioritisation involve symmetry in the costs of omission and commission. Conservation prioritisation is most effective when uncertainty is accounted for, and some planning software deals with it explicitly, in as far as it is possible to characterise the uncertainty (Chapter 11). Many SDM methods can produce estimated confidence intervals around their predictions, although this might involve substantial computation, e.g., using bootstrap methods. Several of the more traditional regression-based methods are able to produce estimates of parameter uncertainty more directly (Table 6.2), but see Van Niel *et al.* (2004) for limitations. Bayesian methods automatically provide information on parameter uncertainly (McCarthy *et al.* 2007). Even if the prioritisation software used for conservation planning does not deal explicitly with uncertainty information, analyses can be run separately using the upper and lower bounds of SDM predictions, so that the sensitivity of results can be evaluated. Alternatively, the sensitivity of spatial priorities to different modelling methods or to variations in choices within the modelling process (such as selection of predictor variables, treatment of species data) could be worthwhile. Because uncertainty is hard to deal with it is often ignored, but we expect that there will be increasing realisation of its importance and the benefits of exploring its impacts on prioritisation. Examples of research about impacts of species distribution uncertainty on prioritisation include Moilanen *et al.* (2006a, b).

6.6 Example: a marine fish in New Zealand's exclusive economic zone (EEZ)

As a worked example of a SDM we present models for the distribution of a small deep-sea fish species, *Mora moro,* in the oceans to the east of New Zealand. We chose this species and dataset as an illustrative example because related models and marine reserve planning applications are published and provide further detail (Leathwick *et al.* 2005; Leathwick *et al.* in review). For this study we restrict the region to the Chatham Rise and surrounds (Figure 6.3), subsampling the ~9000 trawl sites there to provide 1000 presence-absence records as a training data set for modelling and 8142 for evaluation. Although this is a data-rich example, we take the unusual step of only using a small proportion of the data for training because most spatial prioritisation applications do not have such large data sets. *M. moro* occurred in 25% of trawls. A set of 11 predictors relevant to marine benthic species were prepared with grid cells 1km by 1km, and collinearity problems were reduced by restricting the set of predictors to those having pairwise Pearson correlations less than 0.85. Correlations of depth with temperature and salinity were controlled by using residuals from GLMs fitted with natural splines (Leathwick *et al.* 2005). The predictors included estimates of trawl depth, temperature and salinity at the sea floor, primary productivity at the ocean surface, and zones of ocean mixing and tidal currents.

Model development, fit and predictive ability are detailed in Table 6.4 and Figures 6.3 and 6.4. The models from the three methods were broadly similar but with some noticeable differences. We show fitted functions for three variables (Figure 6.4). Note the way that MARS functions comprise piece-wise linear segments, compared with smooth fits in GAMS and highly detailed functions in BRT. The map from BRT (Figure 6.3a) is slightly less smeared than those for GAM and MARS, reflecting the sharper transitions in BRT fitted functions (Figure 6.4). The ability of flexible models such as BRT to predict steep gradients in responses and interactions often lead to improved predictive performance. GAMs might fit sharper transitions too, if more degrees of freedom were allowed per smooth, but we sought to avoid overfitting because there were only 245 records of presence. All models performed well when tested on excluded sites, with AUCs ranging from 0.91 to 0.94 and 42-47% deviance explained (Table 6.4). The ordering of the predictive performance of methods (BRT > GAM > MARS) is typical of other comparisons of these methods (e.g. Elith *et al.* 2006).

Table 6.4. Model details for the case study.

	Model fitting details	Evaluation on independent data
BRT	Used the gbm library in R, and custom code (Elith *et al.* in press); applied 10-fold cross-validation to identify the optimal number of trees (800), with learning rate = 0.01 and tree complexity of 5. Model identified relative importance of top five predictors as: depth (50%), temperature residuals(16%), sea surface temperature anomaly (7%), sea surface temperature gradient (6%), tidal current velocity (6%), chlorophyll-*a*(5%).	AUC = 0.93; % deviance explained = 47.0
GAM	Used the gam library in R. Fitted four models, with choice of variables broadly informed by the BRT and MARS models, and used AIC to identify the best – this had 6 predictors (depth, sea surface temperature anomaly, temperature residuals, tidal current velocity , sea surface temperature gradient, dissolved organic matter), each fitted with a cubic spline smoother with 4 degrees of freedom (df). Alternative models varied in df per smoother and number of variables.	AUC = 0.92; % deviance explained = 44.6
MARS	Used the mda library in R and custom code (Elith and Leathwick 2007). Allowed default settings. These selected 6 variables and assessed relative importance (high to low) as: depth, temperature residuals, dissolved organic matter, sea surface temperature gradient, sea surface temperature anomaly, salinity residuals.	AUC = 0.91; % deviance explained = 41.7

