Biodiversity and Conservation **11**: 2275–2307, 2002.
© 2002 *Kluwer Academic Publishers. Printed in the Netherlands.*

Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling

SIMON FERRIER[1,*], GRAHAM WATSON[1,2], JENNIE PEARCE[1,3] and MICHAEL DRIELSMA[1]
*[1]New South Wales National Parks and Wildlife Service, P.O. Box 402, Armidale, NSW 2350, Australia; [2]Present address: Department of Land and Water Conservation, P.O. Box 245, University of New England, Armidale, NSW 2351, Australia; [3]Present address: Canadian Forest Service, 1219 Queens St East, Sault Ste. Marie, Ontario, Canada P6A 2E5; *Author for correspondence (e-mail: simon.ferrier@ npws.nsw.gov.au)*

Received 15 November 2001; accepted in revised form 10 May 2002

Abstract. Statistical modelling of biological survey data in relation to remotely mapped environmental variables is a powerful technique for making more effective use of sparse data in regional conservation planning. Application of such modelling to planning in the northeast New South Wales (NSW) region of Australia represents one of the most extensive and longest running case studies of this approach anywhere in the world. Since the early 1980s, statistical modelling has been used to extrapolate distributions of over 2300 species of plants and animals, and a wide variety of higher-level communities and assemblages. These modelled distributions have played a pivotal role in a series of major land-use planning processes, culminating in extensive additions to the region's protected area system. This paper provides an overview of the analytical methodology used to model distributions of individual species in northeast NSW, including approaches to: (1) developing a basic integrated statistical and geographical information system (GIS) framework to facilitate automated fitting and extrapolation of species models; (2) extending this basic approach to incorporate consideration of spatial autocorrelation, land-cover mapping and expert knowledge; and (3) evaluating the performance of species modelling, both in terms of predictive accuracy and in terms of the effectiveness with which such models function as general surrogates for biodiversity.

Key words: Biodiversity, Northeast New South Wales, Regional conservation planning, Statistical modelling, Surrogates

Introduction

A fundamental challenge confronting all efforts to retain biodiversity through *in situ* protection is deciding which areas are most worthy of conservation action. The total area of land that can be set aside, or otherwise managed, for conservation is often severely limited by social and economic constraints. Care must therefore be taken to direct scarce resources to areas of highest conservation priority, defined in terms of both conservation value and degree of threat (Margules and Pressey 2000). Assessments at a global or continental scale can help to focus attention on broad regions (e.g. ecoregions) of particular conservation concern (e.g. Olson and Dinerstein 1998;

Myers et al. 2000). However, a more detailed assessment is usually required within each of these regions to guide decisions about the actual location of conservation areas – whether these be strict reserves or areas protected by a range of other measures, including multiple-use management zones and conservation incentives and controls on private land. This process of prioritising conservation action within rather than between regions is referred to here as 'regional conservation planning' (Margules and Redhead 1995; Dinerstein et al. 2000; Groves et al. 2000).

To achieve conservation of biodiversity, regional planning must not only identify conservation areas that will include or 'represent' as many elements of biodiversity as possible, but must also ensure that these areas are sufficiently large, well connected and well replicated to promote long-term persistence of the diversity they encompass (Smith et al. 1993; Cowling et al. 1999). As a result of nearly two decades of research and development work on 'systematic' conservation planning techniques, there now exists a wide range of approaches, algorithms and software packages for designing systems of conservation areas that are representative of the biodiversity of a given region (see reviews by Pressey et al. 1993; Margules and Pressey 2000). A prerequisite for using any of these approaches is the existence of information on the spatial distribution of biodiversity. Unfortunately, such information is usually grossly incomplete. Most entities of biodiversity – particularly at the species and genetic levels – have not yet been discovered, let alone had their distributions mapped at a spatial scale appropriate for regional conservation planning.

A widely applied solution to this problem is to use those entities for which we do have distributional information as 'surrogates' for spatial pattern in biodiversity as a whole (Noss 1990; Humphries et al. 1995; Vane-Wright 1996; Ferrier 1997; Margules and Pressey 2000). Such surrogates commonly include species of particular ecological or social significance (e.g. threatened, focal or flagship species; Lambeck 1997; Simberloff 1998; Caro and O'Doherty 1999) or all species within one or more indicator groups (e.g. beetles, Anderson and Ashe 2000, or butterflies, Kremen 1994). However, even for these surrogate species or groups of species, available information on fine-scaled spatial distribution is usually far from complete. Knowledge of species distributions is derived primarily from locational records – i.e. a species is observed or collected at a particular geographical location. For most regions, the geographical coverage of such information is sparse (Margules and Austin 1994; Lawes and Piper 1998; Maddock and du Plessis 1999; Soberón et al. 2000). Survey or collection sites are often separated by extensive tracts of unsurveyed land. Furthermore, the location of these sites is often biased towards population centres and access routes.

One way of filling geographical gaps in information on species distributions is to use available biological survey data to derive statistical models relating species presence or abundance to remotely mapped environmental variables – e.g. terrain, climate, substrate or land-cover variables. By integrating such modelling with geographical information system (GIS) technology, biological distributions can be extrapolated across large regions, thereby providing geographically complete information for a wide range of environmental applications. The popularity of this approach has increased dramatically in recent years, as evidenced by a rapidly growing scientific literature on statistical and related techniques for modelling biological distributions, and on the application of these techniques to environmental planning and management (see reviews by Franklin 1995; Austin 1998; Guisan and Zimmermann 2000).

This paper adds to the existing literature by providing an overview of modelling work conducted in the northeast New South Wales (NSW) region of Australia. This particular case of the application of distributional modelling to regional conservation planning is one of the most extensive to date anywhere in the world. Statistical modelling and GIS-based extrapolation of species distributions was first applied in the region in the early 1980s and has since been used to model distributions of over 2300 individual species of plants and animals, in addition to modelling distributions of a variety of higher-level communities and assemblages. During the past 6 years, these modelled biological distributions have played a pivotal role in a series of government-led planning processes, resulting in major additions to the region's protected area system. The extensive biological and environmental datasets established for northeast NSW have also been employed as a test-bed for research on the performance of alternative modelling approaches and techniques.

This paper provides a broad overview of the analytical methodology used to model distributions of individual species in northeast NSW. A companion paper (Ferrier et al. 2002; this issue) describes the approaches used to model distributions of communities and assemblages. We start by outlining the planning context within which the modelling work was performed, and the specific roles that distributional modelling played in this planning. Next we describe how statistical modelling was integrated with GIS technology to provide a basic, yet robust, approach to modelling species distributions by linking biological survey data to remotely mapped environmental variables. We indicate a number of ways in which this modelling was refined by extending the basic approach to incorporate consideration of spatial autocorrelation, land-cover mapping and expert knowledge. We then describe how the performance of species models was evaluated, both in terms of predictive accuracy and in terms of the effectiveness with which such models function as general surrogates for biodiversity. Readers interested in accessing more detailed information on any particular aspect of the work are encouraged to consult the cited papers and reports.

