Journal of Applied Ecology 2004
41, 193–200

SPECIAL PROFILE

New paradigms for modelling species distributions?

S. P. RUSHTON*, S. J. ORMEROD† and G. KERBY

**Centre for Land Use and Water Resources, University of Newcastle upon Tyne, St Thomas Street, Newcastle upon Tyne, UK; and †School of Biosciences, Cardiff University, Cardiff, UK*

Summary

1. The management of both desirable and undesirable species requires an understanding of the factors determining their distribution. Quantitative distribution models offer simple methods for formulating the species–habitat link and the means not only for predicting where species should occur, but also for understanding the factors involved. Generalized linear modelling, in particular, links the incidence of species to habitat variables, and has increasingly formed the backbone of the modelling approaches used. New 'data technologies', such as remote sensing and geographical information systems, have further broadened these modelling applications to almost any ecological system and any species for which there are distribution data.

2. Many previous approaches have aimed to identify the most parsimonious model with the best suite of predictors, selected on the basis of null hypothesis testing. However, information-theoretic approaches based on Akaike's information criterion allow the selection of a best approximating model or a subset of models from a set of candidates. Information-theoretic approaches require a deeper understanding of the biology of the system modelled and may well become an improved paradigm for species distribution modelling.

3. *Synthesis and applications*. This special profile of six papers demonstrates the development in methodology used in species distribution modelling. The papers show how information-theoretic approaches can be coupled with emerging data technologies to address issues of conservation significance. With conservation biology and applied ecology at the forefront of many of the basic science developments so far, we expect these methods to pervade other areas of ecological research more fully in future.

Key-words: conservation biology, information-theoretic approaches, models

Journal of Applied Ecology (2004) **41**, 193–200

Introduction

Quantifying the relationships between the distributions of species and their abiotic or biotic environments has a long history in ecological research. While understanding where species occur is a fundamental ecological requirement, prediction of occurrence is essential for much conservation and population management. This is particularly the case for endangered species, where knowing what determines distribution is a necessary precursor for schemes to mitigate decline or to create new populations through reintroduction. Given our focus on reporting research with an explicitly applied significance, it is perhaps no surprise that a sizeable proportion of manuscripts received by the *Journal of Applied Ecology* has been concerned with modelling species–habitat relationships and distributions. Early in the 1990s, *Journal of Applied Ecology* published one of the key, seminal papers that introduced significant new methodology as well as novel applications to real conservation problems (Augustine, Mugglestone & Buckland 1993). Since 2000, we have published at least 14 papers that consider species distribution modelling (Table 1), and many others have addressed methodological issues associated with data handling or data collection for modelling . While there has been an ongoing interest in the development and application of modelling approaches, the last 3 years have seen fundamental changes in the methodology involved with this form of modelling, with the development of information-theoretic approaches. These developments have been apparent in published papers, and the six papers in this special profile focus on some of the key issues. Not only do they illustrate how widely new data technologies, such as remotely sensed imagery, have pervaded applied ecology but also they reveal how the philosophy of modelling is changing. Five of the

Correspondence: Steven.Rushton@newcastle.ac.uk

Table 1. Methodologies and approaches used in modelling species distribution in articles published in the *Journal of Applied Ecology* since 2000