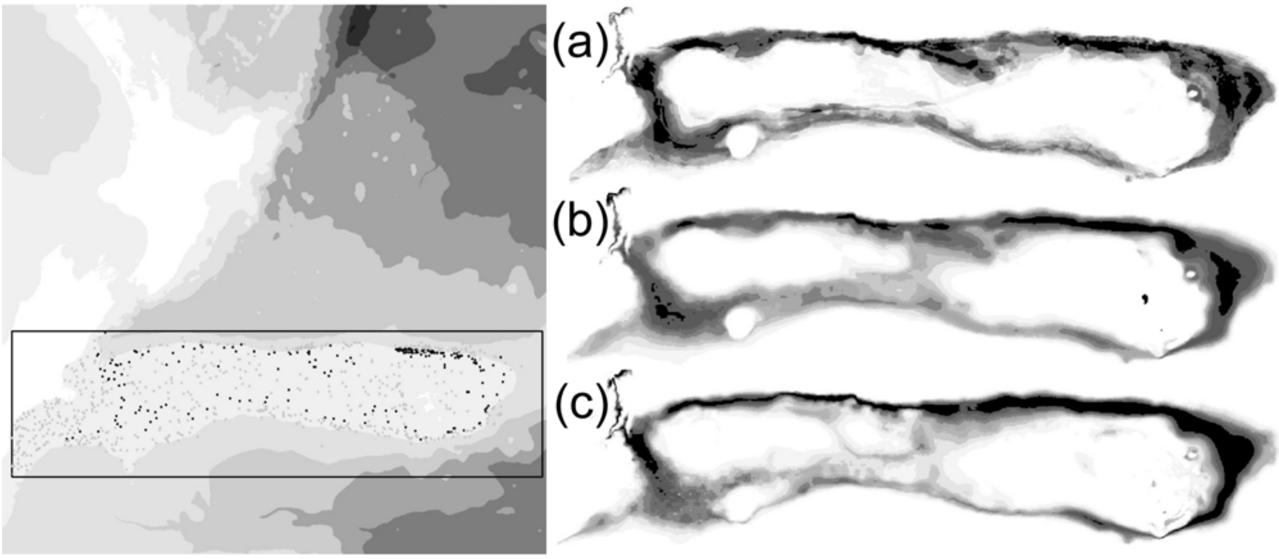


Figure 6.3. : Map of the study area (left; within box) and presence (black) and absence (grey) records used for modelling. Predictions (right), from 3 models: (a) boosted regression trees (BRT); (b) Multivariate adaptive regression splines (MARS) and (c) Generalised additive model (GAM). The grey scale indicates increasing probability of a presence, from white to black