Biodiversity modelling and conservation planning in northeast New South Wales

The planning context

The work described in this paper concerns three adjoining bioregions (Thackway and Creswell 1997) in northeast NSW – the NSW North Coast Bioregion, the New England

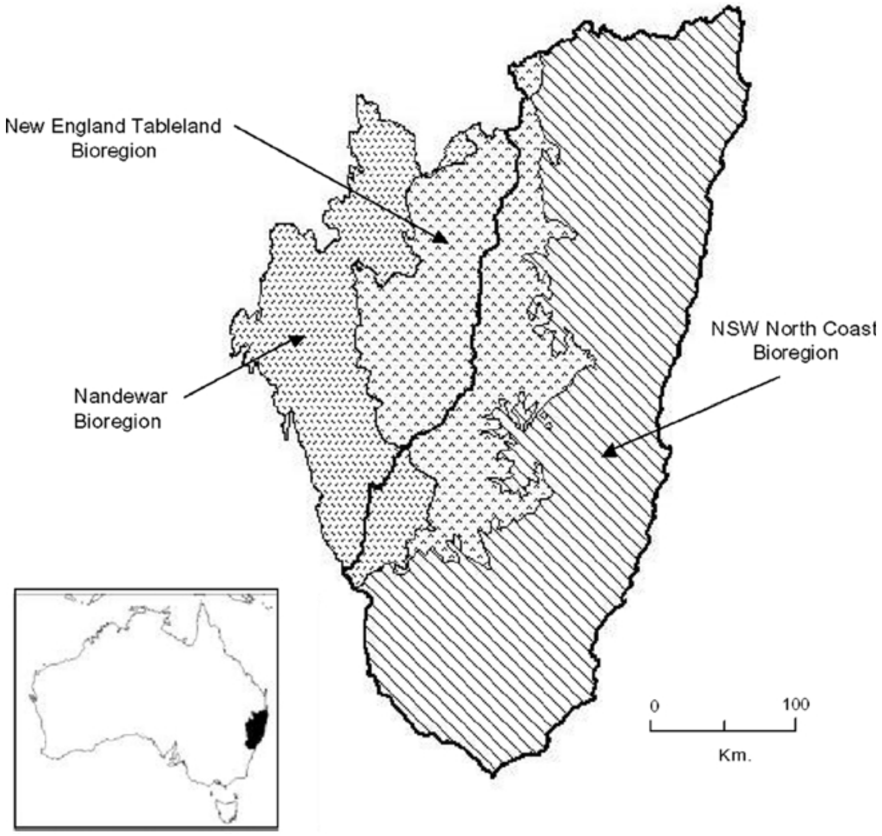

Figure 1. The three bioregions addressed by this study. The bold line delineates the area within which most modelling work was conducted.

Table 1 and Bioregion and the Nandewar Bioregion (Figure 1). The combined area of these bioregions is 106 400 km^2, of which almost 50% has been cleared of native vegetation. The vegetation of northeast NSW is a mosaic of rainforest, moist and dry eucalypt forest and dry eucalypt woodland, with smaller areas of swamp, heath, scrub and native grassland. Most of the rainforest and taller eucalypt forest is confined to the NSW North Coast Bioregion and the eastern half of the New England Table 1 and Bioregion. These forests are recognised as constituting one of the major centres of floristic diversity on the Australian continent, and support one of Australia's richest and most diverse vertebrate faunas, including a high number of endemic and endangered species (Ferrier et al. 2000b). These same forests are also a major source of native timber, and have therefore been the focus of a long-running conflict between the needs of commercial forest harvesting and the protection of biodiversity, old growth and wilderness values. Much of the biological survey and modelling work conducted in northeast NSW has therefore been concentrated within these forested areas (Figure 1).

During the past decade, most assessment and planning initiatives for Australia's forests have been underpinned by a National Forest Policy Statement developed jointly by the Commonwealth, State and Territory Governments (Anonymous 1992). In addressing the issue of nature conservation the statement proposed that "parts of the public native forest estate will continue to be set aside in dedicated nature conservation reserve systems to protect native forest communities, based on principles of comprehensiveness, adequacy and representativeness" and that "there will be complementary management outside reserves".

The development of a comprehensive, adequate and representative reserve system within the public forests of northeast NSW has proceeded in three main stages: (1) The Deferred Forest Assessment completed in 1995 evaluated existing levels of reservation of biodiversity, old growth and wilderness values and identified areas of forest that were to be deferred from logging pending more rigorous evaluation. (2) The Interim Forest Assessment Process completed in 1996 refined the areas identified by the Deferred Forest Assessment through a more intensive analysis of all available data. (3) The Comprehensive Regional Assessment, conducted over a 3-year period (1996–1998), was the most rigorous of the three assessment stages and led to the identification and gazettal of extensive additions to the reserve system as part of a Regional Forest Agreement between the State and Commonwealth Governments.

Since the completion of the Comprehensive Regional Assessment, the datasets and analytical approaches developed in northeast NSW have continued to be employed in a number of ongoing assessment and planning activities. The most notable of these are the NSW Regional Vegetation Planning process which, unlike the processes described above, is focusing on private (freehold) land rather than public land, and a bioregional conservation assessment within the Nandewar Bioregion (Figure 1).

Data sources

Work on the environmental and biological databases that have underpinned the above assessments commenced many years before the first planning decisions were made by the Deferred Forest Assessment in 1995. In the late 1980s the NSW National Parks and Wildlife Service (NSW NPWS) initiated the establishment of an environmental GIS database for northeast NSW, containing mapped and modelled layers pertaining to topography, climate, substrate, vegetation cover and disturbance. This database was consolidated as part of the North East Forests Biodiversity Study conducted by NSW NPWS between 1991 and 1994 (Ferrier et al. 2000a), and has been further refined during the series of forest assessment processes conducted since 1995 (Ferrier 2000). The database was initially established within the Environmental Resource Mapping System (E-RMS), a PC-based GIS package developed in-house by NSW NPWS (Ferrier 1992b), but was later transferred to ArcView (ESRI). Most environmental layers in the database are currently stored at a 1 ha (100 m × 100 m) grid-cell resolution.

A fine-scaled digital elevation model (DEM) based on 1:25 000 topographic data was used to derive a number of topographic indices including slope, aspect, wetness (or compound topographic) index, topographic position and ruggedness. The DEM was also coupled with ESOCLIM climate-surface models (Hutchinson et al. 1997) to derive estimates of long-term mean temperature, rainfall, solar radiation (adjusted for surrounding terrain) and evaporation (also terrain-adjusted). Soil fertility was modelled as a function of mapped lithology and soil landscapes, guided by expert opinion and information extracted from soil surveys and geochemical analyses. Soil depth was modelled as a function of lithology, terrain and climate using depth data obtained from all available soil surveys. Modelled soil depth was further combined with monthly rainfall and evaporation surfaces to derive a soil moisture index based on a simple water balance model.

Vegetation cover was mapped at two spatial scales: (1) broad-scaled (1:100 000) mapping of structural systems derived from Landsat TM imagery; and (2) fine-scaled (1:25 000) mapping of floristic types derived from interpretation of aerial photography. The fine-scaled mapping was initially confined to forests on public land, but was later extended to cover vegetation on all land tenures as part of the Comprehensive Regional Assessment. This assessment also divided all mapped areas of forest into growth stages reflecting the intensity of, and time since, logging and other disturbance.