Authors	Taxon	Species data	Habitat predictors	GIS	Model	Model assessment
Collingham *et al.* (2000)	Weeds	Biological recording	Remote-sensed imagery	Yes	Stepwise logistic regression	Kappa statistics
Cowley *et al.* (2000)	Lepidoptera	Transect survey	Mapped habitat data	No	Logistic regression	Kappa statistics
Milsom *et al.* (2000)	Birds	Field survey	Mapped habitat data and field survey	No	Generalized linear mixed model (logistic and autologistic)	–
Manel, Buckton & Ormerod (2000)	Birds Invertebrates	Field survey	Field survey	No	Logistic regression using AIC	–
Bradbury *et al.* (2000)	Birds	Field survey	Field survey	No	Log-linear and logistic regression	Threshold classification
Gates & Donald (2000)	Birds	Biological recording	Mapped habitat data	No	Logistic regression	–
Jaberg & Guisan (2001)	Bats	Augmented biological records	Mapped habitat data	Yes	Poisson regression	Kappa statistics
Manel, Williams & Ormerod (2001)	Invertebrates	Field survey	Field survey	No	Logistic regression using AIC	Kappa statistics, ROC plots, jack-knifing
Pearce *et al.* (2001)	Mammals, reptiles, birds	Field survey	Remote-sensed imagery	Yes	Generalized additive model (logistic)	Modified *Z*-test
Osborne, Alonso & Bryant (2001)	Birds	Field survey	Remote-sensed imagery	Yes	Logistic and autologistic regression	ROC plots
Suárez-Seoane, Osborne & Alonso (2002)	Birds	Field survey	Remote-sensed imagery	Yes	Generalized additive model	Threshold classification
Ambrosini *et al.* (2002)	Birds	Field survey	Field survey	No	Logistic and linear regression with quasi-like lihood	Kappa statistics
Schadt *et al.* (2002)	Mammals	Radio-tracking	Remote-sensed imagery	Yes	Logistic regression	ROC plots
Holloway, Griffiths & Richardson (2003)	Lepidoptera	Field survey	Mapped habitat data	Yes	Rule base	–
Vaughan *et al.* (2003)	Mammals	Questionnaire	Land classes and farm census	No	Ordinal logistic regression	Concordance between predicted and observed
Cabeza *et al.* (2004)	Lepidoptera	Transect survey	Mapped habitat data	No	Logistic regression	Kappa statistics
Engler, Guisan & Rechsteiner (2004)	Plants	Biological recording	Mapped climatic and terrain data	Yes	Logistic regression	Kappa and ROC plots
Jeganathan *et al.* (2004)	Birds	Bird sign (field survey)	Remote-sensed imagery	Yes	Logistic and autologistic regression	–
Gibson *et al.* (2004)	Birds	Bird song (field survey)	Remote-sensed imagery	Yes	Logistic regression and information-theoretic	ROC plots
Johnson, Seip & Boyce (2004)	Mammals	Radio-tracking	Mapped habitat data	Yes	Logistic regression and information-theoretic	*K*-fold cross-validation
Frair *et al.* (2004)	–	Remote sensing (GPS)	Field survey	Yes	Logistic regression and information-theoretic	ROC plots

papers utilize information-theoretic approaches that are different to the formal hypothesis-testing approach presented in many past papers. We consider here the significance of these recent changes, how *Journal of Applied Ecology* has responded and how they may shape applied ecology in the future.

All modelling studies have three basic components: a data set describing the incidence or abundance of the species of interest and a data set of putative explanatory variables; a mathematical model that relates the species data to the explanatory variables; and an assessment of the utility of the model developed in terms of a validation exercise or an assessment of model robustness. Recent publications in *Journal of Applied Ecology* have all followed this basic formula, but each has placed different emphasis on the three components. While these differences reflect the interests of individual authors, they also reflect author responses to the underlying ecology of the species that have been studied and an implicit recognition that assumptions have to be made at all stages in the modelling process.

The species and habitat data sets

One of the major issues in species habitat and distribution modelling is getting data that are of the correct 'scope' in both time and space (Vaughan & Ormerod 2003). In an ideal world, the target species would be sedentary at a fixed point in space, and its ecological requirements well known and measurable at the same spatio-temporal scales. In reality, measurement of the known potential predictor variables may be difficult and species may have ecological requirements that are unknown or are immeasurable. These issues have very severe implications for the likely success of the modelling effort, irrespective of the approaches used. In many respects, collecting the species (response) data set is more difficult than the associated habitat variables, simply because the target species may move around the landscape. Identifying where the individuals are and what they are using as resources in the landscape is problematic. Consideration of the recent publications in this and other issues of *Journal of Applied Ecology* shows that these data sets fall into two basic types: those that have been collected as part of a survey designed to provide information on target species with the main aim of modelling species–habitat relationships; and those that have been generated as a result of other exercises not specifically associated with distribution modelling. In some groups, such as birds and butterflies (Bradbury *et al*. 2000; Cowley *et al*. 2000; Milsom *et al*. 2000; Cabeza *et al*. 2004), there is often a standard sampling methodology where a transect walked through the landscape allows collation of contacts with the species. Some methodologies, like the Common Bird Census (CBC) used in the study of yellowhammers *Emberiza citrinella* L. by Bradbury *et al*. (2000), have a long history and are widely used to monitor changes in bird populations. Where species are very rare, standard sampling methodologies are often not available and data can be difficult to collect simply because there are insufficient individuals to sample or eventually to model. In these cases, indicators of species presence rather than true species encounters have been used as surrogate response variables. Two papers in the special profile address these issues: Gibson *et al*. (2004) located bristlebirds *Dasyornis broadbenti caryochrous* (McCoy 1867) in south-western Victoria, Australia, on the basis of calls, while Jeganathan *et al*. (2004) used the presence of tracks in sand to estimate the incidence of the critically endangered Jerdon's courser *Rhinoptilus bitorquatus* (Blyth) in Andra Pradesh, India. Survey work where species are rare can also be very expensive, and this has provided a strong financial incentive for analysing data derived from casual and non-systematic surveys. Vaughan *et al*. (2003) circumvented the cost issues of sampling by using a questionnaire distributed to land owners to provide distribution data for their modelling. The main problem with this approach (as they acknowledge) is that biased sampling may result if the response rate is not the same for the different categories of interest on the questionnaire. However, they were able to demonstrate that non-respondents were similar to respondents (Vaughan *et al*. 2003). The use of questionnaires in collecting data of this type has considerable potential and warrants further study.