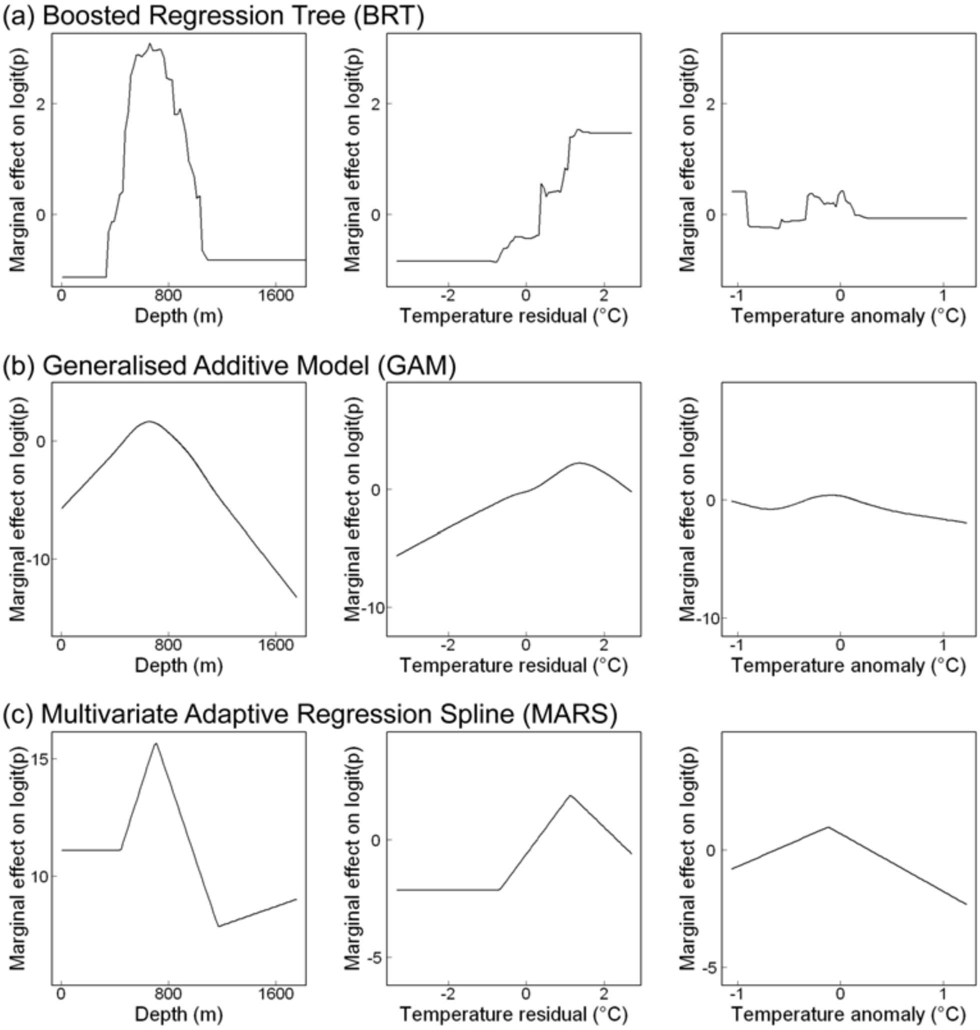


Figure 6.4: Fitted functions for three variables (columns) from the same three models as in Figure 6.3.

6.7 Limitations and future directions

This chapter has focused on the statistical and ecological assumptions and bases for SDMs, emphasising their influence on the likely quality of final predictions. Unbiased samples across geographic and environmental space, proximate predictors, models with sufficient flexibility and complexity, and a sound understanding of geographic patterns of disturbance and competition all contribute to good models. Since data are typically less than perfect and knowledge more limited than ideal, the art of building good SDMs lies in being realistic about data and model deficiencies. Where possible these deficiencies should be addressed. Those that cannot be addressed should be reflected in realistic characterisation of the uncertainty accompanying predictions.

There are limitations inherent to any modelling approach. For statistically-based SDM models to predict well, the response variable (the species data) has to have a consistent and stable relationship with measured predictors. At a new site in the same region this is likely to be a reasonable assumption, provided that the predictors are functionally well related to the species, and the model is effectively describing the observations. At a new site in a more distant region, or at a different time, this is a more problematic assumption, which is likely to become even more tenuous if the predictors have only indirect relationships with the species. Predictions that extrapolate into unsampled geographic regions are much less likely to be reliable than those that interpolate to new sites in the same region.

Prediction becomes even less reliable when extrapolation is required to locations that lie outside of the sampled environmental range, and there is little work addressing the ecological, let alone the algorithmic issues, associated with this. Most attempts to date approach the problem by controlling or characterising the uncertainty in predictions – e.g. the consensus modelling of Thuiller (2003) (and see Chapter 13), but this will provide only minimal advantages if the same assumptions are violated across the full suite of fitted models.

Process-based models, by contrast, focus on modelling causal relationships through more fundamental understanding of mechanistic relationships (B. Phillips *et al.* in review). While the usual argument is that this makes them more robust when predicting to new circumstances, it strictly only applies where the model is based on fundamental physical processes that remain constant regardless of the broader context in which they are applied. Population models can identify critical life stages, density dependence and interaction with competitors (Chapter 9), and the broader metapopulation structure (Chapter 8). These approaches offer some clear advantages, but they tend to be time-consuming to construct and require detailed information. While we acknowledge that statistical models have drawbacks, at this point in time they provide a method that can be readily implemented when predictions are required of distributions of multiple species. In addition to their predictive function, they can also provide useful insight into the likely factors driving the distributions of species. Paying proper attention to issues described here can substantially improve their effectiveness in prioritising conservation effort..