In 1991 NSW NPWS initiated an extensive program of flora and fauna surveys in northeast NSW, as part of the North East Forests Biodiversity Study (Hines et al. 2000). The surveys were designed to collect data that would supplement existing biological datasets. Survey sites were located according to an environmental stratification based on the GIS layers described above, thereby ensuring that sites were well spread across the environmental variation of the region. The surveys targeted mainly vascular flora and vertebrate fauna, although a subset of the fauna sites was also surveyed for ground-dwelling arthropods (ants, beetles and spiders) in a joint project with the Australian Museum. By 1995 over 277 000 locational records had been assembled for 4207 species of vascular plants, vertebrates and ground-dwelling arthropods. These data formed the basis for most of the modelling work described in this paper.

Additional flora and fauna surveys conducted during the Comprehensive Regional Assessment (1996–1998) were designed to fill remaining environmental and geographical gaps in the coverage of survey sites throughout the forests of northeast NSW. Since 1999 most survey effort has been redirected to the eucalypt woodlands of the Nandewar Bioregion and the western half of the New England Tableland Bioregion.

The role of modelling

Despite the extent of biological survey data collected during the North East Forests Biodiversity Study and subsequent assessments these data did not, on their own, provide sufficient spatial coverage for conservation planning. Most planning decisions needed to be made at the scale of individual forestry compartments with an average area of approximately 200 ha. Only a small proportion of these compartments had been subjected to any kind of direct biological survey. Statistical modelling was therefore used to extrapolate biological distributions across unsurveyed parts of the region by modelling relationships between available biological survey data and remotely mapped environmental layers (Ferrier and Smith 1990; Ferrier 1991, 1997; Ferrier and Watson 1997; Ferrier et al. 2000d).

Decisions about the location of new forest reserves in northeast NSW were guided by a set of agreed regional protection targets for various entities of biodiversity, old growth forest and wilderness (Commonwealth of Australia 1997). In the case of biodiversity, protection targets were specified for two types of entities – (1) forest communities (or ‘ecosystems’) that served as a general surrogate, or ‘coarse-filter’ (*sensu* Noss 1987), for biodiversity as a whole, and (2) individual species of particular conservation concern (i.e. ‘fine-filter’ species) (Figure 2). Protection targets for communities were specified as a percentage of the original (preclearing) area of each community, while targets for individual species were specified as a minimum viable habitat area based on expert opinion and, where appropriate, population viability analysis. Statistical modelling played a role in mapping the spatial distribution of both types of entities.

Modelled distributions of species and communities were integrated with spatial data on other conservation entities (old growth forest, wilderness) and socio-economic values (Figure 2), within an interactive decision-support system developed

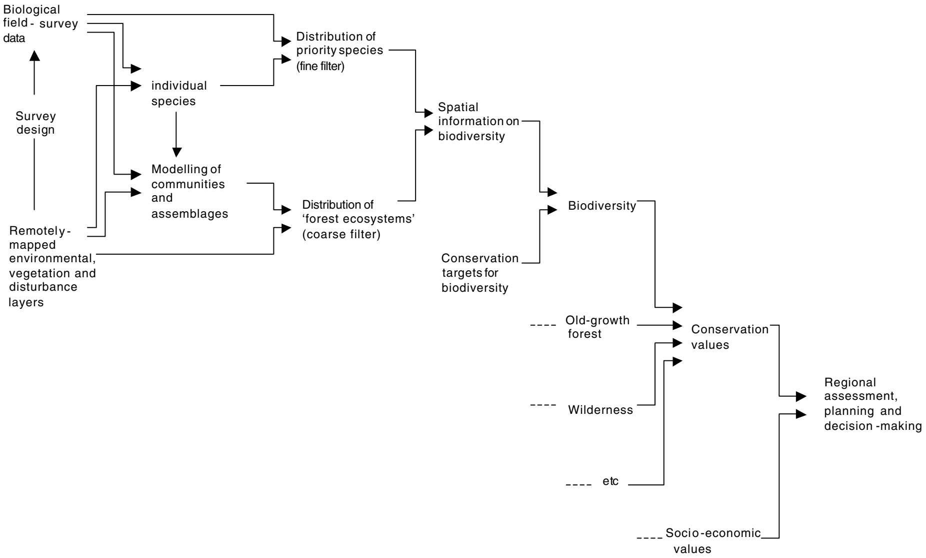
Biological field - survey data
Survey design
Remotely-mapped environmental, vegetation and disturbance layers
individual species
Modelling of communities and assemblages
Distribution of priority species (fine filter)
Distribution of 'forest ecosystems' (coarse filter)
Spatial information on biodiversity
Conservation targets for biodiversity
Biodiversity
Old-growth forest
Wilderness
etc
Conservation values
Socio-economic values
Regional assessment, planning and decision-making

Figure 2. The role of modelling in relation to the overall framework for regional conservation planning in northeast NSW.

by NSW NPWS. This system was employed by teams of negotiators to map conservation priorities across the region, to assess alternative conservation scenarios, and to record land-use decisions (for further information on the decision-support system and the decision-making process see Finkel 1998; Pressey 1998, 1999; Ferrier et al. 2000c).

Developing a robust statistical approach to modelling species distributions

The approach used to model species distributions in northeast NSW needed to be capable of handling a vast quantity of data for a very large number of species. Furthermore, these species needed to be modelled and extrapolated in relation to relatively fine-scaled environmental variables, thereby producing mapped outputs at an appropriate spatial resolution for the planning task at hand. This demanded a high level of automation, both in the fitting of statistical models and in the extrapolation of these models to produce spatial GIS layers for use in conservation planning and decision-making. Despite the need for automation there was also a clear requirement that distributions should be predicted as accurately as possible, and that the level of reliability or uncertainty associated with each model should be clearly communicated to users.

Background to the general approach

The use of statistical modelling to predict potential distributions of species has a relatively long history of application in northeast NSW. In the early 1980s the regional distribution of a single species of conservation concern – the Rufous Scrub-bird *Atrichornis rufescens* – was mapped by using generalised linear modelling (GLM; McCullagh and Nelder 1989) to derive a logistic regression model relating field survey data to coarse-scaled climate, terrain and vegetation variables stored in a rudimentary GIS database (Ferrier 1984, 1991). This appears to be one of the earliest applications anywhere in the world of the integrated use of GLM and GIS to model species distributions. In the late 1980s NSW NPWS further employed GLM to model several other species in northeast NSW, aided by the parallel refinement of environmental GIS layers for the region. During this period considerable effort was also devoted to developing and applying an alternative approach to species modelling based on classification and regression trees (or ‘decision-trees’) (Brieman et al. 1984; Ferrier and Smith 1990; Stockwell et al. 1990; Moore et al. 1991). This work culminated in the development of a predictive modelling module for E-RMS. The module facilitated rapid derivation and spatial extrapolation of decision-tree models within a seamless GIS environment. The decision rules constituting a model could be derived either by automated induction based on the analysis of biological survey data in relation to environmental GIS layers, or by interactive specification based on expert knowledge (Ferrier 1992a).