In countries like the UK there is a large network of amateur biologists and naturalists who collect species records, and for many groups there are formal reporting and data storage schemes. The Biological Records Centre (BRC, Centre for Ecology and Hydrology, Monks Wood, Abbots Ripton, Huntingdon), established in 1964, is the national focus for species recording in the UK and the database contains nearly 12 million records of more than 12 000 species. These data are available for scientific study but the main problem with using such records for species distribution modelling is that they were not originally collected for the purpose of modelling. These data are usually collated over long time periods, often using a variety of methods, often in surveys that are not systematic. Zero records, where recorders search but find nothing, are particularly poorly recorded. In some of the recent publications in *Journal of Applied Ecology*, authors have augmented recording data by sampling to allow modelling. This undoubtedly increases the utility of survey data for distribution modelling. Collingham *et al*. (2000), Jaberg & Guisan (2001) and Osborne, Alonso & Bryant (2001) all used data sets enhanced by more formal sampling, to the extent that where they had no records for species in their study areas they were confident the species were not there. Other studies have circumvented the zero record problem by generating zero records at random within the landscape. The zero records then form the absences to compare with the presence data in the formal model. Schadt *et al*. (2002) used this approach to compare habitat use by lynx *Lynx lynx* with random non-home ranges sampled from the landscape. Engler,

Guisan & Rechsteiner (2004) take this one stage further and use Ecological Niche Factor Analysis (ENFA) to weight the selection of pseudo-absence in favour of areas of the landscape predicted to have a low likelihood of encountering the species. Whilst this form of data manipulation has an obvious appeal, the implications of introducing bias into the data set need careful consideration.

A perennial problem in distribution modelling is in identifying the appropriate scale at which to sample. In the case of non-systematically collated survey data, the sampling unit is usually some form of grid cell, the size of which is not normally related to any ecological feature of significance to the species concerned. In the UK, species distribution data are collated at a range of spatial scales, typically from 100 m to 10 000 m. Collingham *et al*. (2000) addressed this issue in a study modelling the distribution of weed species at two sampling scales, with predictions made at regional and national levels. They concluded that there was reasonable correspondence in the key predictors at both scales but that scaling down from coarse to fine resolutions did not lead to models that gave a good fit to the observed data. This highlights the difficulties associated with using recorder data rather than random sampling, which is (usually) designed to sample at scales appropriate to the ecological processes determining the species distribution.

Where animals are highly mobile, as with foraging predators, observers have changed the emphasis of sampling from an area-based focus, where the presence or abundance of animals in sample units is recorded per unit area, to a focus on recording exactly where individual animals are in the landscape. The animal then forms the sample unit for collating the key predictor variables. Where the target species has a key resource that is easily identifiable, this can also form a basis for sampling strategy. Ambrosini *et al*. (2002) based their sampling strategy for swallow *Hirundo rustica* L. populations on the presence of farm buildings that provided nest sites. For large mammals, and predators in particular, sampling is often based on radio-tracking methodology, as in the studies of lynx in central Europe by Schadt *et al*. (2002) and mountain caribou *Rangifer tarandus caribou* (Gmelin) in Canada by Johnson, Seip & Boyce (2004). Tracking animals individually is expensive in terms of labour and equipment: Schadt *et al*. (2002) based their research on 3402 radio-locations collected over 3 years, while Johnson, Seip & Boyce (2004) tracked caribou from a fixed-winged aircraft. It is not surprising, therefore, that there has been considerable interest in the use of satellite technology, with target animals carrying global positioning system (GPS) collars that store information on their geospatial position as they move around the landscape. While these techniques reduce the expense associated with tracking animals, they are not without their own constraints. Often there are errors in spatial positioning, and the tracking system has missing data because of habitat-induced interference with the signal between satellite and the GPS collar. Frair *et al*. (2004), in this issue, have shown that these errors are predictable and can be corrected under different sampling designs. This is a valuable contribution on a component of error that has too often been ignored in previous research based on radio-telemetry.