In this chapter we have presented an up-to-date view of species modelling, but acknowledge that this is a rapidly evolving field, and that new methods and resources are constantly appearing. For example, quantile regression (e.g. Vaz *et al.* in press) has useful properties for applications where predictions of potential locations are required, and may prove to be particularly useful for conservation applications. We anticipate further useful modelling approaches emerging from Bayesian and machine learning communities; increasing use of predictors derived from remotely sensed data; and further expansion of species location data in online databases. Many of the principles outlined in this chapter apply to all approaches to species modelling, regardless of the methods and data that are used. Without due regard to these principles, the proliferation of data and modelling methods may pose a threat to the reputation of SDM. In our view, progress will be most productive through improved integration of ecological insight into models, but finding ways to do this reliably when modelling the distributions of many species will be challenging. In the meantime, we would encourage modelling practitioners to maintain a balance of healthy scepticism and caution, while remaining open to new ideas.

Acknowledgements

JE was funded by ARC grant DP0772671 and the Australian Centre of Excellence for Risk Analysis; JL's contribution was funded by New Zealand’s Foundation for Research, Science and Technology under contract C01X0305. We appreciated insightful comments on the manuscript from Kerrie Wilson, Atte Moilanen, Carlo Rondinini, Terry Walshe and Brendan Wintle.

References

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

Austin, M.P. and Smith, T.M. (1989). A new model for the continuum concept. *Vegetatio*, **83,** 35-47.

Ayyub, B.M. (2001). *Elicitation of Expert Opinions for Uncertainty and Risks,* CRC Press, Boca Raton.

Barry, S.C. and Elith, J. (2006). Error and uncertainty in habitat models. *Journal of Applied Ecology,* **43,** 413-423.

Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity,* John Wiley & Sons, New York.

Bessa-Gomes, C. and Petrucci-Fonseca, F. (2003). Using artificial neural networks to assess wolf distribution patterns in Portugal. *Animal Conservation,* **6,** 221-229.

Bio, A.M.F., De Becker, P., De Bie, E., *et al.* (2002). Prediction of plant species distribution in lowland river valleys in Belgium: modelling species response to site conditions. *Biodiversity and Conservation*, **11,** 2189-2216.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning,* Springer.

Booth, G.D., Niccolucci, M.J. and Schuster, E.G. (1994) Intermountain Research Station, USDA Forest Service, Ogden, Utah, USA.

Borcard, D. and Legendre, P. (1994). Environmental control and spatial structure in ecological communities: an example using oribatid mites (Acari, Oribatei). *Environmental and Ecological Statistics,* **1,** 37-61.

Boyce, M.S., Vernier, P.R., Nielsen, S.E., *et al.* (2002). Evaluating resource selection functions. *Ecological Modelling,* **157,** 281-300.

Breiman, L., Friedman, J.H., Olshen, R.A., *et al.* (1984). *Classification and Regression Trees,* Wadsworth International Group, Belmont, California.

Buckland, S.T. (2004). *Advanced distance sampling*, Oxford University Press, Oxford, UK.

Burgman, M. (2005). *Risks and Decisions for Conservation and Environmental Management,* Cambridge University Press, Cambridge, UK.

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

Burgman, M.A., Breininger, D.R., Duncan, B.W., *et al.* (2001). Setting reliability bounds on Habitat Suitability Indices. *Ecological Applications,* **11,** 70-78.

Burgman, M.A. and Fox, J.C. (2003). Bias in species range estimates from minimum convex polygons: implications for conservation and options for improved planning. *Animal Conservation*, **6,** 19-28.

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

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

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

Davis, J. and Goadrich, M. (2006) In *Proceedings of the 23rd International Conference on Machine Learning 2006*Pittsburgh, PA.

De'ath, G. (2007). Boosted trees for ecological modeling and prediction. *Ecology,* **88,** 243-251.

De'ath, G. and Fabricius, K.E. (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology,* **81,** 3178-3192.