With the commencement of the North East Forests Biodiversity Study in 1991, attention shifted to evaluating generalised additive modelling (GAM; Hastie and Tibshirani 1990; Yee and Mitchell 1991) as a possible alternative to GLM and decision-tree modelling. Preliminary trials suggested that GAM offered a good compromise between the statistical rigour and parametric power of GLM and the non-parametric flexibility of decision-tree modelling and other 'machine learning' approaches such as artificial neural networks and genetic algorithms. GAM was therefore adopted as the principal technique used to model species distributions in the North East Forests Biodiversity Study (1991–1994), and in all subsequent conservation assessments in the region.

GAM is an extension of GLM, which is itself an extension of ordinary linear regression. Linear regression fits linear (straight line) functions relating a response (dependent) variable to one or more predictor (independent) variables. A basic assumption of linear regression is that the relationship between the response variable and each of the predictors can be approximated by a straight line. A further assumption is that the variance associated with the response is homogeneous across the full range of response values. GLM relaxes both of these assumptions by providing a class of models that allow non-linearity and heterogeneous variance in response variables. Each class of GLM (e.g. logistic regression for modelling binary response data) is defined in terms of a link function that specifies the relationship between the mean of the response and the linear predictor (the sum of the effects of the individual predictor variables), and a variance function which relates the variance of the response to its mean. Once appropriate link and variance functions have been specified, models are fitted by iteratively reweighted least squares (McCullagh and Nelder 1989).

The flexibility of GLM for modelling species responses to environmental variables is still limited by the linear nature of the predictor employed in the link function. This limitation can be partly relieved by adding polynomial terms – e.g. both linear and quadratic terms can be included for each environmental variable to accommodate symmetric bell-shaped response curves (Austin et al. 1984). GAM provides a much more natural and flexible solution to this problem (Yee and Mitchell 1991; Leathwick 1995; Austin and Meyers 1996). Models fitted using GAM have the same link and variance functions as those fitted using GLM, except that the effect of each predictor variable is specified as a non-parametric smooth function, estimated from the data using techniques originally developed for smoothing scatterplots (most commonly cubic splines). The principal difference between GAM and GLM in modelling species distributions is that GAM allows the survey data to determine the shape of response curves, instead of being constrained by any particular parametric form. In other words, fewer assumptions are made about how species respond to their environment.

The progression from GLM to GAM as the primary basis for modelling species distributions in northeast NSW has also occurred in several other parts of the world, including New Zealand (e.g. Leathwick 1998), the USA (e.g. Franklin 1998) and Europe (e.g. Bio et al. 1998).

All models derived during the North East Forests Biodiversity Study and subsequent assessments were fitted using GAM-based logistic regression, within the S-PLUS statistical package (MathSoft). To facilitate spatial extrapolation of species distributions, special in-house software was developed to interface the S-PLUS modelling functions to GIS software – initially to E-RMS, and later to ArcView. The basic components of this integrated system are depicted in Figure 3 (see also Ferrier et al. 2000d).

To fit models for one or more species, data for all relevant survey sites were extracted from the GIS, and associated database management system (DBMS) tables, and loaded into an S-PLUS data frame (a rectangular matrix). Each row of this data frame represented a surveyed site. The columns of the data frame were arranged in two blocks. The first block contained values for the predictors – both environmental predictors (i.e. values for environmental GIS variables at each site) and any additional covariates relating to the survey effort expended at each site (e.g. number of trap-nights) and the conditions under which each site was surveyed (e.g. time of year, weather conditions). The second block contained a column for each of the species to be modelled, indicating the presence or absence of that species at each surveyed site (0 = absent, 1 = present).

Once an appropriate data frame had been constructed, S-PLUS was used to fit a GAM-based logistic regression model to the data for each species. The predictors included in each model were chosen using a simple forward selection procedure, in which predictors were added to the model one at a time (we note however that forward selection is now generally regarded as being slightly less effective than backward selection of predictors; T.J. Hastie, personal communication). The predictor selected at each step was that which best improved the fit of the model, in terms of reduction in deviance. This process continued until none of the remaining predictors could significantly improve the fit (at the $P < 0.05$ significance level). During the forward selection process all functions for continuous predictors were fitted using cubic smoothing splines with four degrees of freedom. Once all predictors to be included in a model had been selected, each continuous predictor was re-evaluated to determine whether the function fitted with four degrees of freedom could be replaced by a simpler function with three or two degrees of freedom, without incurring a significant ($P < 0.05$) increase in the deviance of the model.

Two types of output were produced for each model: (1) plots (graphs) depicting the fitted functions relating probability of occurrence to each selected predictor; and (2) spatial GIS layers depicting probability of occurrence throughout the entire region, as predicted by the model. The plots of fitted functions were derived using special-purpose software developed in S-PLUS. A separate plot was produced for each predictor. Each plot depicted the expected probability of occurrence of a species in relation to varying values of a given predictor, holding all other predictors constant at their mean values. Probability of occurrence was plotted on a cube-root

