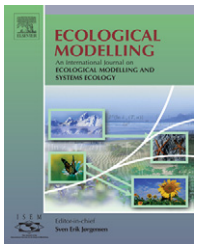
available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/ecolmodel



Review

Species distribution models and ecological theory: A critical assessment and some possible new approaches

*Mike Austin**

CSIRO Sustainable Ecosystems, GPO Box 284, Canberra City, ACT 2601, Australia

ARTICLE INFO

Article history:
Received 28 July 2005
Received in revised form
20 June 2006
Accepted 4 July 2006
Published on line 17 August 2006

Keywords:
Species response curves
Competition
Environmental gradients
Generalized linear model
Generalized additive model
Quantile regression
Structural equation modelling
Geographically weighted regression

ABSTRACT

Given the importance of knowledge of species distribution for conservation and climate change management, continuous and progressive evaluation of the statistical models predicting species distributions is necessary. Current models are evaluated in terms of ecological theory used, the data model accepted and the statistical methods applied. Focus is restricted to Generalised Linear Models (GLM) and Generalised Additive Models (GAM). Certain currently unused regression methods are reviewed for their possible application to species modelling.

A review of recent papers suggests that ecological theory is rarely explicitly considered. Current theory and results support species responses to environmental variables to be unimodal and often skewed though process-based theory is often lacking. Many studies fail to test for unimodal or skewed responses and straight-line relationships are often fitted without justification.

Data resolution (size of sampling unit) determines the nature of the environmental niche models that can be fitted. A synthesis of differing ecophysiological ideas and the use of biophysical processes models could improve the selection of predictor variables. A better conceptual framework is needed for selecting variables.

Comparison of statistical methods is difficult. Predictive success is insufficient and a test of ecological realism is also needed. Evaluation of methods needs artificial data, as there is no knowledge about the true relationships between variables for field data. However, use of artificial data is limited by lack of comprehensive theory.

Three potentially new methods are reviewed. Quantile regression (QR) has potential and a strong theoretical justification in Liebig's law of the minimum. Structural equation modelling (SEM) has an appealing conceptual framework for testing causality but has problems with curvilinear relationships. Geographically weighted regression (GWR) intended to examine spatial non-stationarity of ecological processes requires further evaluation before being used.

Synthesis and applications: explicit theory needs to be incorporated into species response models used in conservation. For example, testing for unimodal skewed responses should be a routine procedure. Clear statements of the ecological theory used, the nature of the data model and sufficient details of the statistical method are needed for current models to be evaluated. New statistical methods need to be evaluated for compatibility with ecological

* Tel.: +61 2 6242 1758; fax: +61 2 6242 1555.
E-mail address: mike.austin@csiro.au.

doi:10.1016/j.ecolmodel.2006.07.005

theory before use in applied ecology. Some recent work with artificial data suggests the combination of ecological knowledge and statistical skill is more important than the precise statistical method used. The potential exists for a synthesis of current species modelling approaches based on their differing ecological insights not their methodology.

Contents

1. Introduction 2
2. Current models for predicting species distributions 3
 2.1. Ecological theory 3
  2.1.1. Shape of species response curve 3
  2.1.2. Types of environmental response 4
 2.2. Data model 4
  2.2.1. Problem of scale and purpose 4
  2.2.2. Selection of biotic variables 5
  2.2.3. Selection of environmental predictors 6
 2.3. Statistical model 6
  2.3.1. Comparison and evaluation of methods 7
  2.3.2. Using artificial data 8
3. Alternative approaches and models 8
 3.1. Liebig's law of the minimum and quantile regression 8
 3.2. Structural equation modelling (SEM) 10
 3.3. Spatial non-stationarity and geographically weighted regression (GWR) 12
4. Conclusion: best practice? 13
 Acknowledgements 15
 References 15

1. Introduction

Statistical regression methods for quantitative prediction of species distributions are central to understanding the realized niche of species and to species conservation in the face of global change. Recently, there have been two significant conferences (Scott et al., 2002; Guisan et al., 2002), many reviews ((Franklin, 1995; Guisan and Zimmermann, 2000; Austin, 2002a; Huston, 2002), numerous methodological comparisons (e.g. Bio et al., 1998; Manel et al., 1999; Miller and Franklin, 2002; Moisen and Frescino, 2002; Munoz and Felicisimo, 2004; Thuiller, 2003; Segurado and Araujo, 2004) and several commentaries (Lehmann et al., 2002a; Elith et al., 2002; Ferrier et al., 2002; Rushton et al., 2004; Guisan and Thuiller, 2005; Elith et al., 2006) on the value, use and application of the methods. An examination of these references and others shows that there is little agreement on appropriate data, methodology or interpretation and little discussion of the conceptual framework on which species predictive models are based.

Comparisons of methods rarely use the same type of data (counts or presence/absence), use the regression method in the same way (multiple linear versus curvilinear terms) or use a common set of predictors. Evaluation of the comparisons by different authors is frequently confounded by these differences in how the methods are applied, i.e. model parameterization (Austin, 2002a). In effect, there are a number of different Kuhnian paradigms operating in this area of research (Kuhn, 1970; Austin, 1999a). Each paradigm consists of an agreed set of facts (e.g. presence data on organisms), a conceptual framework (e.g. niche theory; plants or animals), a restricted set of problems (e.g. climatic control of distribution) and an accepted array of methods (e.g. logistic regression) see also Guisan and Zimmermann (2000). There is also a tendency for confirmatory studies providing supporting evidence rather than tests of the basic assumptions of the paradigm, see Austin (1999a) for examples in community ecology. One clear indicator of the degree to which separate paradigms are operating in this field is the number of common citations in two recent review papers: precisely zero (Austin, 2002a; Rushton et al., 2004). Communication in the widest possible sense between the separate paradigms is clearly a problem and the present author has been one of the culprits contributing to the problem.

Better communication will contribute to progress, but solving problems of a technical or theoretical nature is also necessary. In this review, a three-component framework is used to examine quantitative methods for spatial prediction of species distributions. The components are: (1) an ecological model concerning the ecological theory used or assumed, (2) a data model concerning the type of data used and method of data collection and (3) a statistical model concerning the statistical methods and theory applied (Austin, 2002a). I use this framework to expand on the technical and theoretical problems limiting current practice. I then introduce some methods that have yet to be applied widely in this area and may offer

advantages, and finally offer some suggestions for improving current practice to better integrate ecological theory, statistical modelling and conservation.

Non-statistical methods of prediction such as neural nets (Fitzgerald and Lees, 1992), GARP (Stockwell and Noble, 1992) and climatic envelopes (Pearson and Dawson, 2003) are not considered though see Elith et al. (2006). While the major emphasis is on terrestrial plants, similar issues apply equally to animals (Scott et al., 2002).

2. Current models for predicting species distributions

Logistic regression is a frequently used regression method for modelling species distributions (Guisan and Zimmermann, 2000; Rushton et al., 2004). This is a particular case of Generalised Linear Models (GLM, McCullagh and Nelder, 1989). GLM has been recognised in ecology for some time as having great advantages for dealing with data with different error structures particularly presence/absence data that is the common type of data available for spatial modelling of species distributions (Nicholls, 1989, 1991; Rushton et al., 2004). Generalised Additive Models (GAM, Hastie and Tibshirani, 1990), a powerful extension of GLM are increasingly used for species modelling (Yee and Mitchell, 1991; Leathwick and Whitehead, 2001). Using the three-component framework, three questions can be asked of current papers using the regression methods GLM, and GAM:

What ecological theory is assumed or tested?
Are there any limitations imposed by the nature of the data used?
Are the statistical procedures and methods used compatible with ecological theory?

2.1. *Ecological theory*

Theory in recent papers on species distribution is usually implicit. In an arbitrary review of 20 recent papers using statistical models to predict species distributions, none used the term theory in the text in an ecological context (Journal of Applied Ecology 12 papers 2003–2004; Journal of Biogeography 5 papers 2004; one paper each from Global Change Biology (2003), Ecology Letters (2004) and Global Ecology and Biogeography (2003), * in references). An earlier comprehensive introduction to species modelling provided by Ferrier et al. (2002) does not mention theory except in a statistical context. The implicit theory assumes that species distributions are determined at least in part by environmental variables, and that reasonable approximations for these variables can be estimated. Explicit theory regarding species response to environmental gradients and resources exists (Giller, 1984; Huston, 2002). Niche theory as applied to both plants and animals assumes symmetric Gaussian-shaped unimodal curves. In plant community ecology, niche theory has an intimate relationship with the continuum concept (Austin and Smith, 1989). Current evidence supports the occurrence of unimodal response curves with various skewed asymmetric or symmetric shapes for plants (Austin, 2005).

2.1.1. *Shape of species response curve*

Species modellers, when applying GLM use models linear in the parameters (on the logit transformed scale), but usually fit only linear (straight-line), quadratic or cubic polynomial functions. Modellers using presence/absence data as part of their data model and GLM as their statistical method often do not recognise the need to define the type of functional response based on ecological theory. McPherson et al. (2004) modelling bird species in South Africa using environmental predictors derived from satellite data do not mention the functional form of their logistic regression. Fourteen of the 20 recent references reviewed used GLM. Five used straight-line models without justification (e.g. Gibson et al., 2004). Five used quadratic functions that assume symmetric unimodal responses are ecologically appropriate and other possibilities need not be investigated (e.g. Jeganathan et al., 2004; Venier et al., 2004). Mathematically, u-shaped as well as bell-shaped responses and truncated versions of these functions are assumed possible but skewed curves are considered to be inappropriate. Three papers used cubic polynomials (Thuiller, 2003; Bustamante and Seoane, 2004; Bhattarai et al., 2004). This is consistent with theory, which assumes skewed curves are possible, but cubic polynomials represent a very restricted family of skewed curves that may not accord with ecological expectations and have undesirable properties. The function may fit most of the data well but predict badly towards the limits of the data, for example, predicting a low altitude tree species above the tree line (Austin et al., 1990). Four of the 20 recent references used GAMS to overcome the problem of response functions being ill specified by theory beyond a unimodal asymmetric shape. Two were primarily concerned with comparison of methods for modelling species distributions (Thuiller, 2003; Segurado and Araujo, 2004) and two used the method for specific problems (Clarke et al., 2003; Thuiller et al., 2004). This statistical method fits a smoothing spline defined by the data and was introduced into the ecological literature by Yee and Mitchell (1991). GAM is now recognised as a versatile method for species modelling (Guisan and Zimmermann, 2000; Guisan et al., 2002; Thuiller, 2003; Segurado and Araujo, 2004) but the ecological niche theory used is often determined by the default degrees of freedom specified for the response rather than any explicit theory or hypothesis. There is an urgent need for explicit statements about the niche theory assumed in papers on species distribution modelling.

Huntley et al. (2004) use a modified version of niche theory based on Huntley et al. (1995). The approach adopted is locally weighted regression (LOWESS Cleveland and Devlin, 1988) of presence/absence data. Their version of niche theory derives from their choice of smoothing window. The authors state that the consequence of their narrow smoothing window is that the species' response curves are "spikey" and irregular in shape. They argue that this better defines the range limits of the species and that such a response is more realistic: "Furthermore, given that the fitted surface represents the "realized" distribution of the species as determined not only by its own inherent responses to the environmental gradients but also as the outcome of interactions with numerous other species, each of which is responding in an individualistic manner to these and possibly other gradients, it is likely that it is

inherently rough. The conventional "smooth" model of species response along ecological gradients relates to a "fundamental" property of the species that may rarely be expressed in nature" (Huntley et al., 1995). This is a theoretical position that deserves investigation. It is also an example of where choice of statistical procedures can also determine the ecological theory used, which then remains untested.

Truncation of the species response curve at the observed upper and lower limits of the environmental predictor can confuse discussion of the frequency of different types of response curve, e.g. Austin and Nicholls (1997). Species position along an environmental gradient has been shown to influence the shape of response detected (Bio, 2000; Rydgren et al., 2003). Conclusions about the response curve of species can only be unambiguously determined if the sampled environmental gradient clearly exceeds the upper and lower limits of the species occurrence.