Dennis, R.L.H. and Hardy, P.B. (1999). Targeting squares for survey: predicting species richness and incidence of species for a butterfly atlas. *Global Ecology & Biogeography,* **8,** 443-454.

Dodd, L.E. and Pepe, M.S. (2003). Partial AUC estimation and regression. *Biometrics,* **59,** 614-623.

Dormann, C.F., McPherson, J.M., Araujo, M.B., *et al.* (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography,* **30,** 609-628.

Dormann, C.F., Reineking, B. and Schroder, B. (in prep). Consequences of pre-processing on model outcome in species distribution analyses.

Drake, J.M., Randin, C. and Guisan, A. (2006). Modelling ecological niches with support vector machines. *Journal of Applied Ecology,* **43,** 424-432.

Eken, G., Bennun, L., Brooks, T.M., *et al.* (2004). Key biodiversity areas as site conservation targets. *BioScience,* **54,** 1110-1118.

Elith, J., Burgman, M.A. and Regan, H.M. (2002). Mapping epistemic uncertainties and vague concepts in predictions of species distribution. *Ecological Modelling,* **157,** 313-329.

Elith, J., Graham, C.H., Anderson, R.P., *et al.* (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography,* **29,** 129-151.

Elith, J. and Leathwick, J.R. (2007). Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Diversity and Distributions,* **13,** 165-175.

Elith, J., Leathwick, J.R. and Hastie, T. (in press). A working guide to boosted regression trees. *Journal of Animal Ecology*.

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

Ferrier, S., Watson, G., Pearce, J., *et al.* (2002). Extended statistical approaches to modelling spatial pattern in biodiversity: the north-east New South Wales experience. I. Species-level modelling. *Biodiversity and Conservation,* **11,** 2275-2307.

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

Flury, B.D. and Levri, E.P. (1999). Periodic logistic regression. *Ecology,* **80,** 2254-2260.

Friedman, J.H. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics,* **19,** 1-141.

Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics,* **29,** 1189–1232.

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

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

Guisan, A. and Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology Letters,* **8,** 993-1009.

Harrell, F.E. (2001). *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression and Survival Analysis,* Springer Verlag, New York.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, Chapman and Hall, London.

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

Hirzel, A.H., Hausser, J., Chessel, D., *et al.* (2002). Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? *Ecology,* **83,** 2027-2036.

Huerta, M.A.O. (2007). Fragmentation patterns and implications for biodiversity conservation in three biosphere reserves and surrounding regional environments, northeastern Mexico. *Biological Conservation,* **134,** 83-95.

Insightful Corporation (2007). Splus* - statistical software.

Jarvis, A., Ferguson, M.E., Williams, D.E., *et al.* (2003). Biogeography of wild Arachis: Assessing conservation status and setting future priorities. *Crop Science,* **43,** 1100-1108.

Kohavi, R. (1995) In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence.* Morgan Kaufmann, San Mateo, CA, pp. 1137-1143.

Kremen, C., Cameron, A., Moilanen, A., Phillips, S.J., Thomas, C.D., Beentje, H., Dransfield, J., Fisher, B.L., Glaw, F., Good, T.C., Harper, G.J., Hijmans, R.J., Lees, D.C., Louis, E., Jr., Nussbaum, R.A., Raxworthy, C.J., Razafimpahanana, A., Schatz, G.E., Vences, M., Vieites, D.R., Wright, P.C., & Zjhra, M.L. (2008) Aligning Conservation Priorities Across Taxa in Madagascar with High-Resolution Planning Tools *Science*, **320**, 222–226.

Leathwick, J.R. (1998). Are New Zealand’s *Nothofagus* species in equilibrium with their environment? *Journal of Vegetation Science*, **9**, 719–732.

Leathwick, J., Moilanen, A., Francis, M., *et al.* (in review). Design and evaluation of large-scale marine protected areas. *Conservation Letters*.

Leathwick, J.R. (2001). New Zealand's potential forest pattern as predicted from current species-environment relationships. *New Zealand Journal of Botany,* **39,** 447-464.

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

Leathwick, J.R., Elith, J., Chadderton, L., *et al.* (in press). Dispersal, disturbance, and the contrasting biogeographies of New Zealand's diadromous and non-diadromous fish species. *Journal of Biogeography*.