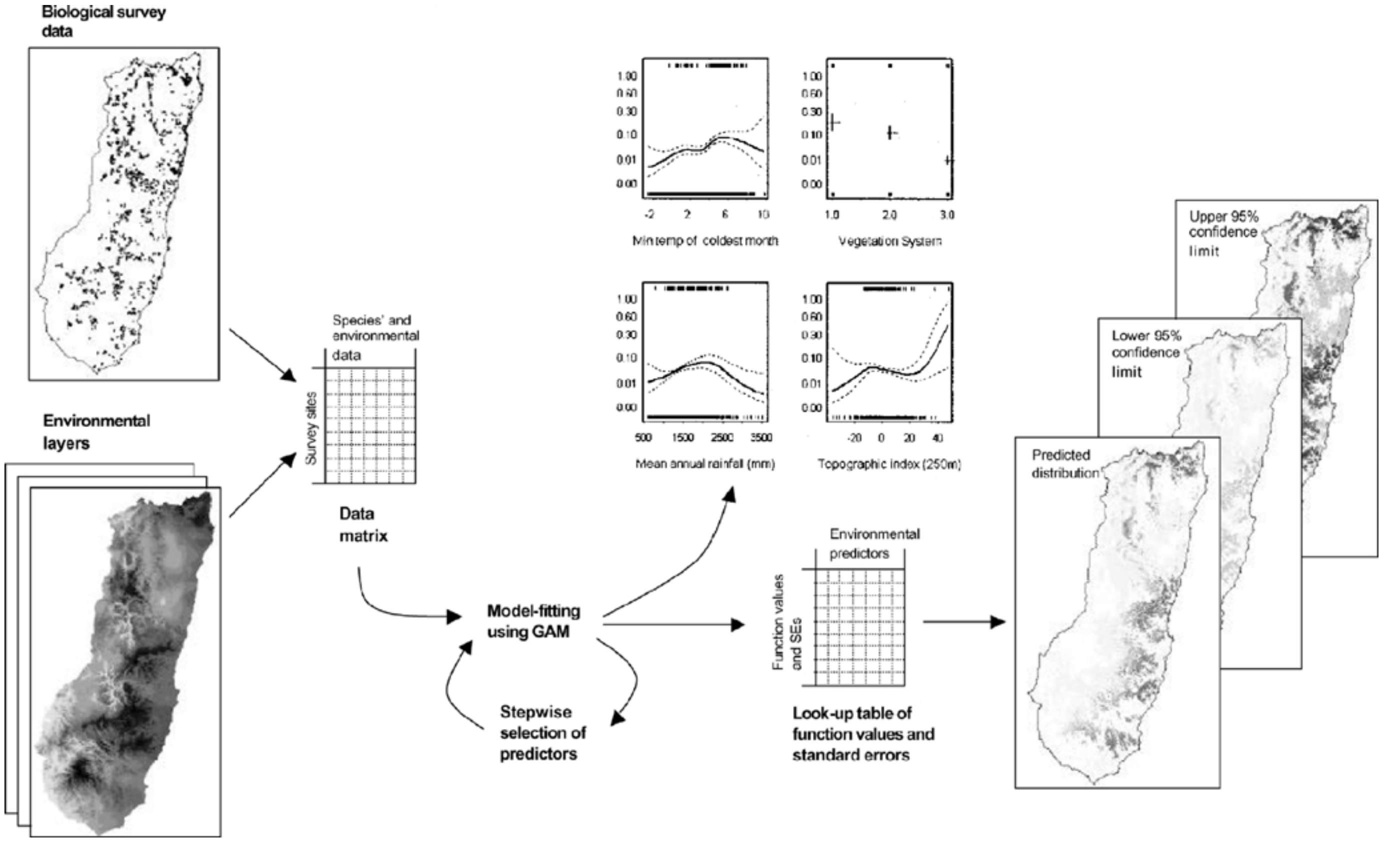

Figure 3. The basic integrated process used to model species distribution in northeast NSW.

scale to emphasise variation at the lower end of the probability range. This facilitated ready interpretation of environmental relationships of rarer species that were often indiscernible if plotted on an untransformed probability scale. The distribution of surveyed sites in relation to the predictor was indicated by two 'rugs' of tick-marks, one at the top of the graph for sites at which the species was recorded as present, and the other at the bottom of the graph for sites at which the species was recorded as absent. Each plot also depicted the upper and lower 95% confidence limits for the displayed function.

The derivation of spatial GIS layers depicting predicted probability of occurrence presented a special challenge. When models are fitted using GLM, predictions can be readily derived and mapped within a GIS because, in this case, predicted probability of occurrence is defined as a mathematical function of predictors stored as spatial layers within the GIS. However, the smooth functions fitted using GAM cannot easily be specified as mathematical formulae (Hastie and Tibshirani 1990). This means that probability of occurrence, as predicted by such a model, cannot be derived and mapped as a simple mathematical function of GIS layers representing the predictors. S-PLUS has the capability to predict the response for a model fitted using GAM, given any specified set of values for the predictors. While in theory this capability could be used to make a prediction for every grid cell in the region of interest, in practice this approach is far too slow and cumbersome for a large region containing many millions of grid cells.

An alternative approach was developed during the North East Forests Biodiversity Study that enabled much faster spatial extrapolation of species models fitted using GAM. This involved using S-PLUS to prepare a 'look-up table' of function values and associated standard errors for a discrete set of values for each predictor. For continuous predictors a specified number of values (at least 30) was sampled evenly across the range of the predictor. This look-up table was then passed to the GIS, thereby allowing the predicted probability of occurrence (and associated confidence limits) for each grid cell to be calculated by simply extracting the appropriate function values and standard errors from the table. Where the value for a predictor in a given grid cell fell between two values in the look-up table, the required function value and standard error were estimated by simple linear interpolation.

Three mapped probability surfaces were derived and stored within the GIS for each modelled species. One surface contained the predicted probability of occurrence for each and every grid cell in the region. The other two surfaces contained estimates of the upper and lower 95% confidence limits for this predicted probability. As well as portraying the magnitude of error associated with predictions in different parts of the region, the confidence limit surfaces provide important information for specific planning and management activities. For example, the lower confidence limit is of particular relevance to activities that need to identify areas where we are highly confident a species exists – e.g. selection of conservation reserves. On the other hand, the upper confidence limit is useful for activities that require confidence in the absence of a species – e.g. environmental impact assessment. The confidence limits were estimated directly from the standard errors provided by S-PLUS. While some early investigation was conducted into using bootstrapping to derive more robust confidence limits for mapped predictions (Ferrier and Watson 1994), the computation time required by this approach rendered it impractical for routine application.

The modelling process just described was largely automated, and was designed to allow large number of species to be modelled in batches. Between 1991 and 1994 the North East Forests Biodiversity Study derived distributional models for 1684 vascular plant species (both canopy and understorey species) and 713 vertebrate animal species (amphibians, reptiles, birds and mammals). Some species of particular conservation concern were later remodelled as part of the Comprehensive Regional Assessment using expanded survey datasets and refined environmental layers (NSW NPWS 1999). This remodelling also made greater use of expert knowledge in developing and refining models (see the section below on 'incorporating expert knowledge').

The integrated modelling process and software routines developed during the North East Forests Biodiversity Study were later adopted as the foundation for a more generic species modelling system (SPMODEL) developed by Environment Australia, an Australian Commonwealth Government agency (Watson 1996; Bennett et al. 1997). Through this system many of the modelling techniques pioneered in northeast NSW have now been applied more widely to conservation assessments in other parts of Australia, particularly to Comprehensive Regional Assessments in southeast NSW, Queensland, Western Australia and Victoria (e.g. Gioia and Pigott 2000). Some of the specific techniques employed originally in the northeast NSW modelling work (e.g. the use of look-up tables for spatial extrapolation) have also been incorporated into another more recently developed system for species modelling – GRASP (Lehmann et al. 2002).

Modelling 'presence-only' data

The modelling process described above was designed primarily for survey datasets in which each species of interest is recorded as either present or absent at each of a set of surveyed sites – i.e. 'presence/absence' data. This type of data is typically collected only by rigorously designed surveys such as those conducted in northeast NSW. Many regions of the world lack sufficient presence/absence data to model species distributions reliably. Even in northeast NSW, despite the extensive survey effort of the past decade, insufficient data have been generated for a number of species of particular conservation concern. These species are often very rare and/or difficult to detect and have therefore yielded few presence records in the presence/absence dataset, even after surveying many hundreds, or thousands, of sites. However, for some of these species, a larger number of presence records were collated from museum and herbarium collections, and *ad hoc* field observations. Such data are often referred to as ‘presence-only’ data because, while they indicate locations at which a species has been recorded as present, they provide no indication of the other locations that were searched unsuccessfully. Without this additional information, variation in survey effort between different environments and geographical areas cannot be readily controlled, or adjusted, for in the fitting of distributional models. Real relationships between a species and its environment can be easily confounded by spurious patterns resulting from sampling bias.

The need to model distributions of rarer species in northeast NSW often presented an interesting dilemma. Say, for example, a species was recorded at only two sites in the presence/absence survey dataset, but had been recorded at another 70 locations in museum collections. Modelling of the presence/absence data for this species would achieve little, given the small number of presence records, but might it still be worth attempting to derive a model from the presence-only data? In many regions of the world this choice is even starker, because presence/absence data are often non-existent (even for common species) and any modelling must therefore rely solely on presence-only data. The strategy adopted in northeast NSW was to employ presence-only data to model only those species for which a reasonable model could not be derived from presence/absence data alone.