The collection of habitat data for species distribution modelling underwent dramatic change in the 1990s when remote-sensed imagery derived from satellites became widely available. This, coupled with the increased use of geographical information systems (GIS) to store and manipulate spatial data, led to an expansion in species distribution modelling. As with tracking individual animals, this was largely because the costs associated with data collection were minimized. These technologies have revolutionized conservation biology and the impacts have been far reaching. Ecologists are no longer restricted to small study areas: Johnson, Seip & Boyce (2004) used a sample region of some 38 800 km^2, an area that would have been impossible to survey on the ground. Remote sensing has also meant that research has become more feasible in remote areas and in habitats in which it is technically difficult to collect large amounts of habitat information (Jeganathan *et al*. 2004). While satellite imagery is available for most of the land surface of the planet, the utility of such data for modelling is determined to a large extent by the ecology of the target species. Typically, satellite-derived data have a fixed resolution (this may depend on the wavelength of radiation sampled) and, as with much biological recording data, they are not collected specifically for modelling species distributions. In many cases these data can only be used as surrogates for habitat predictors. Classifications of data derived from satellite imagery, such as the CEH Land Cover map (Collingham *et al*. 2000) and CORINE land classification (Schadt *et al*. 2002), have been used to provide predictor variables, while other authors have used ordinal variables to represent features such as vegetation density (Osborne, Alonso & Bryant 2001). Jeganathan *et al*. (2004) offer a particularly valuable contribution in illustrating how readily available data such as satellite imagery can be filtered to provide more meaningful habitat predictors. This paper (Jeganathan *et al*. 2004) also illustrates a more important point about the need to understand the biology of the species in defining habitat predictors. These authors approached the habitat predictor problem from the species' viewpoint, by first identifying habitat features and then relating these directly to the satellite imagery. This is likely to generate more meaningful models than simple data mining in which species' data are related to any conveniently available data set.

Development of the species–habitat model

The majority of ecological modelling is based on generalized linear modelling (GLM) approaches, although simple rule-based approaches have been used in recent species–habitat research on butterflies (Holloway, Griffiths & Richardson 2003) and generalized additive models

have occasionally been used (Pearce *et al*. 2001). A linear model is an equation that contains mathematical variables, parameters and random variables that are linear in the parameters and the random variables (Crawley 1993). A GLM has three components: (i) the linear predictor, effectively relating the response variable to the explanatory variables through a (ii) link function, which relates the linear predictor to the expected value of the response, and (iii) an error structure. Of the GLM techniques, logistic regression is the most frequently used modelling approach in species distribution modelling, because a single record of presence or absence of the target species can be considered to be a binomial trial with a sample size of 1. In this type of GLM the link function is logit and the error structure is assumed to be binomial. Logistic regression has considerable appeal to ecologists because presence–absence data are comparatively easy to collect in the field, even when the zero data set has to be created *post hoc* by a different sampling strategy. Eight of the papers in Table 1 as well as the six in this special profile (Cabeza *et al.* 2004; Engler, Guisan & Rechsteiner 2004; Friar *et al*. 2004; Gibson *et al.* 2004; Jeganathan *et al.* 2004; Johnson, Seip & Boyce 2004) have used logistic regression. However, these models, like any other, should not be used uncritically. Two key assumptions in any GLM application are that the data used as predictors are adequate (in the sense that they are true variables determining the species distribution pattern) and that the error structure is appropriate for the data. The first of these assumptions becomes very important if the predictor variables used in modelling are only surrogates for true predictors, as is the case with data derived from remote-sensed imagery. In a logistic regression model the error model can be accepted as appropriate if the residual deviance (unexplained variation) after model fitting is equivalent to the number of degrees of freedom. If the residual deviance is much greater than the degrees of freedom, the data are 'overdispersed'. Overdispersion can arise because there is a structural failure in the model, such as failing to include key predictor variables that are actually driving the response variable, or because the error model is inappropriate for the data. Bradbury *et al*. (2000) developed models relating the incidence of yellowhammers to habitat features and were able to show that the error structure for their models was appropriate. The predictor variables used in this study were key determinants of habitat suitability and not surrogates. Jaberg & Guisan (2001) undertook a formal test for overdispersion in relating the incidence of bats to habitat features. While comparatively easy to undertake, these analyses are much less frequently reported than they should be. One alternative when data are overdispersed is to use quasi-likelihood approaches, as demonstrated by Ambrosini *et al*. (2002). While this approach has appeal, Aitkin *et al*. (1989) argue that it is only appropriate with small amounts of overdispersion and it is not a substitute for a better defined model.