2.1.2. *Types of environmental response*

Species responses will also depend on the nature of the environmental predictor and the associated ecological processes. Plant growth shows a "limiting factor" response to light and skewed unimodal response to temperature. There is an abundance of knowledge about the ecophysiological and biophysical processes that govern the relationships between species and their environment. This knowledge can be used to choose potential variables to describe species distributions (Huntley et al., 1995; Guisan and Zimmermann, 2000; Austin, 2005). Three approaches can be identified from the literature: (a) a conceptual framework based on known biophysical processes which allows consideration and selection of appropriate environmental predictors recognising three types of environmental variables indirect, direct and resource variables (Austin and Smith, 1989; Huston, 1994, 2002; Guisan and Zimmermann, 2000); (b) an alternative framework where choice of predictors is based on ecophysiological knowledge emphasising frost tolerance and growing day degrees (Prentice et al., 1992; Huntley et al., 1995); (c) environmental predictors are selected on the basis of availability and experience that the variables show correlations with species distributions and may act as surrogates for more proximal variables. Many studies adopt the third approach. Reviewing the selected papers, eight were found to have chosen environmental predictors with the implicit assumption that they were relevant, five had variables selected for a specific problem, e.g. predicting animal/vehicle accidents (Malo et al., 2004), only five explicitly considered or asserted the relevance of the predictors. Careful selection of predictors utilising existing knowledge of physiology and environmental processes would improve the interpretability and evaluation of current statistical models.

There does not appear to be a consensus on either the necessity for explicit ecological theory or what would constitute appropriate theory, when investigating species' geographical distributions (Austin, 1999b). Synthesis of current concepts into a more comprehensive theoretical framework should be possible. The complex interdependency between theory, data and statistics is clear (Huston, 2002). In the next section, this interdependency is examined where the choice of data is the primary concern.

2.2. *Data model*

A data model may have many components including definition of sampling frame, survey design, choice of attributes, attribute measurement and precision, compatibility with statistical methods and the roles of a relational database and geographic information system data management. However, four components currently figure most prominently in species spatial modelling papers, purpose, scale of study, availability of data and selection of attributes. Pragmatically, purpose, availability of data and the cost of surveys limit the types of data models that can be adopted. One strategy is to collate plot data from existing surveys adopting a minimum common dataset (e.g. Austin et al., 1990). The dataset may then consist of presence/absence data for tree species, i.e. species for which identification is likely to be reliable from plots of a specified area and known location. However, the location of existing data is likely to be biased for the purpose of the new study, though it is possible to supplement existing data with new surveys designed to compensate for the bias (Cawsey et al., 2002). A second strategy is to use location records obtained from Herbaria, Museums or Atlases (e.g. Thuiller et al., 2003a; Huntley et al., 2004; Venier et al., 2004; Graham et al., 2004). Of the relevant papers in the reference set (16), seven used survey data, six used atlas data and three used survey data aggregated to grid cells. Sampling bias and the frequent restriction to presence-only data are significant data model problems for atlas and gridcell data. Kadmon et al. (2003, 2004) have examined number of presences, and climatic and roadside biases on the performance of a climatic envelope model, i.e. using presence-only data. They conclude that 50–75 presences are sufficient to obtain accurate estimates of species distribution. Climatic bias has a negative effect on accuracy but the impact of roadside bias is much less. However, the road network in Israel is not climatically biased (Kadmon et al., 2004). Elith et al. (2006) in an extensive comparison of modelling methods using numerous datasets of presence data, confirm the predictive success of using presence data. Predictive success varies markedly between the datasets. Sampling bias will vary with species and the location of the data; in some areas climatic bias can be expected to be correlated with roadside bias. Further work is required on the use of Herbarium records.

2.2.1. *Problem of scale and purpose*

Scale and purpose are key determinants of the data model adopted, but theoretical considerations remain important in the choice of scale, biotic and environmental data. The scale at which data are available can severely restrict the purposes for which the data can be used or place caveats on the usefulness of the results for the intended purpose. Two important aspects of scale are extent and resolution (Whittaker et al., 2001; Huston, 2002; Ricklefs, 2004; Guisan and Thuiller, 2005). Extent refers to the area over which a study is carried out, while resolution is the size of the sampling unit at which the data are recorded. For example, if the purpose is to investigate the environmental realized niche of a species (Austin et al., 1990) then the extent of the study should range beyond the observed environmental limits of the species. If this is not the case, then the species responses are truncated and the actual

shape cannot be determined (Austin et al., 1994; Austin and Nicholls, 1997).

Resolution governs what variables can be measured and what processes can be hypothesised to operate in determining species distribution and abundance. Leathwick and Austin (2001) modelled tree species distribution in New Zealand where the extent was the three main islands, an area of ca. 270,000 km^2, and the resolution was a plot size of 0.4 ha for 80% of the 14,540 plots used, rest 0.04 ha. This level of resolution allowed the investigation of interactions between competition from the dominant *Nothofagus* species and environmental predictors. This contrasts with studies where resolution can limit the conclusions drawn. For example, Huntley et al. (2004) have as their general purpose to "model relationships between species distributions and climate... to use these models to predict how species potential distributions may be altered in response to potential future climate scenarios" (p. 418). Their specific purpose is to determine "if climate is indeed more influential in determining the distribution of plants than animals" (p. 418). This follows from a conclusion by Austin (2002b, p. 81) that "The ecological theory that determines the success of predictive species modelling differs radically between plant and animal ecology.... The physical environment in terms of climate and soils is clearly more important for plants". Huntley et al. (2004) compared the modelling success of individual species models for higher plants, birds and insects and concluded that predictive success is independent of trophic level and hence predictive models for plants and animals are not fundamentally different. Their conclusions are correct given their purpose and the scale and nature of their data model. The hypothesis was tested by modelling the geographical distribution of species in large areas of Europe. The species presence/absence data were derived from atlases and the three bioclimatic variables from maps. These data were then estimated for 50 km by 50 km UTM grid cells. The modelling used the theory and methods (Huntley et al., 1995) discussed above. Their conclusion is only applicable to the data model used. Using only climatic predictors at the level of resolution of 2500 km^2, only climatic effects will be detected. The results discussed in Austin (2002b) referred to studies where the extent was equivalent to the level of resolution of Huntley et al. approximately 5000 km^2, however the resolution in Austin (2002b) was 0.1 ha for plants and ca. 10 ha for the fauna. At this resolution, plant competition and animal mobility and territories will impact on distribution and interact with climate variables. The differences in scale limit the types of generalisations that can be made, and more attention needs to be given to this topic.

Scale is a problem which has received more attention from ecologists interested in spatial patterns of species richness; though see Huston (2002) for recent discussion of this topic. Two reviews, Whittaker et al. (2001) and Ricklefs (2004), examine many issues regarding biogeographical patterns that are also relevant to modelling individual species.

2.2.2. *Selection of biotic variables*

There are two aspects of the selection of biotic variables of particular current interest: the nature and measurement of the dependent biotic variable, and whether other biotic variables should be incorporated into the models as predictors.

There are three types of biotic data usually considered in spatial prediction, various measures of abundance, presence/absence data and presence-only data. The development of GLM and GAM regression methods has meant that most measures of abundance and presence/absence can now be accommodated. The use of presence data where there is no equivalent absence data is a current technical issue of importance, given the large amounts of presence-only data available (e.g. Hirzel et al., 2001; Hirzel et al., 2002; Zaneiwski et al., 2002; Engler et al., 2004; Brotons et al., 2004). The potential and pitfalls of presence-only data is the subject of much current research, and the outcome will have important practical implications for modelling for climate change and conservation evaluation (Graham et al., 2004; Elith et al., 2006).

One type of abundance data that is widely available is vegetation survey data from phytosociological relevees. However, Guisan and Harrell (2000) have shown that the cover/abundance scales used for estimating species abundance in relevees require ordinal regression techniques as the values are ranks not continuous variates. The data can be converted to presence/absence and used with logistic regression (Coudun and Gegout, 2005). Databases with large numbers of relevees for large regions are available, e.g. Gegout et al. (2005). Their extent and resolution are more suitable for niche modelling than data derived from atlases, allowing local soil variables and competition from dominant species to be included. Boyce et al. (2002) emphasise that mobile animals may not be using the entire suitable habitat at any one time and modelling their habitat requires an appropriate data model and special resource selection functions. The choice of attribute measurement for the response variable as part of the data model is a key issue at the present time.

One aspect of modelling the realized niche that is rarely incorporated is the role of biotic processes, for example competition and predation. While the importance of these processes is widely recognised, their importance for modelling the spatial distribution of species has had only limited examination. There is a long-standing debate on the importance of extrinsic and intrinsic factors in controlling animal populations (Krebs, 2001, p. 283) and another on the importance of trophic level interactions on population control (Krebs, 2001, p. 495). Response to climate as an extrinsic factor can be treated as a simple correlation analysis for prediction purposes (Huntley et al., 2004), but application of the results to changed conditions is questionable and changes in trophic interactions may be critical for predicting responses to climate change (Harrington et al., 1999). Where the spatial resolution is suitable, the choice of predictors may need to include estimates of prey abundance, nesting sites and territories for successful modelling. For example, Pausas et al. (1995) used eucalypt foliage nutrient content as a surrogate for food quality and a tree hole index as a surrogate for nesting sites for a model of arboreal marsupial species richness. At the resolution where biotic processes become important in modelling the species environmental niche, then biotic predictors will be needed. However, such predictors are frequently not available as GIS layers and so cannot be used for spatial prediction. The development of suitable spatial surrogates for such variables from more distal variables is an area that needs more investigation.

The use of biotic predictors for plants has mainly centred on introducing competition from other similar species. Leathwick and Austin (2001) modelled the distribution of tree species in New Zealand improving the fit of the GAM models by including tree density for the dominant tree genus *Nothofagus* as competition terms in models. They also demonstrated that competitive influence is a function of the environment; interaction terms between *Nothofagus* density and the principal environmental predictors, mean annual temperature and water deficit, further improved the models. In this case, suitable GIS layers were available for predicting species distribution. Austin (2002a) discusses earlier examples of the use of competition predictors based on dominant species in regression models; see also Leathwick (2002). Further examination of methods for incorporating competition terms into spatial modelling is needed.

2.2.3. *Selection of environmental predictors*

How predictors are selected depends on the ecological and biophysical processes thought to influence the biota and again the availability of data and the purpose of the model. Two approaches should be considered if we are to move away from using all possible predictors and use existing knowledge to best advantage. These are the nature of the potential predictors whether indirect or direct, and whether we have suitable ecophysiological knowledge for choosing predictors.

Recognition of indirect, direct and resource variables (Austin and Smith, 1989; Huston, 1994; Guisan and Zimmermann, 2000) can have a profound impact on how each variable is used in the modelling approach. Indirect variables such as altitude and latitude can only have a correlation with organisms through their correlation with variables such as temperature and rainfall that can have a physiological impact on organisms. Because the correlation between indirect variables and more direct variables is location specific and need not be linear, there is no theoretical expectation regarding the shape of species responses to indirect variables. Temperature and rainfall are direct variables, while resource variables are those which are consumed by organisms, e.g. nitrogen for plants or prey for carnivores. There are theoretical expectations/hypotheses about the shapes of response to these types of variables (Austin and Smith, 1989). Plant response to soil nutrients is expected to be hyperbolic where species abundance increases to a level beyond which there is no further increase, a limiting factor response. Evaluation of species niche models should incorporate a test of whether the shape of the response to an environmental predictor is consistent with expected ecological theory.

Rainfall, while having a direct effect on organisms is a distal variable, where the proximal variable might be water availability at the root hair for plants. Biophysical processes link indirect variables to rainfall and from there via water balance models to estimates of moisture stress for plants (Austin, 2005). Recognition of the nature of the predictor helps to define the type of response to be expected. The frequent inclusion of slope and aspect in species modelling studies is an example of indirect variables where the physical relationship with solar radiation is well known and can be calculated using trigonometric functions (e.g. Dubayah and Rich, 1995). Austin (2002a) reviews earlier studies where progressive incorporation of more direct and proximal predictors using water balance models rather than slope and aspect significantly improved regression models describing the environmental niche of eucalypt species. Leathwick in a series of papers has demonstrated the value of such derived proximal predictors for examining climate change and forest equilibrium at the scale of New Zealand (Leathwick, 1995, 1998; Leathwick et al., 1996; Leathwick and Whitehead, 2001).