In the North East Forests Biodiversity Study presence-only data were modelled using a modified version of the GAM-based approach employed for presence/absence data. This involved fitting a logistic regression model to a data frame containing all sites where a species had been recorded as present, in addition to a set of ‘pseudo-absence’ sites (Ferrier and Watson 1997; Ferrier et al. 2000d). The latter was generated by locating sites randomly across the total geographical area, or ‘domain’, of interest. Depending on the source of presence data, this domain might be variously defined as the entire study region, or as some ‘surveyable’ portion of the region – e.g. all areas of forest within a specified distance of access routes – in an attempt to address likely spatial bias in survey coverage. The pseudo-absence sites were not intended as a sample of sites at which the species was truly absent, but rather as a sample of all sites within the domain. The approach is therefore analogous to that commonly employed to analyse ‘used-vs.-available’ data in the development of resource selection models (Manly et al. 1993; Boyce et al. 2002).

A sufficiently large sample of pseudo-absence sites was generated (typically 1000 sites) to provide reasonable representation of the environmental variation exhibited by the domain. However, when GAM was used to fit a logistic regression model to the combined presence and pseudo-absence data, each pseudo-absence site was down-weighted in the analysis to emulate an equal number of presences and absences. This was achieved by using the case-weighting option provided by S-PLUS. For example, if 1000 pseudo-absence sites were being used, then each presence site would be assigned a weight of one while each pseudo-absence site would be assigned a weight of $n/1000$, where n is the number of presence sites. This weighting facilitated the estimation of approximate degrees of freedom, deviances and significance levels for the fitting of presence-only models. The weighting also enabled predictions to be expressed in terms of a relative index of likelihood of occurrence ranging from 0 to 1. These predictions indicated those parts of the region where a species is most likely to occur, but provided no estimation of the actual probability of occurrence. Zaniewski et al. (2002) have further evaluated and refined this presence-only modelling technique using data on New Zealand ferns.

Refining species models by extending the basic approach

To improve the accuracy of modelled species distributions in northeast NSW, the basic approach described above was in some cases extended to: (1) address spatial autocorrelation in the distribution of species; (2) incorporate mapped land-cover attributes, derived from interpretation of aerial photography or satellite imagery, as additional predictors alongside abiotic environmental variables; and (3) integrate expert knowledge of species distributions and habitat requirements.

Addressing spatial autocorrelation

The basic modelling process depicted in Figure 3 assumes that the probability of a species occurring in a given grid cell is determined purely by the environmental characteristics, or ‘habitat suitability’, of that cell. Two cells sharing the same values for all environmental variables will therefore be predicted to have the same probability of occurrence, regardless of where these cells are located geographically. In other words, even though predictions are mapped onto geographical space, geography is not considered directly in the fitting of models. Models are fitted entirely within environmental space. In the real world, however, species distributions may exhibit spatial pattern or ‘autocorrelation’ (Legendre 1993) that cannot be explained purely in terms of the environment or habitat at each grid cell in a region (no matter how finely and accurately this environment is measured). Incorporating consideration of spatial autocorrelation into modelling of species distributions is a challenge that has attracted increasing research interest in recent years (e.g. Augustin et al. 1996; Miller and Franklin 2002). As part of the modelling work conducted in northeast NSW, attention has been focused on two specific problems relating to spatial autocorrelation, at two quite different spatial scales.

At the coarse biogeographical scale, range limits associated with historical dispersal barriers or competition may result in a species being absent from areas of apparently suitable habitat. In northeast NSW this problem was addressed initially by simply including latitude as an additional predictor in species models, given that most range limits in this region are latitudinal. Early trials were also conducted in using a more sophisticated approach in which geographical space was incorporated into models as a two-dimensional predictor (Ferrier and Watson 1994). This was achieved by modelling the relationship between probability of occurrence and geographical location as a two-dimensional smooth surface fitted using 'loess' (Cleveland et al. 1992), a local regression smoothing technique. The principal advantage of this approach over simply employing latitude and/or longitude as one-dimensional predictors was that it allowed for more complex interactions between these variables in shaping the distribution of species. The approach is effectively a non-parametric equivalent of using polynomial regression or trend surfaces to incorporate geographical space into parametric multiple regression models (Legendre 1993).

An example of the application of the approach in northeast NSW is provided in Figure 4. Here the distribution of a species of reduced-limbed skink (*Ophioscincus*

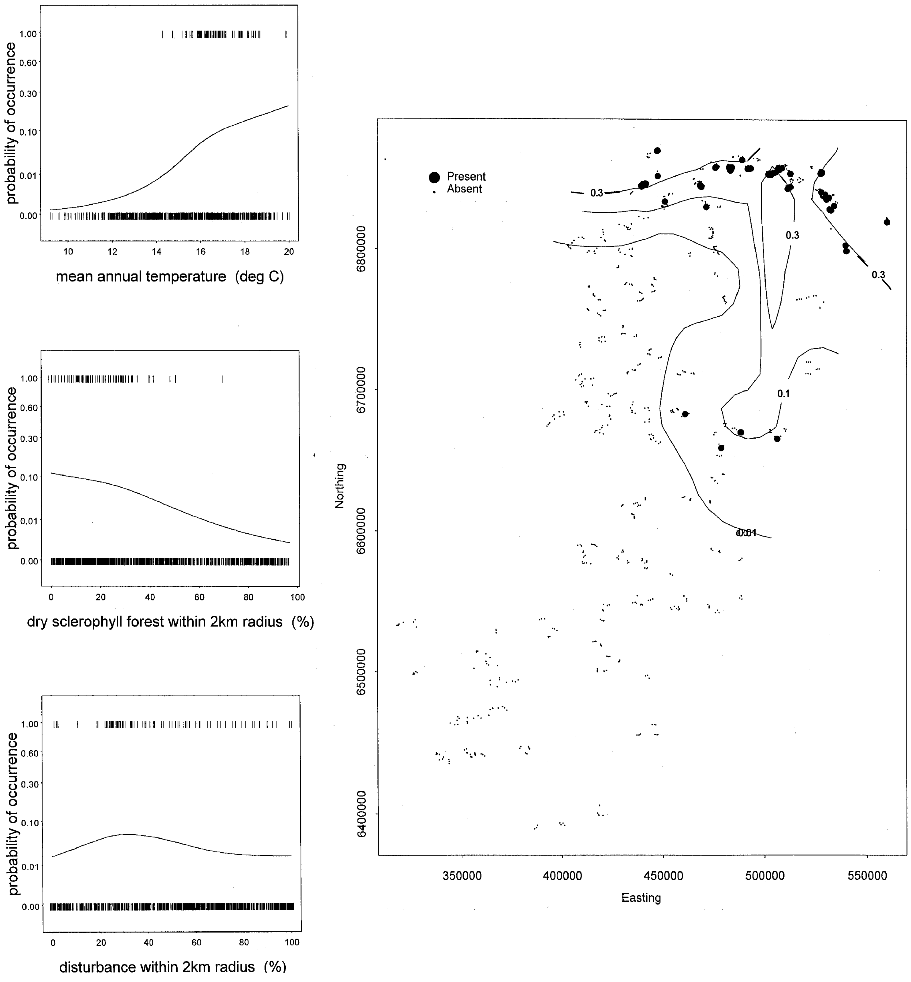

Figure 4. A GAM-based logistic regression model for the reduced-limbed skink *Ophioscincus truncatus*, employing geographical space as a two-dimensional predictor (loess surface) alongside three one-dimensional environmental predictors.