Most, if not all, modelling inevitably has to be based on data collected from the field. Given the expense of undertaking data collection, many data sets are collected over small areas. In these cases the data sets often show spatial autocorrelation or some other form of non-independence. Many of the recent papers published in *Journal of Applied Ecology* have been marked by one or both of these. Authors have responded differently to these features in their data. The predictor variables collected by Ambrosini *et al*. (2002) were strongly correlated because they were recording the cover of all habitats around a fixed point, and as such they were subject to the unit sum constraint. These authors removed variables from their analyses on the basis of the extent to which the variables were correlated. Augustine, Mugglestone & Buckland (1993) developed an autologistic approach based on the Gibbs sampler, that allowed for the inclusion of autoregressive components in the model, and this has been used extensively. Osborne, Alonso & Bryant (2001) used the same methodology to develop large-scale distribution maps for birds in Spain. Milsom *et al*. (2000) used an autoregressive component to allow for the fact that the marshes on which they recorded wading birds were more similar if they were in the same land holding, but they abandoned the approach in favour of a generalized linear mixed model where they defined land holding as a random effect.

A common feature of much species distribution modelling is that there are often many candidate predictor variables. To the non-statistician, a surfeit of predictors looks like a good thing. In reality, however, finding the 'best' model amongst the many possibilities is not an easy task. The number of candidate models increases by two factorial the number of predictors available and the 'best' model identified may depend critically on the route taken to find it. With a large suite of predictor variables, it is also possible to 'overfit' to the extent that models often perform very well in the context of the data set used to create them but fail to be robust when used elsewhere. Overfitting obviously has major implications for the applied value of the work, as models only have real utility if they have a general application in areas other than those from which they were created. Collingham *et al*. (2000) attempted to circumvent this problem by using both stepwise forwards inclusion and stepwise backwards deletion of variables in their models.

More recently there has been an increasing trend to use Akaike's information criterion (AIC) in model selection. AIC forms the basis of information-theoretic approaches to modelling and is considered by Gibson *et al*. (2004) in this special profile, but its use in the present GLM context has largely been associated with identifying variables for inclusion or exclusion in models. AIC is actually equivalent to twice the log-likelihood of the model fitted plus two times the number of parameters estimated in its formation. Given that the model with the smallest log-likelihood is considered to be that with the best fit, the addition of two times the number of parameters means that AIC effectively includes a penalty

for adding predictor variables to the model. Thus, AIC aids in identifying the most parsimonious model amongst a set. These approaches have been used in vertebrate studies particularly. Manel, Williams & Ormerod (2001) used it for birds and Jaberg & Guisan (2001), Schadt *et al*. (2002) and Johnson, Seip & Boyce (2004) for mammals. Studies of large organisms typically generate data sets that are overparameterized in relation to the range of potential predictor variables. The main reason for this is the cost associated with collecting data for the target species. For example, radio-tracking requires monitoring individual animals over a long time period (Schadt *et al*. 2002). Furthermore, large animals are often rare and the effort associated with collecting data may be huge. Overparameterization can be reduced by data simplification. Manel, Williams & Ormerod (2001), Gates & Donald (2000) and Suárez-Seoane, Osborne & Alonso (2002) used principal components analysis (PCA) to create axes that summarized trends in habitat characteristics for use as predictors in their analyses of bird incidence, but other approaches are in development.