Huntley and colleagues (Huntley et al., 1995, 2004) following Prentice et al. (1992) select their climatic predictors using a different physiological model. Three bioclimatic variables were selected based on well-known roles in imposing constraints on species distributions. The variables were mean temperature of the coldest month, annual sum of degree-days above 5 °C and Priestley-Taylor's α (an estimate of the annual ratio of actual to potential evapotranspiration) (Prentice et al., 1992). These represent direct variables acting as surrogates for cold tolerance, growing conditions and moisture stress. Resource variables were not included. The selection of the temperature variables has a better physiological rationale than the usual selection of mean annual temperature with a skewed unimodal response curve (cf. Austin et al., 1994). However, the shape of these temperature responses is unspecified beyond the possibility of a threshold effect (Huntley et al., 1995). For example, is there an optimal number of growing day-degrees for a species above which there is no further influence? In addition, the predictions resulting from using either growing degree-days or mean annual temperature may be very similar at a regional scale as these variables are often highly correlated. Growing degree-days were shown by Pausas et al. (1997) to have a quadratic relationship to mean annual temperature with an r^2 of 0.989 for 98 meteorological stations in New South Wales, Australia. A synthesis of the different approaches to selection and use of environmental predictors described above may yield more robust and ecologically more rational species models.

Error in the environmental variables is ignored in conventional regression analysis. Such error can have profound effects on the outcomes of models. Van Neil et al. (2004) have recently drawn attention to the influence of error in digital elevation models (DEM) on environmental variables. Measures such as slope and aspect are usually calculated from a DEM and then incorporated into a geographical information system (GIS) from which they are retrieved for modelling. The authors conclude that a direct variable, solar radiation may be less prone to error than the indirect variables from which it is calculated aspect and slope. The influence of error in the DEM on species models has been further investigated by Van Neil and Austin (in press). Most studies derive their environmental predictors from a GIS. The original errors generated in producing the estimates for the GIS need careful evaluation before predictors are used for modelling.

2.3. *Statistical model*

Ecologists are very dependent on collaboration with statisticians for the introduction of new statistical theory and methodology into ecology, e.g. the introduction of GAM by Yee and Mitchell (1991). Improvements continue to be made with respect to GAM (Wood and Augustin, 2002; Yee and Mackenzie,

2002). The number of different approaches and techniques discussed in Hastie and Tibshirani (1990) make it clear, however, that the full array of statistical methods has yet to be incorporated in the ecological modelling of species distributions, see also Elith et al. (2006). There are three issues regarding the evaluation of the statistical model used for spatial prediction of plant and animal species: (1) how to compare the numerous statistical methods available (2) how should the success of the modelling be assessed? (3) how should the compatibility of the statistical model with the ecological model be evaluated?

2.3.1. *Comparison and evaluation of methods*

Comparison of modelling methods is a problem as new methods are continually being introduced. Multivariate adaptive regression splines (MARS, Friedman, 1991) is one method with potential because of its handling interactions between the predictors (Moisen and Frescino, 2002; Munoz and Felicisimo, 2004). Leathwick et al. (2005) have recently compared GAM and MARS methods incorporating a number of novel procedures applied to freshwater fish. They conclude that the two methods give similar results but that MARS has computational advantages. Leathwick et al. (in press) have examined the use of boosted regression trees (BRT, Friedman et al., 2000) as a method for modelling demersal species richness. They conclude that BRT gives superior predictive performance compared to GAM even when the latter incorporates interaction terms. Phillips et al. (2006) have recently introduced a maximum entropy method (MAXENT). A distinctly different method, generalized dissimilarity modelling (GDM) has been introduced by Ferrier et al. (2002). Elith et al. (2006) provides a description of each of these and other methods, further references and information on available software. Other recent comparisons of methods include Maggini et al. (2006), Moisen et al. (2006), and Drake et al. (2006); all differ in data models and selection of statistical models.

The comparisons of methods undertaken by different authors are rarely if ever comparable. For example, two recent comparisons of methods (Araujo et al., 2005; Elith et al., 2006) compare 4 and 16 methods respectively but only two are in common. They have different purposes, evaluation of models using presence/absence data from different time periods for predicting climate change (Araujo et al., 2005) and using regression methods with presence-only data to maximise use of herbarium and museum records of organisms. Both use GLM with the polynomial expansion x, x^2, x^3, hence any possible comparison is based on the particular procedure not the general method. There is no reason why GLM could not have used the sequence x, square-root x, see Austin and Cunningham (1981). Comparisons of GLM with other methods are confounded with the particular polynomial function used with GLM in these papers. Similarly, both groups of authors use GAM with four degrees of freedom for smoothing though this is not the only option, see Wamelink et al. (2005) for an example of alternatives.

Araujo et al. (2005) emphasise the need to use independent data for evaluation. They describe three methods of evaluation (they use the term validation): resubstitution where the same data is used to calibrate the model and measure the fit; data-splitting where the data is split into two at random, a calibration set and a evaluation set, and independent validation where a totally independent data set from a different region is used. Data-splitting is the current preferred method but individual observations may still show spatial auto-correlation (Araujo et al., 2005). The third alternative does not seem plausible. Separate regions with the same species complement, ranges and combinations of environmental predictors and ecological history simply do not occur. Elith et al. (2006) adopt a useful compromise. They calibrate with one set of data then evaluate the fit with totally separate data collected independently from the same region. Given their purpose of evaluating the use of presence data, the calibration set is presence-only and the evaluation set is presence/absence, however the results can be expected to apply to other data.

The comparative evaluation of Elith et al. (2006) is the most comprehensive to date. Three groups of methods are recognised with different levels of predictive success, the newest methods BRT, GDM and MAXENT are the best, followed by MARS, GLM, GAM, and new version of GARP (OM-GARP), while other methods, e.g. GARP, LIVES, BIOCLIM and DOMAIN are less satisfactory. The evaluation is based on AUC the area under the receiver operating characteristic (ROC) curve and the point biserial correlation (Elith et al., 2006). Both measure predictive success using an independent data set.

Current best practice for assessing model success for presence/absence data is AUC (Pearce and Ferrier, 2000; Rushton et al., 2004; Thuiller, 2003), while a number of different measures are used for quantitative data (Moisen and Frescino, 2002; Moisen et al., 2006). However, all the procedures depend on the relationship between observed and predicted values; that is on predictive success not on explanatory value. Is the shape of the predicted response curve for an environmental predictor ecologically rational? While this question can be addressed for an individual species model, when large numbers of species are modelled in one publication, presentation of such graphs is not possible, limiting evaluation of the modelling results. Elith et al. (2006) in their online appendix do provide maps of the predicted distributions for several models for a few species. It is apparent that models with the same or very similar AUC value may predict very different patterns of distribution. Reliance on AUC as a sufficient test of model success needs to be re-examined (Termansen et al., 2006).

When maps of observed and predicted geographical distributions are presented, evaluation of the environmental predictor model is still not possible (Thuiller, 2003; Thuiller et al., 2003a; Huntley et al., 2004). Some provide the degree of the polynomial fitted for each predictor (Bustamante and Seoane, 2004; Thuiller et al., 2003b) but not their values or signs. A few provide the model coefficients so that the response shapes could be reconstructed (Venier et al., 2004). Lehmann et al. (2002a) and Elith et al. (2005) provide suggestions on practical options for presenting results graphically for evaluation. Maggini et al. (2006) and Van Neil and Austin (in press) are unusual in presenting both response curves and map predictions. Even if predictive success is high, this does not necessarily mean that the shape is rational. The fitting of a cubic polynomial for predictor may account well for a skewed response at low values but also predict high probabilities of occurrence at high predictor values due to the second inflection point in the curve. This will occur if observations at high predictor levels are sparse even if there are no presences in that region.

Austin et al. (2006) provide examples. Ultimately, the decision rests on whether prediction is the sole purpose or whether ecological rationality is needed when using models for estimating the impact of climate change and similar purposes.

2.3.2. *Using artificial data*

The major difficulty with evaluating statistical methods and their compatibility with ecological theory is that the true model is unknown. Comparative evaluations on real data are unsatisfactory because two statistical methods may give different models but both may be half-right. The objection to artificial data is that current theory regarding species response curves is simplistic and unrealistic. A statistical model which fails to recover the correct structure even from data constructed on the basis of a simple theory is unlikely to recover useful information from real data. Artificial data allows questions such as: are the correct predictor variables selected, are the correct response shapes found, what are the most appropriate statistical tests to use and what are the implications of using direct or indirect environmental predictors (Austin et al., 2006)?

For example, Bio (2000) generated artificial data sets with three predictors and species response shapes assuming bell-shaped responses to examine relative performance of the model selection criteria. She found that the likelihood ratio test performed better than Akaike's information criterion (AIC) and the Bayesian information criterion at recovering the true model. The use of AIC produced more complicated models than the true model. This generalisation however was sensitive to the position of the species on the simulated gradient. Maggini et al. (2006) also examine this issue with similar conclusions.

Austin et al. (2006) consider the use of artificial data for evaluating statistical models where the true model is known in detail (Austin et al., 1995 provide greater detail on data construction). They generated sets of artificial data based on two theories of species response shape, and two sets of predictors direct and indirect with explicit biophysical relationships between them. The two theoretical models were Swan/ter Braak model which assumes equally spaced bell-shaped resonse curves along environmental gradients and the Ellenberg/Minchin model which allows for a range of skewed response curves (Austin et al., 2006). The environmental gradients assumed were based on direct variables and their associated indirect variables. For example, aspect and slope determine the radiation climate of a site: using radiation (a direct variable), or aspect, (an indirect variable) will give rise to very different types of model. Fitting radiation as a predictor may result in a unimodal quadratic response function for a species. The equivalent model using aspect in place of radiation requires a bimodal response curve for the same species (Austin et al., 2006). Results also indicate that a random variable can easily be incorporated into a GLM or GAM model as a significant predictor. However, inspection of the shape of the response and its standard error would lead to recognition of the problem. The response is flat and not obviously different from zero. Skilled analysis of results and residuals is necessary to "discover" the "true model" for an individual species. The fitting of large numbers of models for numerous species with default settings for the method is not a process which will find the most appropriate ecological model. Conventional modelling with indirect variables was found to be less successful than with direct variables and could lead to irrational response curves.

The main conclusion from the study described above is that successful recovery of the true model depends more on the ecological insight and statistical skill of the modellers than the particular statistical modelling method used (Austin et al., 2006). However, the work raises significant issues of how to design and evaluate comparative studies of this kind where success is not simply based on predictive ability.

3. Alternative approaches and models

The review above concerns the current paradigms being used in the spatial prediction of species distributions and estimation of their realised environmental niche. The question needs to be addressed, are there other approaches being used in ecology and statistics which could improve our research? Are there statistical methods that have been neglected but are more consistent with ecological theory, our knowledge of biophysical processes or the problem of spatial autocorrelation? Three approaches requiring more attention are discussed below.

3.1. *Liebig's law of the minimum and quantile regression*

Huston (2002), in an introductory essay at the conference on "Predicting Species Occurrences: Issues of accuracy and scale" (Scott et al., 2002), drew attention to the impact of assuming Liebig's Law of the Minimum is operating when modelling species response to environmental predictors and then predicting their spatial distribution. Van der Ploeg et al. (1999) provide a modern account of the origins of the Law. One statement of the Law would be "plant growth will be limited by the nutrient in shortest supply even when other nutrients are abundant, giving rise to a hyperbolic response curves to individual nutrients". If other resources are less than optimal for some observations as is usual for field observations as opposed to experiments, observed species performance will be less than the maximal possible response to the first resource. In fact, typical regression analyses fitted to the mean values may not even reflect the true shape of the response to the first resource.