truncatus) has been modelled using GAM, employing geographical space as a two-dimensional predictor alongside three one-dimensional environmental predictors. All of the predictors are integrated within a single multivariate model – i.e. the smooth surface relating probability of occurrence to geographical location controls for the effects of the environmental variables, while the smooth functions fitted to the environmental variables control for the effects of geographical location. The model suggests that this species is confined to the northeastern portion of the study area, despite the presence of potentially suitable habitat further south. Incorporating geographical space as a two-dimensional predictor in this manner greatly increases the computation time required to fit models. The approach was therefore not applied routinely during the northeast Forests Biodiversity Study. However, it was later refined and incorporated into the SPMODEL system (Watson 1996) and was employed successfully during the Comprehensive Regional Assessment to derive refined models for a number of priority faunal species (NSW NPWS 1999). The approach offers a powerful and flexible means of addressing coarse-scaled spatial autocorrelation in modelling of species distributions, and is therefore worthy of further investigation and refinement.

At finer spatial scales, patchiness in the distribution of species within apparently suitable habitat can result from a wide range of biological and historical factors, some of which operate in a deterministic manner while others are relatively stochastic. While some of this variation may therefore be predictable, a portion will always remain as unexplained 'noise'. In northeast NSW attention was focused on modelling distributional patchiness as a function of the spatial context of individual grid cells. For many species – particularly faunal species with large home ranges (e.g. owls, marsupial gliders) – the probability of occurrence within a given grid cell is likely to depend not only on the suitability of habitat within that cell but also on the suitability and spatial configuration of habitat in neighbouring cells. According to ecological principles of metapopulation dynamics (Hanski 1999b), such species are more likely to occur in large well-connected patches of suitable habitat than in small isolated patches.

As an initial approach to factoring spatial context into faunal models for northeast NSW, several 'contextual indices' were derived from the available environmental predictors. These indices measured characteristics of the environment within a specified radius of each grid cell of interest or 'focal cell' – e.g. the proportion of cells within a 500 m radius that contain rainforest, or the mean level of disturbance within a 2000 m radius. An inverse-distance weighting was applied in calculating each index, thereby ensuring that cells close to the focal cell had a greater effect than cells further away. Contextual indices were added to the list of environmental variables considered as candidate predictors when fitting faunal models. These indices featured prominently in models for a large number of species, particularly those with large home ranges. Two of the environmental predictors included in the model depicted in Figure 4 are contextual indices.

The strategy used to address spatial context in faunal habitat modelling in north-east NSW has recently been refined through employment of autologistic regression. Our particular approach is an extension of that proposed by Augustin et al. (1996) in which a term for autocorrelation – the 'autocovariate' – is incorporated as an additional predictor in a standard logistic regression model. In Augustin et al. 's original approach the autocovariate is estimated as a weighted average of the observed or predicted occurrence of a species in all grid cells within a specified neighbourhood, where the weight applied to each grid cell is simply the inverse of the Euclidean distance from the focal cell. This approach assumes that the effect of a neighbouring cell is purely a function of distance. Yet, in reality, this effect is likely to depend not only on the distance between cells but also on the nature of intervening habitat. A neighbouring cell containing suitable habitat is likely to have less effect if it is separated from the focal cell by a barrier of unsuitable habitat than if the two cells are linked by a continuous block of suitable habitat. The weighting of neighbouring cells in the derivation of autocovariates for faunal modelling should therefore reflect connectedness (or accessibility) rather than simply distance.

In our approach the autocovariate for focal cell i is defined more generally as:

$$\text{autocov}_i = \sum_{j=1}^{k_i} \exp(-\alpha d_{ij})\hat{p}_j$$

where k_i is the number of cells within a neighbourhood (square or circle) of specified size centred on the focal cell, d_{ij} is the 'effective distance' between cell j and the focal cell, p_j is the predicted probability of occurrence in cell j, and α is a constant determining the effect of distance on isolation for the species of interest (Ferrier et al. 1999b; Drielsma and Ferrier, in preparation).

This formulation closely resembles that sometimes employed in metapopulation ecology to measure the 'habitat neighbourhood' around a specified location (Hanski 1999a,b). We estimate effective distances between cells by first assigning an 'impedence multiplier' to every grid cell in the region. This multiplier determines what contribution a cell lying on the path between two other cells will make to the effective distance between those cells. For example, if the path between two cells crosses a 100 m cell with a multiplier of 1.5 then 150 m will be added to the effective distance. The multiplier can be derived in many different ways but, to date, has usually been specified as some inverse function of predicted probability of occurrence. In other words, grid cells predicted to be of higher habitat suitability are assigned lower values for the multiplier, and therefore contribute less to effective distances, than do cells of lower habitat suitability. Once values for the multiplier have been derived, the effective distance between any given pair of cells is estimated by searching for the shortest path (in terms of effective distance) between those cells. The effective distance associated with this potentially convoluted path thereby provides a measure of connectedness between a focal cell and each of the other cells in a specified neighbourhood.

As for any autologistic modelling in which the autocovariate is based on predicted rather than observed response values, models derived using our approach must be fitted iteratively. Initially, GAM is used to fit a model without the autocovariate – i.e. using the environmental predictors alone. The mapped probability surface derived from this initial model is used to calculate values for the autocovariate at each surveyed site. The model is refitted, incorporating the autocovariate alongside the other environmental predictors, and is then used to derive a new mapped probability surface from which new values for the autocovariate are calculated. This process is repeated iteratively until convergence. An example is provided in Figure 5. We have optimised estimation of the autocovariate for GIS datasets containing very large numbers of grid cells by employing shortest-path search algorithms derived from graph theory (Drielsma and Ferrier, in preparation).

Application of this autologistic modelling approach in northeast NSW has been largely experimental to date. However, early results of testing suggest that the technique represents a substantial improvement over autologistic approaches that ignore connectivity in the estimation of autocovariates. One possible shortcoming of the approach is that expert ecological knowledge is required to estimate some of the parameters used in calculating the autocovariate – i.e. the constant α, and the function relating impedence to predicted probability of occurrence. Future attention needs to be given to the possibility of employing global optimisation techniques (e.g. simulated annealing), in conjunction with GAM, to estimate these parameters from the survey data as part of the model fitting process.