Leathwick, J.R., Elith, J. and Hastie, T. (2006). Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological Modelling,* **199,** 188-196.

Leathwick, J.R., Rowe, D., Richardson, J., *et al.* (2005). Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. *Freshwater Biology,* **50,** 2034-2052.

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

Lehmann, A., Overton, J.M. and Leathwick, J.R. (2002). GRASP: generalized regression analysis and spatial prediction. *Ecological Modelling,* **157,** 189-207.

Lei, F.M., Qu, Y.H., Tang, Q.Q., *et al.* (2003). Priorities for the conservation of avian biodiversity in China based on the distribution patterns of endemic bird genera. *Biodiversity and Conservation,* **12,** 2487-2501.

Linke, S., Pressey, R.L., Bailey, R.C., *et al.* (2007). Management options for river conservation planning: condition and conservation revisited. *Freshwater Biology,* **52,** 918–938.

Liu, C., Berry, P.M., Dawson, T.P., *et al.* (2005). Selecting thresholds of occurrence in the prediction of species distributions. *Ecography,* **28,** 385-393.

Loiselle, B.A., Howell, C.A., Graham, C.H., *et al.* (2003). Avoiding pitfalls of using species distribution models in conservation planning. *Conservation Biology,* **17,** 1591-1600.

Longley, P., Goodchild, M.F., Maguire, D.J., *et al.* (2005). *Geographic Information Systems and Science,* John Wiley & Sons.

Mackey, B.G. and Lindenmayer, D.B. (2001). Towards a hierarchical framework for modelling the spatial distribution of animals. *Journal of Biogeography,* **28,** 1147-1166.

Manly, B.F.J., McDonald, L.L., Thomas, D.L., *et al.* (2002). *Resource selection by animals - statistical design and analysis for field studies. 2nd Edition.,* Kluwer Academic, Dordrecht.

McCarthy, M.A. (2007). *Bayesian Methods for Ecology,* Cambridge University Press, Cambridge.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models,* Chapman and Hall, London.

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

Moilanen, A., Leathwick, J.R. and Elith, J. (2008). A method for spatial freshwater conservation prioritization. *Freshwater Biology,* **53,** 577–592.

Moilanen, A., Runge, M.C., Elith, J., *et al.* (2006). Planning for robust reserve networks using uncertainty analysis. *Ecological Modelling,* **199,** 115-124.

Moilanen, A., Wintle, B., Elith, J., *et al.* (2006). Uncertainty analysis for regional-scale reserve selection. *Conservation Biology,* **20,** 1688-1697.

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

Pawar, S., Koo, M.S., Kelley, C., *et al.* (2007). Conservation assessment and prioritization of areas in Northeast India: Priorities for amphibians and reptiles. *Biological Conservation,* **136,** 346-361.

Pearce, J. and Ferrier, S. (2000). Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling,* **133,** 225-245.

Pearce, J.L., Cherry, K., Drielsma, M., *et al.* (2001). Incorporating expert knowledge and fine-scale vegetation mapping into statistical modelling of faunal distribution. *Journal of Applied Ecology,* **38,** 412-424.

Pearson, R.G., Dawson, T.P., Berry, P.M., *et al.* (2002). SPECIES: A Spatial Evaluation of Climate Impact on the Envelope of Species. *Ecological Modelling,* **154,** 289-300.

Peterson, A.T., Papes, M. and Eaton, M. (2007). Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. *Ecography,* **30,** 550-560.

Phillips, B., Chipperfield, J.D. and Kearney, M.R. (in review). Approaches to modelling the spread of invasive species.

Phillips, S.J., Anderson, R.P. and Schapire, R.E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling,* **190,** 231-259.

Phillips, S.J., Dudik, M., Elith, J., *et al.* (in review). Sample Selection Bias and Presence-Only Models Of Species Distributions. *Ecological Applications.*

Pielke Jnr, R.A. (2003). The role of models in prediction for decision In Canham, C., Cole, J. and Lauenroth, W. K., eds., *Models in Ecosystem Science.* Princeton University Press.

Prasad, A.M., Iverson, L.R. and Liaw, A. (2006). Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems,* **9,** 181-199.

Pressey, R.L. (2004). Conservation Planning and Biodiversity: Assembling the Best Data for the Job. *Conservation Biology,* **18,** 1677-1681.