A number of authors have been addressing this type of response problem in a variety of ecological contexts. Kaiser et al. (1994) drew attention to the "Law of the Minimum" problem with respect to phosphorus limitations to algal biomass as measured by chlorophyll content in lakes, arguing that simple least-squares linear regression was inappropriate. Thomson et al. (1996) also argued that: "Conventional correlation analysis... fundamentally conflicts with the basic concept of limiting factors" in a study of the spatial distribution of *Erythronium grandiflorum* (Glacier lily) in relation to soil properties and gopher disturbance. Scharf et al. (1998), in investigating patterns between prey size and predator size in animal populations, considered a related problem estimating the upper and lower boundaries of scatter plots of the two variables. These three groups of authors all used

different analytical approaches, while acknowledging that their methods were not entirely satisfactory. Cade et al. (1999) then presented the use of quantile regression (see also Scharf et al., 1998) for the purpose of estimating envelope curves or "factor-ceiling responses" (Thomson et al., 1996). Cade and Noon (2003) provide a clear exposition of the statistical method and its potential for ecological analysis of observational data for both plants and animals. Figure one provides an artificial example, showing the expected response under the equivalent of experimental conditions (Fig. 1a) and then as expected in observational field studies (Fig. 1b). Estimates of the slopes of the responses based on the 10th or 50th (approximates the least-squares regression) quantiles would be much lower than those from the 90th would. Huston (2002) shows a similar but more complex example. This approach is

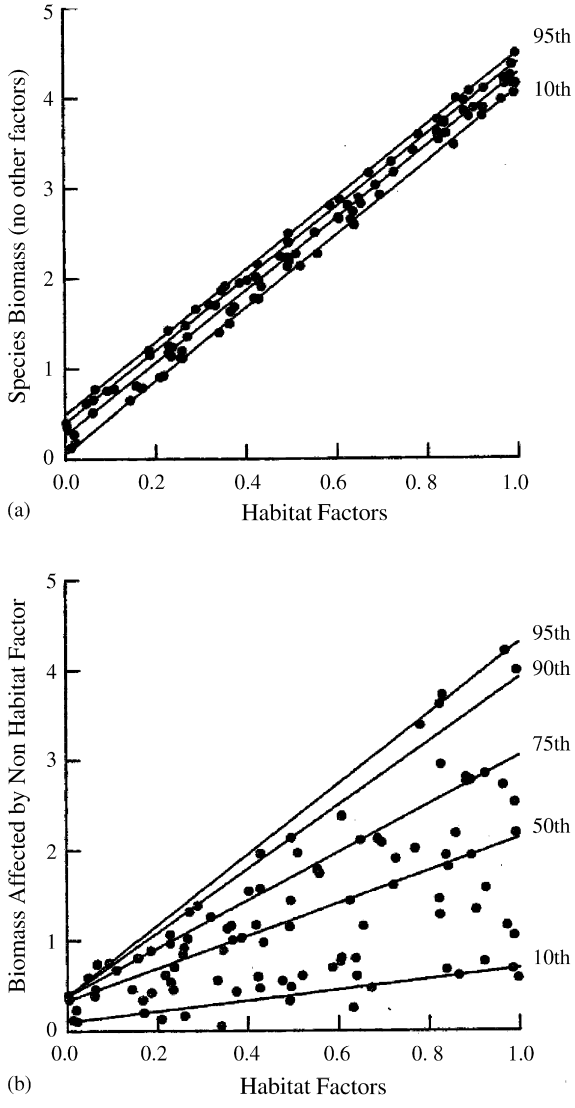


Fig. 1 – Artificial example of biomass relationship with habitat condition showing 10th, 50th, 75th, 90th and 95th quantiles estimated from quantile regression. (a) Biomass/habitat relationship uninfluenced by non-habitat factors. (b) Biomass/habitat relationship when influenced by non-habitat limiting factors. From Cade et al. (1999) with permission of Ecological Society of America.

not simply about the failure to specify an important predictor in the regression model but also that such variables almost invariably reduce the abundance of the dependent variable and may thus obscure the nature of the relationship.

Cade and Guo (2000) used quantile regression to study seedling survival of desert annuals. By estimating envelope response curves (95th and 99th quantiles) of final summer seedling density against initial winter density, they were able to interpret seedling survival as being determined by seed supply at low initial densities and by competitive self-thinning at high densities, consistent with a particular mechanistic model. This result could not have been obtained from conventional statistical analysis of the scatter plots due to the numerous low survival values resulting from other unrecorded limiting factors. Krause-Jensen et al. (2000) examined Eelgrass (*Zostera marina*) abundance and growth along a water depth gradient using a less satisfactory method for estimating "upper boundaries" (Blackburn et al., 1992) than quantile regression. They did, however, fit quadratic functions for Eelgrass biomass and cover in relation to the indirect environmental gradient of depth. Knight and Ackerly (2002) investigated variation in the average nuclear DNA content of species across a direct environmental gradient, July maximum temperature (Fig. 2). The scatter plot (Fig. 2a) shows a unimodal envelope curve with other limiting factors reducing DNA content at intermediate temperatures. The changing relationship is clearly summarised by the plot of the changes in sign, value and significance of the quadratic term when regressions for various quantiles are calculated (Fig. 2b). While the normal least-squares polynomial was significant with a negative quadratic coefficient, the strength of the unimodal response was only captured when the small values of DNA influenced by the other non-specified limiting factors were down-weighted in the upper quantile regressions.

Schroder et al. (2005) apparently provide the first example of the application of quantile regression to species abundance data in relation to environmental gradients. They compare quantile regression curves for 95% quantiles with the mean response curves using the non-linear response functions of Huisman et al. (1993). The response curves for fen plant species in relation to single predictors like annual flooding duration and phosphate show some dramatic differences between the quantile and mean curves. The quantile response curves appear more ecologically rational without abrupt thresholds and unexpected shapes. The authors do not make a link with Liebig's Law.

In fact, Liebig's Law is not the only ecological hypothesis that has been put forward to explain species physiological responses to nutrients in general (Rubio et al., 2003). These authors contrast Liebig's Law with the "multiple limitation hypothesis" (MLH Bloom et al., 1985) which says a plant's adaptive growth will result in all resources limiting plant growth simultaneously. The basic assumption of MLH is that resources are substitutable for each other at least to some extent (Bloom et al., 1985; Rubio et al., 2003). Rubio et al. (2003) tested experimentally whether responses to pairs of mineral nutrients were consistent with either hypothesis. They found that it depended which pairs of nutrients were compared, some showed a Liebig response, some a MLH response and some were indeterminate. If Liebig's law operates for

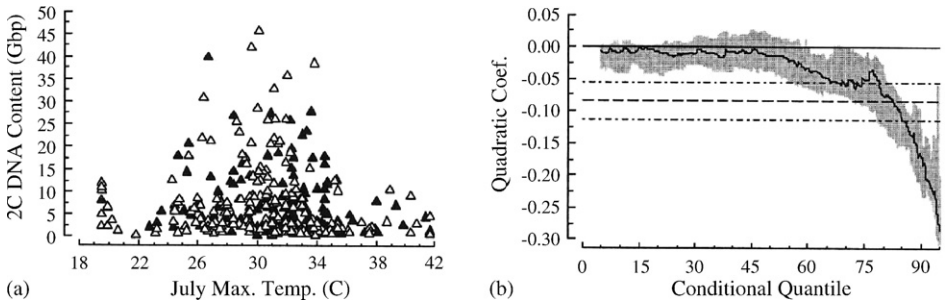


Fig. 2 – (a) The relationship between cell nuclear DNA content (2CDNA) of species and average July maximum temperature within the range of the species. (b) The change in value of the coefficient for the quadratic term (solid dark line) in the progressive quantile regressions calculated for the data in (a). Note the abrupt change above the 75th quantile. The single dashed line is the estimate for the quadratic coefficient for the least squares regression for the total data and the double dashed lines are the 95% confidence limits. The grey area represents the 95% confidence interval for the quantile regression estimates. From Knight and Ackerly (2002) with permission of Blackwell Publishers.

any predictor then there is a strong justification for quantile regression. However, the shape of a species response when expressed as a mathematical function implies an ecological theory and the opportunity to test one or more associated hypotheses. The papers of Huston, 2002, Knight and Ackerly (2002) and Schroder et al. (2005) provide a strong case for the use of quantile regression for modelling species environmental responses. The experimental study of Rubio et al. (2003) contrasting ecophysiological theories demonstrates that the choice of mathematical function should not depend on default options in a software package, nor assume a specific theory like Liebig's Law applies in all cases.

3.2. *Structural equation modelling (SEM)*

Structural equation modelling has been advocated and used in a variety of ecological contexts, ecological genetics and evolution (Mitchell, 1992, 1994), ecosystem function and toxicology (Johnson et al., 1991), comparative ecophysiology (Shipley and Lechowicz, 2000), trophic interactions (Marquez et al., 2004), plant species recruitment (Garrido et al., 2005) and rare species conservation (Iriondo et al., 2003). Vile et al. (2006) have applied SEM to study changes in species functional traits during old-field succession. Arhonditsis et al., have combined SEM with Bayesian analysis to examine the role of abiotic and biotic processes on phytoplankton dynamics and water clarity in two lakes. McCune and Grace (2002) provide a detailed introduction to its use in ecology. It does not appear to have been used to model the spatial distribution of individual species.

Shipley (2000) provides a general definition: "SEM models represent translations of a series of hypothesised cause-effect relationships between variables into a composite hypothesis concerning patterns of statistical dependencies". He presents the advantages of SEM as "... can test models that include variables that cannot be directly observed and measured (so-called latent variables) and for which one must rely on observed indicator variables that contain measurement errors". Potentially, SEM overcomes many of the problems of conventional multiple regression. For example, in conventional regression, environmental predictors are assumed to be measured without error. By explicitly recognising that correlations between variables may reflect causal pathways and that such variables may have both direct and indirect effects on a dependent variable, SEM can differentiate between alternative regression models. In fact, an SEM model of a complex set of pathways describing how environmental variables (e.g. Fig. 3) may affect each other and the dependent variable can be tested for consistency with the observed data. A hypothesised set of causal pathways can be rejected if it is not consistent with the observations.

Shipley (2000) lists the disadvantages of SEM as functional relationships must be linear, non-multivariate normal data are difficult to treat, and large sample sizes are needed. SEM would provide a means of incorporating knowledge about indirect, direct and resource variables (Austin and Smith, 1989) into a hypothesis about the causal pathways linking environmental variables, biotic influences, e.g. competition and herbivory with the distribution of species. The possibility of estimating explicitly latent variables also has considerable potential. Latent variables could be estimates of variables more proximal in the causal path than those we can measure (Austin, 2005) and use in multiple regressions.

However, if we equate species richness per plot to the abundance of a species, considering it to be controlled by the same biotic and abiotic variables then a significant example has been published, Grace and Pugesek (1997), which is further explained in McCune and Grace (2002). These authors examine plant species richness as a function of plant biomass, disturbance and abiotic variables in a coastal wetland. They recognised that abiotic variables such as soil salinity might have a direct effect on species richness and an indirect effect via an effect on plant biomass per plot, which then affects species richness, possibly by reducing light beneath the canopy. They established an initial SEM model based on such hypotheses (Fig. 3a), partitioning the correlations between the variables to simultaneously fit the entire data set, not just the dependent variable species richness. The terminology used is complex. The following description is based on Grace and Pugesek (1997). Measured variables (e.g. soil carbon) are referred to as indicators of the latent variables (e.g. soil infertility). The relationships between the indicator variables and latent variables constitute the measurement model,