Model assessment

The third component of species distribution modelling is model assessment. This is not only an important check on the adequacy of the models developed but it is an important factor determining their utility. In an applied sense, models have their greatest utility when they can be used predictively and not simply as a means of exploring putative relationships in a data set. There are several ways in which species distribution models can be assessed and recent papers in *Journal of Applied Ecology* not only illustrate the breadth of approaches but they have also been innovative in implementing new approaches. The standard method for assessing all linear models is to test the null hypothesis, that the regression coefficients estimated from the model are no different to zero. While there has been substantial criticism over the use of hypothesis testing of this sort in distribution modelling (Burnham & Anderson 2002) and authors have recognized that these approaches may not be appropriate (Milsom *et al*. 2000), the Wald test (Aitkin *et al*. 1989) is still the most frequently used method for assessing the significance of variables when included in the model. More sophisticated approaches have been based on jack-knifing, where authors have sampled the data set and fitted models repeatedly to assess how robust the parameter estimates are to change in the input data set.

The simplest way to assess any model is to compare model predictions with observed data. Assessment of logistic regression models is complicated by the fact that the predictions from such models are proportions in the range 0–1 while most test data against which they may be compared comprise records of 1 or 0. In effect, model assessment involves comparing probabilities with categories. One obvious approach is to use thresholds in the predictions, above and below which presence and absence are defined. There are a number of metrics that can be used to compare model predictions and observed data using thresholds and, of these, the kappa statistic has been increasingly used in model testing in recent *Journal of Applied Ecology* papers (Cowley *et al*. 2000; Jaberg & Guisan 2001; Manel, Williams & Ormerod 2001; Ambrosini *et al*. 2002). This statistic is, however, sensitive to sample size and fails if one class (the presences or absences) exceeds the other (Fielding & Bell 1997). *Journal of Applied Ecology* has seen a recent increase in the use of threshold-independent approaches (Manel, Williams & Ormerod 2001; Osborne, Alonso & Bryant 2001; Schadt *et al*. 2002; Suárez-Seoane, Osborne & Alonso 2002), such as receiver operator characteristic statistics (ROC plots), in the assessment of logistic regression models. These are based on plotting the true positives against the false positive fractions for a range of thresholds in prediction probability. The area under the curve for a ROC plot is taken as a measure of the accuracy of the model that is not dependent on a single threshold. ROC plots have now entered the toolbox of species distribution modelling and in this special profile are used by Frair *et al*. (2004) in assessing the model classification accuracy of positioning animals in the landscape, Jeganathan *et al*. (2004) in their models of Jerdon's courser, Gibson *et al*. (2004) in their information-theoretic approach and Engler *et al.* (2004) in their study of *Eryngium alpinum* L.

Information-theoretic approaches: a new paradigm?

While papers in *Journal of Applied Ecology* have been shaping the debate and developing best practice on species distribution modelling over several years, there have been dramatic changes in the philosophy behind the modelling process. In what may well be seen as a paradigm shift, Burnham & Anderson (2002) have argued for a rethink on modelling approaches. Most, if not all, modelling has been based on a null hypothesis-testing approach. Predictor variables are accepted within a model if there is a suitable decline in residual deviance after their inclusion, or if the regression coefficient is significantly different from zero. Burnham & Anderson (2002) stressed that hypothesis testing has a very important role in the design and analysis of experiments where the scientist has control over both the response and the predictor variables in the search for any causal link, and where there is randomization of treatment and controls. They conclude that the value of this approach in the analysis of observational studies is less clear. When there are many potential predictor variables, model selection is usually based on stepwise consideration of variables, and reliance on hypothesis testing forces ecologists into a decision-making approach focusing on arbitrary levels of statistical significance. Decisions to include or exclude variables in the final model are based on significance levels, which are by definition always

arbitrary. The model with the least residual deviance is selected as the best. It is possible to overfit models to the extent that they appear to explain variation in the observed data set, but perform poorly when used in other circumstances. Cherry (1998) made the point that the null hypothesis in such models, that the regression coefficients are equal to zero, is in most cases uninteresting relative to the need to know how much the individual variables determine the relationship. Being able to compare the relative contribution of predictor variables to the overall species distribution provides more information and leads to greater understanding of the underlying ecological processes driving it. In a sense, this is a shift from a focus on the formal statistical basis for a particular model linking species to potential predictor variables, to one orientated more towards the processes themselves.