Pulliam, H.R. (1988). Sources, sinks and population regulation. *American Naturalist,* **132,** 652-661.

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

R (2007). http://www.r-project.org/.

Ray, N. and Burgman, M.A. (2006). Subjective uncertainties in habitat suitability maps. *Ecological Modelling,* **195,** 172-186.

Reese, G.C., Wilson, K.R., Hoeting, J.A., *et al.* (2005). Factors affecting species distribution predictions: a simulation modeling experiment. *Ecological Applications,* **15,** 554-564.

Reineking, B. and Schröder, B. (2006). Constrain to perform: regularization of habitat models. *Ecological Modelling,* **193,** 675–690.

Rodrigues, A.S.L. and Brooks, T.M. (2007). Shortcuts for biodiversity conservation planning: the effectiveness of surrogates. *Annual Review of Ecology, Evolution and Systematics,* **38,** 713-737.

Rondinini, C., Stuart, S. and Boitani, L. (2005). Habitat suitability models and the shortfall in conservation planning for African vertebrates. *Conservation Biology,* **19,** 1488-1497.

Schimper, A.F.W. (1903). *Plant-Geography upon a Physiological Basis. English translation by W. R. Fisher,* Clarendon Press, Oxford.

Schulman, L., Toivonen, T. and Ruokolainen, K. (2007). Analysing botanical collecting effort in Amazonia and correcting for it in species range estimation. *Journal of Biogeography,* **34,** 1388-1399.

Soll, J.B. and Klayman, J. (2004). Overconfidence in Interval Estimates. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **30,** 299-314.

Stauffer, H.B., Ralph, C.J. and Miller, S.L. (2004). Ranking habitat for Marbled Murrelets: new conservation approach for species with uncertain detection. *Ecological Applications*, **14,** 1374-1383.

Steyerberg, E.W., Harrell, F.E., Borsboom, G.J.J.M., *et al.* (2001). Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology,* **54,** 774-781.

Thuiller, W. (2003). BIOMOD - Optimizing predictions of species distributions and projecting potential future shifts under global change. *Global Change Biology,* **9,** 1353-1362.

Turner, W., Spector, S., Gardiner, N., *et al.* (2003). Remote sensing for biodiversity science and conservation. *Trends in Ecology & Evolution,* **18,** 306-314.

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

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

Vaz, S., Martin, C.S., Eastwood, P.D., *et al.* (in press). Modelling species distributions using regression quantiles. *Journal of Applied Ecology*.

Venables, W.N. and Dichmont, C.M. (2004). GLMs, GAMs and GLMMs: an overview of theory for applications in fisheries research. *Fisheries Research,* **70,** 319-337.

Whittaker, R.H. (1956). Vegetation of the Great Smoky Mountains. *Ecological Monographs,* **26,** 1-80.

Whittaker, R.M., Levin, S.A. and Root, R.B. (1973). Niche, habitat and ecotope. *American Naturalist,* **107,** 321-338.

Whittingham, M.J., Stephens, P.A., Bradbury, R.B., *et al.* (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology,* **75,** 1182-1189.

Wieczorek, J.R., Guo, Q. and Hijmans, R.J. (2004). The point-radius method for georeferencing point localities and calculating associated uncertainty. *International Journal of Geographic Information Science,* **18,** 745-767.

Wilks, D.S. (1995). *Statistical Methods in the Athmospheric Sciences,* Academic Press.

Wilson, K.A., Westphal, M.I., Possingham, H.P., *et al.* (2005). Sensitivity of conservation planning to uncertainty associated with predicted species distribution data. *Biological Conservation,* **122,** 99-112.

Wintle, B.A. and Bardos, D.C. (2006). Modelling species habitat relationships with spatially autocorrelated observation data. *Ecological Applications,* **16,** 1945-1958.

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

Wintle, B.A., McCarthy, M.M., Kavanagh, R.P., & Burgman, M.A. (2003) The use of Bayesian model averaging to better represent uncertainty in predictions derived from ecological models. *Conservation Biology*, **17**, 1579–1590.

Wood, S.N. (2006). *Generalised Additive Models: An Introduction with R,* Chapman and Hall / CRC Press, Boca Raton, Florida, USA.

Worton, B.J. (1989). Kernel methods for estimating the utilization distribution in home-range studies. *Ecology,* **70,** 164-168.

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