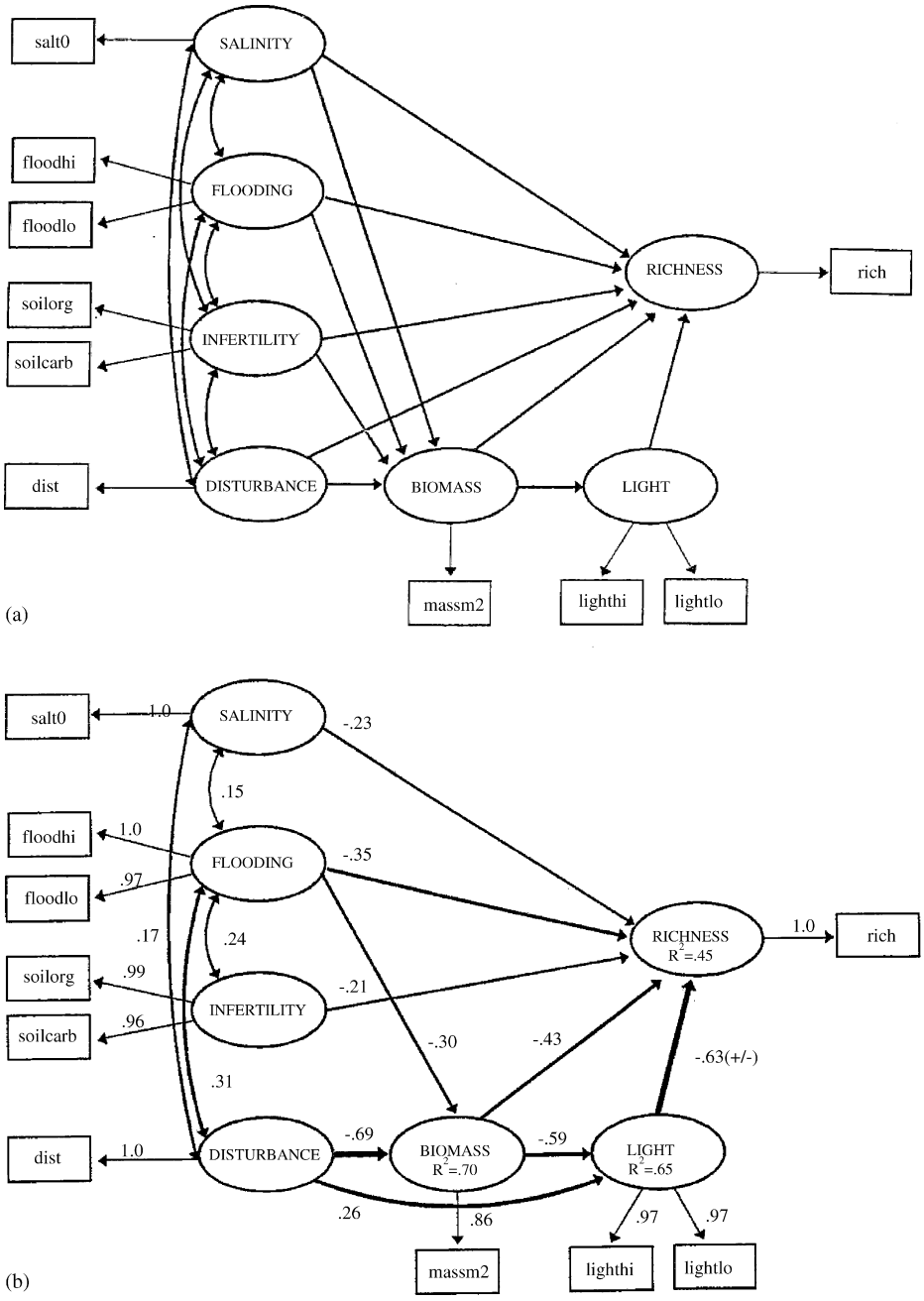


Fig. 3 – Structural equation model (SEM) for species richness in a coastal wetland. (a) Initial conceptual model. Latent variables are enclosed in ellipses and indicated (estimated) by measured or indicator variables shown in boxes. Arrows represent possible path coefficients. See Grace and Pugesek (1997) for further details. (b) Final specific model. The path coefficients represent standardised partial regression coefficients. Arrows between latent variables and indicators represent the degree to which indicators are correlated with latent variables. Pathways between latent variables show the direction, sign and partial regression coefficients. The pluses and minuses behind the path coefficient for the light-to-richness path serve as a reminder that this path has strong positive and negative components since it is the transformation of a hump-shaped relationship. The endogenous variables biomass, light and richness are shown to have 70%, 65% and 45% of their variance explained by the model. (Figs. 3 and 6 from Grace and Pugesek (1997) from American Naturalist with permission).

while the relationships between the latent variables are known as the structural model. There are two kinds of latent variable, exogenous those which only predict other variables (e.g. flooding) and endogenous those which are dependent on other variables (e.g. species richness). For further details of the statistical procedures used, see McCune and Grace (2002). The results of SEM based on Fig. 3a are shown in Fig. 3b. Note that while flooding is estimated to have an indirect effect

through correlation with biomass there is no evidence that salinity does. Disturbance has an effect only through biomass and light at ground level. Species richness is seen as a direct and indirect function of environmental variables, the biotic variable biomass and its dependent variable light, plus an indirect function of disturbance. If abundance of an individual species were substituted for species richness in Fig. 3b, the SEM model would appear an entirely feasible approach to modelling individual species distribution. It would have the added advantage of making explicit the relationships between indirect, direct and resource variables.

The recent papers on the determinants of species richness in different plant communities (Grace and Pugesek, 1997; Weiher, 2003; Weiher et al., 2004) provide interesting results on the relative importance of environmental variables and biomass in influencing species richness in different communities. However, these authors use bivariate curvilinear regression to provide functions to linearise the relationship between the principal predictors and species richness as the dependent variable prior to analysis. This assumes that no other predictor is masking the shape of the relationship. There is an urgent need to evaluate the impact of non-linear relationships (*sensu lato*) and effects arising from concepts like the Law of the Minimum on SEM, before it is widely used in modelling species distribution. Tests with artificial data based on current ecological theory would provide a suitable initial approach.

3.3. *Spatial non-stationarity and geographically weighted regression (GWR)*

Spatial autocorrelation, where the abundance or occurrence of species is correlated with presence and abundance of the species nearby, can affect statistical modelling (Cressie, 1993). Specific account of this has been incorporated into species modelling (Smith, 1994; Leathwick, 1998). More recently, the problem of whether the statistical model remains constant over the spatial extent of a study has been raised (Osborne and Suarez-Seoane, 2002). A statistical procedure, GWR has been developed to examine specifically this issue (Fotheringham et al., 2002). Biologically, this approach could be of importance as it is a local technique that allows the regression model parameters to vary in space. If species are not in equilibrium with their environment, or if the social behaviour of animals changes with location, then statistical models based on local regions may provide more information and better predictions than a global model based on data from the whole study area (Osborne and Suarez-Seoane, 2002; Foody, 2004).

A simple ecological example is provided by Foody (2003) where the normalised difference vegetation index (NDVI), a remotely sensed measure of vegetation productivity, is related to rainfall for North Africa and the Middle East (Fig. 4). An example of avian species richness prediction from three environmental variables (maximum NDVI, mean annual temperature, total annual precipitation) for sub-Saharan Africa shows marked spatial variation in regression coefficients (2004). There are clearly major changes in the local regressions and these vary progressively across southern Africa. There is controversy over the relative importance of GWR versus global

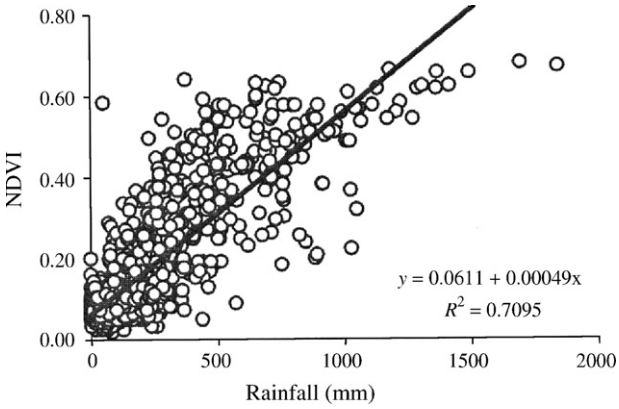


Fig. 4 – Relationship between normalised difference vegetation index (NDVI) and rainfall for 1987 from North Africa and Middle East showing a straight-line ordinary least squares regression. Note curvilinear scatter of data points. (Fig. 2a from Foody (2003) Remote Sensing of Environment with permission).

spatial regression compare Jetz and Rahbek (2002) and Foody (2004) for the same species data and see also Jetz et al. (2005) and Foody (2005a) This regression method has yet to be used for modelling individual species and needs to be reviewed carefully before being used.

I use the NDVI example (Foody, 2003) to consider some of the problems. Figure four shows the straight-line relationship fitted to the global data set. However, one would expect that above a certain rainfall NDVI would be unresponsive to rainfall, an application of Liebig's Law of the Minimum. An equally parsimonious conventional least-squares regression would be to fit a reciprocal function (1/x). This would approximate an ecologically rational response and fit the data better (Fig. 5a). However, as Huston (2002) has pointed out if a limiting factor response is theoretically appropriate then quantile regression model is the statistical model to use (Fig. 5b).

If the expected relationship is curvilinear, the application of GWR poses a problem. Both NDVI and rainfall show spatial autocorrelation. Fitting a suitable conventional regression model to the variables may well result in residuals with no remaining spatial autocorrelation, all other things being equal. However, when a local regression is fitted with observations weighted by distance from the location, bias can result because of the spatial autocorrelation in the predictor. In a high rainfall location, highly weighted observations close to the location will also have high rainfall and conversely in low rainfall locations neighbouring observations will have low rainfall. The consequences in the curvilinear response model could well be as shown in Fig. 5c. In low rainfall regions, a steep linear regression while in high rainfall areas, a flat, non-significant regression, due not to non-stationarity in the relationship but to a curvilinear relationship and spatial autocorrelation in the predictor. Foody (2004, 2005b) investigating bird species richness in Sub-Saharan Africa and Britain, respectively, using NDVI and temperature fits only straight-line functions. Unimodal responses are characteristic of species responses to climatic predictors. In such circumstances, GWR will effectively subsample limited ranges of

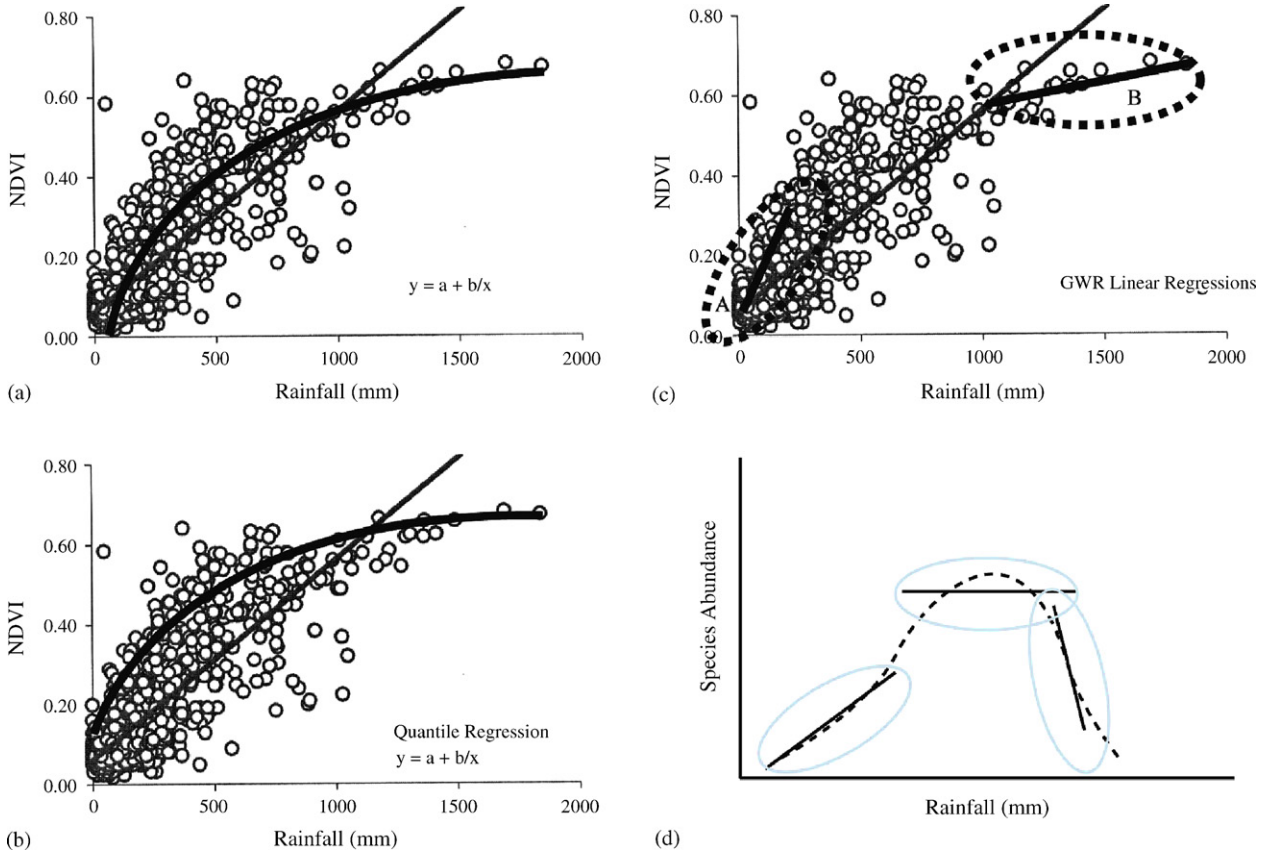


Fig. 5 – Alternative approaches to analysis of normalised difference vegetation index (NDVI) and rainfall for 1987 from North Africa and Middle East. (a) Parsimonious curvilinear regression ($y = a + b/x$). (b) Possible 95% quantile parsimonious curvilinear regression ($y = a + b/x$). (c) Potential linear geographical weighted regressions (GWR) from (A): low rainfall region, and (B): high rainfall region. (d) Potential linear geographical weighted regressions (GWR) for a species showing a unimodal relationship with rainfall from regions with different levels of rainfall (a–c) modified from Foody (2003) see Fig. 4.

the species distribution producing apparent non-stationary in the underlying process (Fig. 5d).