The information-theoretic approach is based on formulating a series of models that rely on an understanding of the system being studied, followed by an assessment of how the different putative model(s) compare(s) to reality. The scientist then selects that (or possibly a small set of) model(s) that is nearer to reality than any of the rest. The basis for comparing models is AIC. The AIC can be used to determine Akaike weights for each model, which are the weights of evidence in favour of each model being the nearest to reality, given the other models being considered. The most obvious feature of this approach is that it is comparative, and leads to the identification of the best amongst a suite of models. More importantly, inference and prediction need not necessarily be based on one (the best) model. Where there are several alternative models, each supported by the data, then the suite of good models should be used in inference and prediction using model-averaging approaches.

As with all modelling, information-theoretic approaches should not be used uncritically. A key issue is having sufficient knowledge of the system being studied so that ecologically meaningful models can be formulated, as these approaches can only identify the best amongst a set that has been identified *a priori*. If the predictor variables are poor representations of the underlying factors then the approach can only identify the best amongst what is effectively a poor set of models. Eberhardt (2003) makes the point that model selection should not only be based on AIC but also on statistics like r^2, which provide a more 'global' measure of how good the model is at explaining the data. He also argues that the development of *a priori* models is not always easy without some form of preliminary analysis of data, and data exploration of this type is usually based on some form of hypothesis testing.

Of the six papers in this special profile, three involve information-theoretic metrics in analysis of data, and Gibson *et al*. (2004) adopted the information-theoretic approach directly advocated by Burnham & Anderson (2002). These authors faced the problem that the essential components of suitable habitat for the rufous bristlebird *Dasyornis broadbenti* were not well defined, so in a sense they were only able to use the information-theoretic approach in an exploratory way. Nonetheless, they provide a very clear illustration of the modelling approach and the application of jack-knifing and model averaging. They identified four models out of 32 as having some support from their data and went on to generate an average model that they implemented as a predictive model in a GIS. This paper (Gibson *et al*. 2004) adopts much of the philosophy advocated by Burnham & Anderson (2002), and considers *post-hoc* evaluation of the model performance. We expect that this paper will provide a good basis for future work in species distribution modelling.

Perspectives

With the increased availability of remote-sensed data, GIS and statistical packages, it is becoming increasingly easy to undertake species distribution modelling. Easy access to data and computer software means that analyses can be undertaken almost as a matter of routine for any species for which there are distribution records. This creates opportunities for applied ecologists to develop management tools for conservation in a way that was unprecedented 15 years ago. The introduction of information-theoretic approaches to species distribution modelling, while innovative in terms of model selection procedures, will probably be of greater significance because it argues for a bottom-up approach to distribution modelling. It requires ecologists to consider more closely what the organisms are doing before even considering the modelling process. This will inevitably mean that applied ecologists will develop models based on ecological principles that will be more robust and of greater utility. The coupling of data collection and information-theoretic modelling philosophy shown in five papers (Cabeza *et al.* 2004; Friar *et al*. 2004; Gibson *et al.* 2004; Jeganathan *et al.* 2004; Johnson, Seip & Boyce 2004) in the special profile is another example of how the application of ecology is increasingly both driving and underpinning the advance of basic science: the need to solve real-world problems in conservation biology has placed applied ecology at the forefront of methodological developments in distribution modelling, which we expect will soon pervade the whole subject of ecology.

References

Aitkin, M., Anderson, D., Francis, B. & Hinde, J. (1989) *Statistical Modelling in GLIM*. Oxford Science Publications, Oxford University Press, Oxford, UK.

Ambrosini, R., Bolzern, A.M., Canova, L., Arieni, S., Møller, A.P. & Saino, N. (2002) The distribution and colony size of barn swallows in relation to agricultural land use. *Journal of Applied Ecology*, **39**, 524–534.

Augustine, N.H., Mugglestone, M.A. & Buckland, S.T. (1993) An autologistic model for the spatial distribution of wildlife. *Journal of Applied Ecology*, **33**, 339–348.

Bradbury, R.B., Kyrkos, A., Morris, A.J., Clark, S.C., Perkins, A.J. & Wilson, J.D. (2000) Habitat associations

and breeding success of yellowhammers on lowland farmland. *Journal of Applied Ecology*, **37**, 789–805.