The NDVI/rainfall example captures a number of issues relevant to spatial modelling of species:

(1) the use of straight-line regression that is linear in the variables is inappropriate, when the data and theory suggest a curvilinear response;
(2) further consideration of the limiting factor theory clearly relevant to NDVI suggests that quantile regression would be the preferred statistical model;
(3) methods of spatial autocorrelation and non-stationarity of processes after allowing for curved responses require further investigation, not least because of their importance for testing the assumption that species distributions are in equilibrium with current environments.

4. Conclusion: best practice?

From the papers cited in this review, it is clear that there is no standard for current best practice when modelling species environmental niche or geographical distribution, whether plant or animal. Numerous incompatibilities between the ecological, data and statistical models can be identified. New ideas on how to proceed, such as Huston's (2002) suggestion of using Liebig's Law of the Minimum and quantile regression, Grace and Pugesek's (1997) ideas on the use of SEM and the role of GWR (Foody, 2004) to investigate spatial dependency, all require further development and investigation before they can be used on a routine basis. Can any recommendations be made?

There are now numerous reports that skewed response curves are frequent (Bio et al., 1998; Ejrnaes, 2000; Rydgren et al., 2003), supporting expectations from ecological theory. Best practice would therefore be to test for such responses using a GAM model or similar procedure and not assume straight line or quadratic functions without explicit theoretical justification. GRASP (Lehmann et al., 2002b) is one software package that provides a series of tools for exploring possible responses before modelling. Recent applications of new statistical procedures for modelling (Moisen et al., 2006; Munoz and Felicisimo, 2004; Elith et al., 2006) expand the potential for examining the complex curves and interactions that may be postulated by ecological theory (Austin and Smith, 1989) or detected by exploratory modelling. Understanding the interrelationship between ecological theory, statistical theory and the relative

performance of statistical models is a complex issue. Artificial data offers one means of examining these issues making explicit both ecological and statistical assumptions (Bio, 2000; Austin et al., 2006). Their results indicate that all statistical procedures should be tested with realistic artificial data before being adopted as current practice. However, Austin et al. (2006) also conclude that ecological insight and statistical skill are more important than the precise methodology used when searching for the true model in artificial data.

Defining best current practice for modelling species distributions faces great problems (Huston, 2002; Cade et al., 2005). Ecological theory suggests that environmental predictors must be evaluated in terms of Liebig's Law of the Minimum, the multiple limitation hypothesis and the expected shape of response. The Law will apply to some resource variables; others may be substituteable (Rubio et al., 2003) However, nutrient resources can also occur in toxic excess. Response to the direct but distal variable, temperature may reflect frost damage at low temperature while high temperature damage reflects protein denaturation. Exactly how predictors are likely to influence species will depend on whether they are indirect, direct or resource variables, proximal or distal, abiotic or biotic (Austin and Smith, 1989; Grace and Pugesek, 1997). Regression modelling rarely if ever examines correlation of variables in terms of process. Recognition of the type of variable can have a dramatic effect on the type of response curve which might be expected (Austin et al., 2006). Setting up a SEM with a detailed hypothetical path analysis (Fig. 3) when selecting predictors would make explicit the nature of the predictors and their likely inter-dependence. The possibility of using known biophysical process knowledge to estimate more proximal latent variables could be assessed against depending on indirect surrogate variables such as slope.

The different ecophysiological assumptions currently influencing selection of environmental predictors (Austin and Smith, 1989; Huntley et al., 1995; Leathwick and Whitehead, 2001; Huston, 2002) need to be justified in more detail, e.g. Bloom et al. (1985) and Rubio et al. (2003). The expectation is that predictors representing light, nutrients, water and temperature will influence plant distribution. Excluding predictors for one of these factors needs to be justified. Ecological judgement will have to be exercised over what to include and what needs to be justified.

One example of exercising ecological judgement when integrating ecological theory with statistical models is introducing competition between plant species. Logically, the most appropriate statistical model would be simultaneous regression where interaction (competition or facilitation) coefficients between all species are estimated along with the environmental predictors (Brzeziecki, 1987). However, the plot size of vegetation data is usually large relative to the size of individual plants and in most cases, rare species will not occur adjacent to each other and hence are unable to compete. In such circumstances, competition coefficients are inappropriate and cannot be calculated. As species abundance increases relative to plot size, the probability of species occurring as neighbours will increase and species interactions become more likely. Identification of the possible occurrence of competition using regression models will be a function of plant sizes, abundances and spatial pattern relative to plot size. This reasoning provides an explanation of the success of those regression models of a species discussed by Austin (2002a) where the model fit increased dramatically when the vegetation was stratified by plant community and the abundance of the dominant species of each community introduced as an additional predictor of the realised niche of the species. Leathwick (2002) has demonstrated both competitive and facilitative interactions between the dominant *Nothofagus* species conditional on environment in New Zealand forests. However, introducing such an ecological process as competition will depend critically on the data model adopted. Choice of plot size relative to the scale of the process and availability of abundance data will determine the feasibility of such modelling.

Examples of using ecological and physiological knowledge to design the ecological, data and statistical models could be multiplied. The outcome of applying these ideas will be models that are more robust, include ecologically more rational responses and better prediction. Similar progress is being made with the data model and statistical model, but agreed standards for best current practice appear unlikely in the near future based on the review of literature presented here.

The three potential areas where progress might be made have been suggested, quantile regression, SEM and GWR. Quantile regression appears to have a sound basis in ecological and statistical theory, and software is available (Cade and Noon, 2003). A version is also available for nonparametric, nonlinear smoothers (Yu and Jones, 1998 quoted in Cade and Noon, 2003, see also Schroder et al., 2005). The only application to species modelling of quantile regression by Schroder et al. (2005) uses only a single predictor variable at a time due to the limited amount of data. A case study of its application to modelling of species spatial distribution using numerous environmental predictors is needed.

Structural equation modelling has an appealing conceptual framework with the possibility of testing whether data is consistent with a hypothesised causal pathway (Shipley, 1999). The limitation of SEM to linear relationships, multivariate normal data and preferably large data sets (Shipley, 2000) seems to restrict its potential for spatial prediction. Grace and Pugesek (1997) and Vile et al. (2006) apply data transformations to linearise individual relationships between variables. Arhonditsis et al. (2006) claim that non-linear relationships can be accommodated. It is beyond the competence of the present reviewer to review the statistical aspects of SEM, but it does appear to have problems dealing with curvilinear relationships and interactions (Schumacker and Marcoulides, 1998; Lee et al., 2004). The previous discussion on GWR (Fig. 5) makes it clear that the method cannot distinguish between non-stationarity of process and curvilinear relationships between predictor and dependent variables. This does not mean it cannot be used for exploratory data analysis; regions of rapid change of parameters still need to be investigated and understood. Spatial autocorrelation remains an important issue for spatial prediction of species distribution.

Improved communications between the existing research paradigms is needed. Certain "rules of thumb" can be applied:

1. Investigate the possibility of curvilinear relationships (*sensu lato*).

2. State explicitly what ecological theory is being assumed or tested.
3. Ensure that data resolution is consistent with theory and the predictors being used.
4. Examine relationships between variables for environmental process interactions in order to derive more proximal predictors.
5. Evaluate new methods with realistic artificial data, i.e. data consistent with current ecological understanding.
6. Use more than one statistical method to build the predictive model.
7. Do not depend solely on prediction success when evaluating species models.
8. Investigate the model residuals for spatial and other patterns.
9. Use independent data to test the models.

Many of these "rules" should be used routinely. Others might use terminology that is more sophisticated when discussing analytical technique but what is important is the compatibility between the ecological questions and the statistical methodology.

These regression models test whether there is a relationship between certain hypothesised environmental predictors and species occurrence. The models provide a predictive equation with which the data are consistent. The models do not imply causation. There is a reasonable presumption that the predictors are surrogates for causal processes based on ecological knowledge. However, disproving such presumptions poses difficulties. Few manipulative experimental tests of regression models have been attempted. McCune et al. (2003) modelled lichen distributions in relation to elevation using logistic regression and a kernel smoother. Subsequently, Antoine and McCune (2004) modelled the vertical height distribution of lichens within the forest canopy. They then compared these distribution curves with those obtained from the biomass growth of lichens derived from transplant experiments within the forest canopy. Two species showed consistent responses, abundance and growth followed similar curves. One species *Lobaria oregana* was inconsistent. Maximum abundance was at 25–30 m but maximum growth was at 40–45 m. Two hypotheses were advanced, competition or establishment. Such experimental tests will only be possible with certain organisms.

Process models are often advocated as an alternative approach to statistical models. Deductive testing of possible processes without adequate description of the pattern is likely to be inefficient. Statistical models of species occurrence offer a solution. They provide a statement of the potentially relevant environmental variables that need to be included in the process models and their relative importance. An iterative procedure between the process and statistical models is needed with each model acting as a test of the other. This approach is implicit in the recent studies of Leathwick (Leathwick and Whitehead, 2001; Leathwick and Austin, 2001; Leathwick, 2002) but needs to be made explicit. The hypothesised causal path model or SEM shown in Fig. 3 represents such an explicit statement. Both Grace (in McCune and Grace, 2002) and Shipley (1999, 2000) have also argued explicitly for this approach. Yet, the potential of SEM for testing models of species distribution will remain uncertain until it is demonstrated that the approach can deal with the numerous curvilinear responses that are commonly found in species/environment relationships (cf. Lee et al., 2004). Species distribution models are important in applied ecology but progress in the development and use will require constant evaluation.

This review is critical of our current progress in achieving an appropriate synthesis of our ecological theory, data models and statistical methods. However, there are many good ideas, methods and appropriate data sets in the different research paradigms operating in the area of species modelling which could contribute to a synthesis. Such a synthesis would improve our predictions and hence management of our natural resources.

Acknowledgements

I thank P. Gibbons, A.O. Nicholls, C.J. Krebs and K. Van Neil for comments on the manuscript and the CSIRO Sustainable Systems librarians for their help with references. This paper formed part of the Riederalp 2004 Workshop on Generalized Regression and Spatial Prediction and I thank the organisers A. Guisan, A. Lehmann, J. Overton, S. Ferrier and R. Aspinall for the invitation to attend.

REFERENCES*

Antoine, M.E., McCune, B., 2004. Contrasting fundamental and realized ecological niches with epiphytic lichen transplants in an old-growth *Pseudotsuga* forest. Bryologist 107, 163–173.

Araujo, M.B., Pearson, R.G., Thuiller, W., Erhard, M., 2005. Validation of species-climate impact models under climate change. Global Change Biol. 11, 1504–1513.

Arhonditsis, G.B., Stow, C.A., Steinberg, L.J., Kenney, M.A., Lathrop, R.C., McBride, S.J., Reckhow, K.H., 2006. Exploring ecological patterns with structural equation modelling and Bayesian analysis. Ecol. Model. 192, 385–409.

Austin, M.P., 1999a. A silent clash of paradigms: some inconsistencies in community ecology. Oikos 86, 170–178.

Austin, M.P., 1999b. The potential contribution of vegetation ecology to biodiversity research. Ecography 22, 465–484.

Austin, M.P., 2002a. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecol. Model. 157, 101–118.

Austin, M.P., 2002b. Case studies of the use of environmental gradients in vegetation and fauna modelling: theory and practice in Australia and New Zealand. In: Scott, J.M., Heglund, P.J., Samson, F., Haufler, J., Morrison, M., Raphael, M., Wall, B. (Eds.), Predicting Species Occurrences: Issues of Accuracy and Scale. Island Press, Covelo, California, pp. 73–82.

Austin, M.P., 2005. Vegetation and environment: discontinuities and continuities. In: van der Maarel, E. (Ed.), Vegetation Ecology. Blackwell Publishing, Oxford, pp. 52–84.

Austin, M.P., Cunningham, R.B., 1981. Observational analysis of environmental gradients. Proc. Ecol. Soc. Austr. 11, 109–119.

Austin, M.P., Nicholls, A.O., 1997. To fix or not to fix the species limits, that is the ecological question: response to Jari Oksanen. J. Veg. Sci. 8, 743–748.

* References used in review in section on current models for predicting species distributions.

Austin, M.P., Smith, T.M., 1989. A new model for the continuum concept. Vegetatio 83, 35–47.
Austin, M.P., Nicholls, A.O., Margules, C.R., 1990. Measurement of the realised qualitative niche: environmental niches of five *Eucalyptus* species. Ecol. Monogr. 60, 161–177.
Austin, M.P., Belbin, L. Meyers, J.A., Doherty, M.D., Luoto, M. (2006). Evaluation of statistical models for predicting plant species distributions: role of artificial data and theory. Ecol. Model., doi:10.1016/j.ecolmodel.2006.05.023, in press.
Austin, M.P., Nicholls, A.O., Doherty, M.D., Meyers, J.A., 1994. Determining species response functions to an environmental gradient by means of a β-function. J. Veg. Sci. 5, 215–228.
Austin, M.P., Meyers, J.A., Belbin, L., Doherty, M.D., 1995. Modelling of landscape patterns and processes using biological data. Subproject 5: simulated data case study. Consultancy Report for ERIN, CSIRO Wildlife and Ecology, Canberra.
*Bhattarai, K.R., Vetaas, O.R., Grytnes, J.A., 2004. Fern species richness along a central Himalayan elevational gradient, Nepal. J. Biogeogr. 31, 389–400.
Bio, A.M.F., 2000. Does vegetation suit our models? Data and model assumptions and the assessment of species distribution in space. Faculteit Ruimtelijke Wetenschappen Universiteit Utrecht. Nederlandse Geografische Studies 265.
Bio, A., Alkemade, R., Barendregt, A., 1998. Determining alternative models for vegetation response analysis: a non-parametric approach. J. Veg. Sci. 9, 5–16.
Blackburn, T.M., Lawton, J.H., Perry, J.N., 1992. A method of estimating the slope of upper bounds of plots of body size and abundance in natural animal assemblages. Oikos 65, 107–112.
Bloom, A.J., Chapin, F.S., Mooney, H.A., 1985. Resource limitationin plants-an economic analogy. Ann. Rev. Ecol. Syst. 16, 363–392.
Boyce, M.S., Vernier, P.R., Nielson, S.E., Schmiegelow, F.K.A., 2002. Evaluating resource selection functions. Ecol. Model. 157, 281–300.
Brotons, L., Thuiller, W., Araujo, M.B., Hirzel, A.H., 2004. Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. Ecography 27, 437–448.
*Bustamante, J., Seoane, J., 2004. Predicting the distribution of four species of raptors (Aves: Accipitridae) in southern Spain: statistical models work better than existing maps. J. Biogeogr. 31, 295–306.
Brzeziecki, B., 1987. Analysis of vegetation–environment relationships using a simultaneous equations model. Vegetatio 71, 175–184.
Cade, B.S., Guo, Q., 2000. Estimating effects of constraints on plant performance with regression quantiles. Oikos 91, 245–254.
Cade, B.S., Noon, B.R., 2003. A gentle introduction to quantile regression for ecologists. Front. Ecol. Environ. 1, 412–420.
Cade, B.S., Noon, B.R., Flather, C.H., 2005. Quantile regression reveals hidden bias and uncertainty in habitat models. Ecology 86, 786–800.
Cade, B.S., Terrell, J.W., Schroeder, R.L., 1999. Estimating effects of limiting factors with regression quantiles. Ecology 80, 311–323.
Cawsey, E.M., Austin, M.P., Baker, B.L., 2002. Regional vegetation mapping in Australia: a case study in the practical use of statistical modelling. Biodivers. Conserv. 11, 2239–2274.
*Clarke, E.D., Spear, L.B., McCracken, M.L., Marques, F.F.C., Brochers, D.L., Buckland, S.T., Ainley, D.G., 2003. Validating the use of generalized additive models and at-sea surveys to estimate size and temporal trends of seabird populations. J. Appl. Ecol. 40, 278–292.
Cleveland, W.S., Devlin, S.J., 1988. Locally weighted regression: an approach to regression analysis by local fitting. J. Am. Stat. Assoc. 83, 596–610.
Coudun, C., Gegout, J., 2005. Ecological behaviour of herbaceous forest species along a pH gradient: a comparison between oceanic and semicontinental regions in northern France. Global Ecol. Biogeogr. 14, 263–270.
Cressie, N.A.C., 1993. Statistics for Spatial Data. Wiley, New York.
Drake, J.M., Randin, C., Guisan, A., 2006. Modelling ecological niches with support vector machines. J. Appl. Ecol. 43, 424–432.
Dubayah, R., Rich, P.M., 1995. Topographic solar radiation models for GIS. Int. J. Geogr. Inf. Syst. 9, 405–419.
Elith, E., Burgman, M.A., Regan, H.M., 2002. Mapping epistemic uncertainties and vague concepts in predictions of species distribution. Ecol. Model. 157, 313–329.
Elith, J., Ferrier, S., Huettmann, F., Leathwick, J., 2005. The evaluation strip: a new and robust method for plotting predicted responses from species distribution models. Ecol. Model. 186, 280–289.
Elith, J., et al., 2006. Novel methods improve prediction of species distributions from occurrence data. Ecography 29, 129–151.
*Engler, R., Guisan, A., Rechsteiner, L., 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. J. Appl. Ecol. 41, 263–274.
Ejrnaes, R., 2000. Can we trust gradients extracted by detrended correspondence analysis? J. Veg. Sci. 11, 565–572.
Ferrier, S., Watson, G., Pearce, J., Drielsma, M., 2002. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. 1. Species-level modelling. Biodivers. Conserv. 11, 2275–2307.
Fitzgerald, R.W., Lees, B.G., 1992. The application of neural networks to the floristic classification of remote sensing and GIS data in complex terrain. In: Proceedings of the XVII Congress of the International Society for Photogrammetry and Remote Sensing, Washington, USA, pp. 570–573.
Foody, G.M., 2003. Geographical weighting as a further refinement to regression modelling: An example focused on the NDVI-rainfall relationship. Remote Sens. Environ. 88, 283–293.
Foody, G.M., 2004. Spatial nonstationarity and scale-dependency in the relationship between species richness and environmental determinants for the sub-Saharan endemic avifauna. Global Ecol. Biogeogr. 13, 315–320.
Foody, G.M., 2005a. Clarifications on local and global data analysis. Global Ecol. Biogeogr. 14, 99–100.
Foody, G.M., 2005b. Mapping the richness and composition of British breeding birds from coarse spatial resolution satellite sensor imagery. Int. J. Remote Sens. 26, 3943–3956.
Fotheringham, A.S., Brunsdon, C., Charlton, M., 2002. Geographical Weighted Regression: The Analysis of Spatially Relationships. Wiley, Chichester.
Franklin, J., 1995. Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. Prog. Phys. Geogr. 19, 474–499.
Friedman, J.H., 1991. Multivariate adaptive regression splines. Ann. Stat. 19, 1–141 (with discussion).
Friedman, J.H., Hastie, T., Tibshirani, R., 2000. Additve logistic regression: a statistical view of boosting. Ann. Stat. 28, 337–407.
Garrido, J.L., Rey, P.J., Herrera, C.M., 2005. Pre- and post-germination determinants of spatial variation in recruitment in the perennial herb *Helleborus foetidus* L. (Ranunculaceae). J. Ecol. 93, 60–66.
Gegout, J., Coudon, C., Bailly, G., Jabiol, B., 2005. EcoPlant: A forest site database linking floristic data with soil and climate variables. J. Veg. Sci. 16, 257–260.
*Gibson, L.A., Wilson, B.A., Cahill, D.M., Hill, J., 2004. Spatial prediction of rufous bristlebird habitat in a coastal heathland: a GIS-based approach. J. Appl. Ecol. 41, 213–223.
Giller, J., 1984. Community Structure and the Niche. Chapman and Hall, London.
Grace, J.B., Pugesek, B.H., 1997. A structural equation model of plant species richness and its application to a coastal wetland. Am. Nat. 149, 436–460.

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C., Peterson, A.T., 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends Ecol. Evol. 19, 497–503.

Guisan, A., Harrell, F.E., 2000. Ordinal response regression models in ecology. J. Veg. Sci. 11, 617–626.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135, 147–186.

Guisan, A., Edwards, T.C., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol. Model. 157, 89–100.

Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models? Ecol. Lett. 8, 993–1009.

Harrington, R., Woiwod, I., Sparks, T., 1999. Climate change and trophic interactions. Trends Ecol. Evol. 14, 146–150.

Hastie, T., Tibshirani, R., 1990. Generalised Additive Models. Chapman and Hall, London.

Hirzel, A., Helfer, V., Metral, F., 2001. Assessing habitat-suitability models with a virtual species. Ecol. Model. 145, 111–121.

Hirzel, A.H., Hausser, J., Chessel, D., Perrin, N., 2002. Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? Ecology 83, 2027–2036.

Huisman, J., Olff, H., Fresco, L.F.M., 1993. A hierarchical set of models for species response analysis. J. Veg. Sci. 4, 37–46.

Huntley, B., Berry, P.M., Cramer, W.P., McDonald, A.P., 1995. Modelling present and potential future ranges of some European higher plants using climate response surfaces. J. Biogeogr. 22, 967–1001.

*Huntley, B., Green, R.E., Collingham, Y.C., Hill, J.K., Willis, S.G., Bartlein, P.J., Cramer, W., Hagemeijer, W.J.M., Thomas, C.J., 2004. The performance of models relating species geographical distributions to climate is independent of trophic level. Ecol. Lett. 7, 417–426.

Huston, M.A., 1994. Biological Diversity: The Coexistence of Species on Changing Landscapes. Cambridge University Press, Cambridge.

Huston, M.A., 2002. Introductory essay: critical issues for improving predictions. In: Scott, J.M., Heglund, P.J., Samson, F., Haufler, J., Morrison, M., Raphael, M., Wall, B. (Eds.), Predicting Species Occurrences: Issues of Accuracy and Scale. Island Press, Covelo, California, pp. 7–21.

Iriondo, J.M., Albert, M.J., Escudero, A., 2003. Structural equation modelling: an alternative for assessing causal relationships in threatened plant populations. Biol. Conserv. 113, 367–377.

*Jeganathan, P., Green, R.E., Norris, K., Vogiatzakis, I.N., Bartsch, A., Wotton, S.R., Bowden, C.G.R., Griffiths, G.H., Pain, D., Rahmani, A.R., 2004. Modelling habitat selection and distribution of the critically endangered Jerdon's courser *Rhinoptilus bitorquatus* in scrub jungle: an application of a new tracking method. J. Appl. Ecol. 41, 224–237.

Jetz, W., Rahbek, C., 2002. Geographic range size and determinants of avian species richness. Science 297, 1548–1551.

Jetz, W., Rahbek, C., Lichstein, J.W., 2005. Local and global approaches to spatial data analysis in ecology. Global Ecol. Biogeogr. 14, 97–98.

Johnson, M.L., Huggins, D.G., DeNoyelles, F., 1991. Ecosystem modelling with LISREL: a new approach for measuring direct and indirect effects. Ecol. Appl. 1, 383–398.

Kadmon, R., Farber, O., Danin, A., 2003. A systematic analysis of factors affecting the performance of climatic envelope models. Ecol. Appl. 13, 853–867.

Kadmon, R., Farber, O., Danin, A., 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. Ecol. Appl. 14, 401–413.

Kaiser, M.S., Speckman, P.L., Jones, J.R., 1994. Statistical models for limiting nutrient relations in inland waters. J. Am. Stat. Assoc. 89, 410–423.

Knight, C.A., Ackerly, D.D., 2002. Variation in nuclear DNA content across environmental gradients: a quantile regression analysis. Ecol. Lett. 5, 66–76.

Krause-Jensen, D., Middelboe, A.L., Sand-Jensen, K., Christensen, P.B., 2000. Eelgrass, *Zostera marina*, growth along depth gradients; upper boundaries of the variation as a powerful predictive tool. Oikos 91, 233–244.

Krebs, C.J., 2001. Ecology; The Experimental Analysis of Distribution and Abundance, fifth ed. Benjamin Cummings, San Francisco.

Kuhn, T.S., 1970. The Structure of Scientific Revolutions, second ed. The University of Chicago Press, Chicago.