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

Cabeza, M., Araújo, M.B., Wilson, R.J., Thomas, C.D., Cowley, M.J.R. & Moilanen, A. (2004) Combining probabilities of occurrence with spatial reserve design. *Journal of Applied Ecology*, **41**, 252–262.

Cherry, S. (1998) Statistical test in publications of the Wildlife Society. *Wildlife Society Bulletin*, **26**, 947–953.

Collingham, Y.C., Wadsworth, R.A., Huntley, B. & Hulme, P.E. (2000) Predicting the spatial distribution of non-indigenous riparian weeds: issues of spatial scale and extent. *Journal of Applied Ecology*, **37** (Supplement), 13–27.

Cowley, M.J.R., Wilson, R.J., Leon-Cortes, J.L., Guitierrez, D., Bulman, C.R. & Thomas, C.D. (2000) Habitat-based statistical models for predicting the spatial distribution of butterflies and day-flying moths in a fragmented landscape. *Journal of Applied Ecology*, **37**, 60–72.

Crawley, M. (1993) *GLIM for Ecologists*. Blackwell Scientific Publications, Oxford, UK.

Eberhardt, E.E. (2003) What should we do about hypothesis testing? *Journal of Wildlife Management*, **67**, 241–247.

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. *Journal of Applied Ecology*, **41**, 263–274.

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

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. *Journal of Applied Ecology*, **41**, 201–212.

Gates, S. & Donald, P.F. (2000) Local extinction of British farmland birds and the prediction of further loss. *Journal of Applied Ecology*, **37**, 806–820.

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. *Journal of Applied Ecology*, **41**, 213–223.

Holloway, G.J., Griffiths, G.H. & Richardson, P. (2003) Conservation strategy maps: a tool to facilitate biodiversity action planning illustrated using the heath fritillary butterfly. *Journal of Applied Ecology*, **40**, 413–422.

Jaberg, C. & Guisan, A. (2001) Modelling the distribution of bats in relation to landscape structure in a temperate mountain environment. *Journal of Applied Ecology*, **38**, 1169–1181.

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. *Journal of Applied Ecology*, **41**, 224–237.

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. *Journal of Applied Ecology*, **41**, 238–251.

Manel, S., Buckton, S.T. & Ormerod, S.J. (2000) Testing large-scale hypotheses using surveys: the effects of land use on the habitats, invertebrates and birds of Himalayan rivers. *Journal of Applied Ecology*, **37**, 756–770.

Manel, S., Williams, H.C. & Ormerod, S.J. (2001) Evaluating presence–absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, **38**, 921–931.

Milsom, T.P., Langton, S.D., Parkin, W.K., Peel, S., Bishop, J.D., Hart, J.D. & Moore, N.P. (2000) Habitat models of bird species' distribution: an aid to the management of coastal grazing marshes. *Journal of Applied Ecology*, **37**, 706–727.

Osborne, P.E., Alonso, J.C. & Bryant, R.G. (2001) Modelling landscape-scale habitat use using GIS and remote sensing: a case study with great bustards. *Journal of Applied Ecology*, **38**, 458–471.

Pearce, J.L., Cherry, K., Drielsma, M., Ferrier, S. & Whish, G. (2001) Incorporating expert opinion and fine-scale vegetation mapping into statistical models of faunal distribution. *Journal of Applied Ecology*, **38**, 412–424.

Schadt, S., Revilla, E., Wiegand, T., Knauer, F., Kaczensky, P., Breitenmoser, U., Bufka, L., Cerveny, J., Koubek, P., Huber, T., Stanisa, C. & Trepl, L. (2002) Assessing the suitability of central European landscapes for the reintroduction of Eurasian lynx. *Journal of Applied Ecology*, **39**, 189–203.

Suárez-Seoane, S., Osborne, P.E. & Alonso, J.C. (2002) Large-scale habitat selection by agricultural steppe birds in Spain: identifying species–habitat responses using generalized additive models. *Journal of Applied Ecology*, **39**, 755–771.

Vaughan, I.P. & Ormerod, S.J. (2003) Modelling the distribution of organisms for conservation: optimising the collection of field data for model development. *Conservation Biology*, **17**, 1601–1611.

Vaughan, N., Lucas, E.-A., Harris, S. & White, P.C.L. (2003) Habitat associations of European hares *Lepus europaeus* in England and Wales: implications for farmland management. *Journal of Applied Ecology*, **40**, 163–175.