Leathwick, J.R., 1995. Climatic relationships of some New Zealand forest tree species. J. Veg. Sci. 6, 237–248.

Leathwick, J.R., 1998. Are New Zealand's *Nothofagus* species in equilibrium with their environment? J. Veg. Sci. 9, 719–732.

Leathwick, J.R., 2002. Intra-generic competition among Nothofagus in New Zealand's primary indigenous forests. Biodivers. Conserv. 11, 2177–2187.

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

Leathwick, J.R., Whitehead, D., 2001. Soil and atmospheric water deficits and the distributions of New Zealand's indigenous tree species. Funct. Ecol. 15, 233–242.

Leathwick, J.R., Elith, J., Hastie, T. Comparative performance of two techniques for statistical modelling of presence–absence data. Ecology, in press.

Leathwick, J.R., Whitehead, D., McLeod, M., 1996. Predicting changes in the composition of New Zealand's indigenous forests in response to global warming: a modelling approach. Environ. Software 11, 81–90.

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

Leathwick, J.R., Elith, J., Francis, M.P., Hastie, T., Taylor, P. Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. Marine Ecol. Prog. Ser., in press.

Lee, S.Y., Song, X.Y., Poon, W.Y., 2004. Comparison of approaches in estimating interactions and quadratic effects of latent variables. Multivariate Behav. Res. 39, 37–67.

Lehmann, A., Overton, J.McC., Austin, M.P., 2002a. Regression models for spatial prediction: their role for biodiversity and conservation. Biodivers. Conserv. 11, 2085–2092.

Lehmann, A., Overton, J.McC., Leathwick, J.R., 2002b. GRASP: generalized regression analysis and spatial prediction. Ecol. Model. 157, 189–207.

Maggini, R., Lehmann, A., Zimmermann, N.E., Guisan, A., 2006. Improving generalized regression analysis for spatial predictions of forest communities. J. Biogeogr., in press.

*Malo, J.E., Suarez, F., Diez, A., 2004. Can we mitigate animal–vehicle accidents using predictive models? J. Appl. Ecol. 41, 701–710.

Manel, S., Dias, J.M., Ormerod, S.J., 1999. Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with Himalayan river bird. Ecol. Model. 120, 337–347.

Marquez, A.L., Real, R., Vargas, J.M., 2004. Dependence of broad-scale geographical variation in fleshy-fruited plant species richness on disperser bird species richness. Global Ecol. Biogeogr. 13, 295–304.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second ed. Chapman and Hall, London.

McCune, B., Grace, J.B., 2002. Analysis of Ecological Communities. MjM Software Design, Oregon, USA.

McCune, B., Berryman, S.D., Cissel, J.H., Gitelman, A.I., 2003. Use of a smoother to forecast occurrence of epiphytic lichens under alternative forest management plans. Ecol. Appl. 13, 1110–1123.

*McPherson, J.M., Jetz, W., Rogers, D.J., 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? J. Appl. Ecol. 41, 811–823.

Miller, J., Franklin, J., 2002. Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. Ecol. Model. 157, 227–247.

Mitchell, R.J., 1992. Testing evolutionary and ecological hypotheses using path analysis and structural equation modelling. Funct. Ecol. 6, 123–129.

Mitchell, R.J., 1994. Effects of floral traits, pollinator visitation, and plant size on *Ipomopsis aggregrata* fruit production. Am. Nat. 143, 870–889.

Moisen, G.G., Frescino, T.S., 2002. Comparing five modelling techniques for predicting forest characteristics. Ecol. Model. 157, 209–225.

Moisen, G.G., Freeman, E.A., Blackard, J.A., Zimmermann, N.E., Edwards Jr., T.C., 2006. Predicting tree species presence and basal area in Utah—a comparison of generalized additive models, stochastic gradient boosting, and tree-based methods. Ecol. Model., in press.

Munoz, J., Felicisimo, A.M., 2004. Comparison of statistical methods commonly used in predictive modelling. J. Veg. Sci. 15, 285–292.

Nicholls, A.O., 1989. How to make biological surveys go further with generalized linear models. Biol. Conserv. 50, 51–76.

Nicholls, A.O., 1991. Examples of the use of generalized linear models in analysis of survey data for conservation evaluation. In: Margules, C.R., Austin, M.P. (Eds.), Nature Conservation: Cost Effective Biological Surveys and Data Analysis. CSIRO, Melbourne, pp. 191–201.

Osborne, P.E., Suarez-Seoane, S., 2002. Should data be partitioned spatially before building large-scale distribution models? Ecol. Model. 157, 249–259.

Pausas, J.G., Braithwaite, L.W., Austin, M.P., 1995. Modelling habitat quality for arboreal marsupials in the South coastal forests of New South Wales, Australia. For. Ecol. Manag. 78, 39–49.

Pausas, J.G., Austin, M.P., Noble, I.R., 1997. A forest simulation model for predicting eucalypt dynamics and habitat quality for arboreal marsupials. Ecol. Appl. 7, 921–933.

Pearce, J., Ferrier, S., 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecol. Modell. 133, 225–245.

Pearson, R.G., Dawson, T.P., 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Global Ecol. Biogeogr. 12, 361–371.

Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modelling of species geographic distributions. Ecol. Model. 190, 231–259.

Prentice, I.C., Cramer, W., Harrison, S.P., Leemans, R., Monserud, R.A., Solomon, A.M., 1992. A global biome model based on plant physiology and dominance, soil properties and climate. J. Biogeogr. 19, 117–134.

Ricklefs, R.E., 2004. A comprehensive framework for global patterns in biodiversity. Ecol. Lett. 7, 1–15.

Rubio, G., Zhu, J., Lynch, J.P., 2003. A critical test of the two prevailing theories of plant resonse to nutrient availability. Am. J. Bot. 90, 143–152.

*Rushton, S.P., Ormerod, S.J., Kerby, G., 2004. New paradigms for modelling species distributions? J. Appl. Ecol. 41, 193–200.

Rydgren, K., Okland, R.H., Okland, T., 2003. Species response curves along environmental gradients. A case study from SE Norwegian swamp forests. J. Veg. Sci. 14, 869–880.

Scharf, F.S., Juanes, F., Sutherland, M., 1998. Inferring ecological relationships from the edges of scatter diagrams: comparison of regression techniques. Ecology 79, 448–460.

Schroder, H.K., Andersen, H.E., Kiehl, K., 2005. Rejecting the mean: estimating the response of fen plant species to environmental factors by non-linear quantile regression. J. Veg. Sci. 16, 373–382.

Schumacker, R.E., Marcoulides, G.A., 1998. Interaction and Nonlinear Effects in Structural Equation Modelling. Lawrence Erlbaum Associates, New Jersey.

Scott, J.M., Heglund, P.J., Haufler, J., Morrison, M., Raphael, M., Wall, B., Samson, F., 2002. Predicting Species Occurrences: Issues of Accuracy and Scale. Island Press, Covelo, California.

*Segurado, P., Araujo, M.B., 2004. An evaluation of methods for modelling species distributions. J. Biogeogr. 31, 1555–1568.

Shipley, B., 1999. Testing causal explanations in organismal biology: causation, correlation and structural equation modelling. Oikos 86, 374–382.

Shipley, B., 2000. Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference. Cambridge University Press, Cambridge.

Shipley, B., Lechowicz, M.J., 2000. The functional co-ordination of leaf morphology, nitrogen concentration, and gas exchange in 40 wetland species. Ecoscience 7, 183–194.

Smith, P.A., 1994. Autocorrelation in logistic regression modelling of species' distributions. Global Ecol. Biogeogr. Lett. 4, 47–61.

Stockwell, D.R.B., Noble, I.R., 1992. Induction of sets of rules from animal distribution data: a robust and informative method of data analysis. Math. Comput. Simul. 33, 385–389.

Termansen, M., McClean, C.J., Preston, C.D., 2006. The use of genetic algorithms and Baysian classification to model species distributions. Ecol. Model. 192, 410–424.

Thomson, J.D., Weiblen, G., Thomson, B.A., Alfaro, S., Legendre, P., 1996. Untangling multiple factors in spatial distributions: lilies, gophers and rocks. Ecology 77, 1698–1715.

*Thuiller, W., 2003. BIOMOD-optimising predictions of species distributions and projecting potential future shifts under global change. Global Change Biol. 9, 1353–1362.

*Thuiller, W., Vayreda, J., Pino, J., Sabate, S., Lavorel, S., Garcia, C., 2003a. Large-scale environmental correlates of forest tree distributions in Catalonia (NE Spain). Global Ecol. Biogeogr. 12, 313–325.

Thuiller, W., Araujo, M.B., Lavorel, S., 2003b. Generalized models vs. classification tree analysis: predicting spatial distributions of plant species at different scales. J. Veg. Sci. 14, 669–680.

*Thuiller, W., Araujo, M.B., Lavorel, S., 2004. Do we need land-cover data to model species distributions in Europe? J. Biogeogr. 31, 353–361.

Van der Ploeg, R.R., Bohm, W., Kirkham, M.B., 1999. On the origin of the theory of mineral nutrition of plants and the law of the minimum. Soil Sci. Soc. Am. J. 63, 1055–1062.

Van Neil, K.P., Laffan, S.W., Lees, B.G., 2004. Effect of error in the DEM on environmental variables for predictive vegetation modelling. J. Veg. Sci. 15, 747–756.

Van Neil, K.P., Austin, M.P. Predictive vegetation modelling for conservation: impact of error propagation from digital elevation data. Ecol. Appl., in press.

*Venier, L.A., Pearce, J., McKee, J.E., McKenny, D.W., Niemi, G.J., 2004. Climate and satellite-derived land cover for predicting breeding bird distribution in the Great Lakes Basin. J. Biogeogr. 31, 315–331.

Vile, D., Shipley, B., Garnier, E., 2006. A structural equation model to integrate changes in functional strategies during old-field succession. Ecology 87, 504–517.
Wamelink, G.W.W., Goedhart, P.W., Van Dobben, H.F., Berendse, F., 2005. Plant species as predictors of soil pH: replacing expert judgement with measurements. J. Veg. Sci. 16, 461–470.
Weiher, E., 2003. Species richness along multiple gradients: testing a general multivariate model in oak savannas. Oikos 101, 311–316.
Weiher, E., Forbes, S., Schauwecker, T., Grace, J.B., 2004. Multivariate control of plant species richness and community biomass in blackland prairie. Oikos 106, 151–157.
Whittaker, R.J., Willis, K.J., Field, R., 2001. Scale and species richness: towards a general, hierarchical theory of species diversity. J. Biogeogr. 28, 453–470.
Wood, S.N., Augustin, N.H., 2002. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. Ecol. Model. 157, 157–178.
Yee, T.W., Mackenzie, M., 2002. Vector generalized additive models in plant ecology. Ecol. Model. 157, 141–156.
Yee, T.W., Mitchell, N.D., 1991. Generalized additive models in plant ecology. J. Veg. Sci. 2, 587–602.
Yu, K., Jones, M.C., 1998. Local linear quantile regression. J. Am. Stat. Assoc. 93, 228–237.
Zaneiwski, A.E., Lehmann, A., Overton, J., 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. Ecol. Model. 157, 261–280.

Further reading

*Amar, A., Arroyo, B., Redpath, S., Thirgoods, S., 2004. Habitat predicts losses of red grouse to individual hen harriers. J. Appl. Ecol. 41, 305–314.
*Cabeza, M., Araujo, M., Wilson, R.J., Thomas, C.D., Cowley, M.J.R., Moilanen, A., 2004. Combining probabilities of occurrence with spatial reserve design. J. Appl. Ecol. 41, 252–262.
*Frair, J.L., Nielsen, S.E., Merrill, E.H., Lele, S.R., Boyce, M.S., Munro, R.H.M., Stenhouse, G.B., Beyer, H.L., 2004. Removing GPS collar bias in habitat selection studies. J. Appl. Ecol. 41, 201–212.
*Heikken, R.K., Luoto, M., Virkkala, R., Rainio, K., 2004. Effects of habitat cover, landscape structure and spatial variables on the abundance of birds in an agricultural-forest mosaic. J. Appl. Ecol. 41, 824–835.
*Johnson, C.J., Seip, D.R., Boyce, M.S., 2004. A quantitative approach to conservation planning: using resource selection functions to map the distribution of mountain caribou at multiple spatial scales. J. Appl. Ecol. 41, 238–251.

* References used in review in section on current models for predicting species distributions.