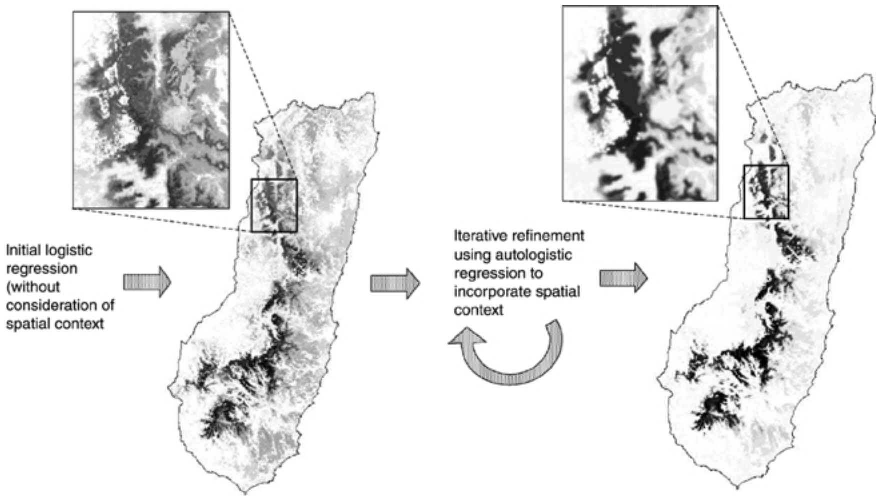
Initial logistic regression (without consideration of spatial context
Iterative refinement using autologistic regression to incorporate spatial context

Figure 5. An example of the application of autologistic regression modelling to incorporate spatial context into a distributional model for the marsupial Greater Glider *Petauroides volans*. Darker levels of grey indicate higher probabilities of occurrence for this species. Note that small isolated fragments of habitat predicted by the original model are effectively downgraded by the autologistic regression modelling.

Incorporating land-cover mapping

Statistical modelling of species distributions is often based purely on abiotic environmental predictors that describe various attributes of terrain, climate and substrate. This approach is predicated on an assumption that these variables are sufficient to explain observed biological distributions. In other words, such modelling assumes a direct deterministic relationship between species distributions and mapped abiotic environmental variables. In reality, however, this relationship may be far from perfect if environmental variables are not mapped at a sufficient level of spatial resolution and accuracy, or if key variables are not considered in the modelling. A particularly challenging example of the latter relates to the role that past disturbance may play in shaping current biological distributions. For example, in northeast NSW the current distribution of rainforest, and therefore all species associated with rainforest, has been shaped in part by relatively stochastic fire events occurring over many thousands of years. Variables describing the current abiotic environment of the region can therefore explain only a proportion of the pattern in rainforest distribution. However, rainforest can be mapped very accurately through interpretation of aerial photography or satellite imagery. Such mapping can greatly improve the accuracy with which distributions of rainforest-associated species are modelled.

Modelling of species distributions in relation to abiotic environmental variables is sometimes viewed as a competing alternative to more traditional techniques of land-cover mapping, particularly mapping of vegetation types. However, in our modelling work in northeast NSW, we have instead viewed abiotic environmental mapping and land-cover mapping as complementary sources of information that can assist in explaining and modelling biological distributions. Variables derived from land-cover mapping were incorporated as predictors in species models, alongside abiotic environmental predictors. In its simplest form this involved treating broad vegetation types mapped from satellite imagery and aerial photography (e.g. rainforest, moist eucalypt forest, dry eucalypt forest) as levels of a factor variable (Ferrier et al. 2000d).

During the Comprehensive Regional Assessment a more refined approach was developed to incorporate land-cover mapping into modelling of faunal species of special conservation concern. This involved deriving 'habitat indices' from fine-scaled (1:25 000) mapping of vegetation types and growth stages (NSW NPWS 1999; Pearce et al. 2001b). Aerial photograph interpretation had been used to divide the region's forests into 110 unique forest types, and each of these types had been further subdivided into seven growth stage classes, yielding a total of 770 potential combinations of forest type and growth stage. It was clearly impractical to incorporate this classification into statistical modelling by simply treating each mapped combination as a separate class of a factor variable. Many of the combinations contained no, or very few, faunal survey sites. Expert opinion was therefore used to amalgamate these combinations into a smaller number of classes for modelling. However, rather than creating a single amalgamated classification, the experts derived 10 different classifications, each relating to a particular habitat attribute. For example, to derive a ‘tree hollow index’ the 770 mapped combinations were amalgamated into a small number of ordered classes based on expert knowledge of the availability of tree hollows in different forest types and growth stages. The other indices (e.g. nectar index, structural complexity index) were derived by amalgamating the classes in different ways.

Integrating expert knowledge

Locational records derived from formal or *ad hoc* surveys are only one potential source of information on the distribution of species. Experts familiar with the fauna and flora of a region often possess a well-developed knowledge of species distributions and habitat requirements. This knowledge is usually acquired from many years of opportunistic or incidental field observation and may be difficult to translate into a discrete set of locational records. Ideally, survey data and expert knowledge should be viewed as complementary, rather than alternative or competing, information sources. In practice, however, the integration of expert knowledge into statistical modelling of species distributions presents many challenges.

All of the modelling work performed in northeast NSW was conducted in close collaboration with teams of ecologists familiar with the region’s flora and fauna. Expert knowledge was incorporated at several stages in the development, interpretation and application of statistical models (NSW NPWS 1999; Ferrier et al. 2000d; Pearce et al. 2001b). The role played by experts in deriving habitat indices for modelling faunal distributions has already been discussed above in the section on ‘incorporating land-cover mapping’. Initially experts were used to critically review the biological survey datasets to identify, and where possible correct, erroneous or anomalous records. Experts also assisted in selecting relevant environmental predictors to be used in modelling each biological group (e.g. understorey plants, reptiles).

Once a statistical model had been fitted to the available data for a given species, expert input was again sought to check, interpret and, where appropriate, refine or modify predictions from the model. Expert refinement was restricted to models for species of special conservation concern, fitted during the Comprehensive Regional Assessment. This refinement usually involved some form of GIS-based manipulation to correct for over-prediction in parts of the region outside the known range of a species, or in habitat types known to be unsuitable for the species. If the predicted distribution of a species produced by statistical modelling appeared particularly anomalous then, in some cases, an alternative GIS-based model was developed based on expert knowledge alone – e.g. ‘species *x* occurs in areas of tall eucalypt forest, with a mean annual rainfall between 1500 and 2000 mm, and a soil fertility index greater than 3’.

Expert opinion also played a crucial role in the final stage of preparing modelled species distributions for use in prioritising and selecting conservation areas in northeast NSW. Population viability analysis had been employed to set reservation targets for species of special conservation concern (Environment Australia 1999). However, these targets were usually specified in terms of some measure of total abundance – e.g. number of breeding females – whereas the distributional models derived for each species predicted only the probability, or relative likelihood, of occurrence. Expert opinion was therefore used to convert probabilities of occurrence into densities (NSW NPWS 1999). This was achieved by first dividing the range of predicted probabilities for each species into four classes corresponding to four levels of habitat quality – core, intermediate, marginal and unsuitable. Each of these habitat classes was then assigned a density or 'carrying capacity' – e.g. number of breeding females per km^2. This conversion, albeit approximate, enabled areas being considered for reservation to be assessed in terms of their potential contribution to population viability targets.

Evaluating performance of species modelling

In parallel with the development and application of species models in northeast NSW, considerable effort was devoted to evaluating the performance of such modelling. This evaluation work provided planners with important information on the level of uncertainty associated with predictions derived from species models, and therefore the appropriate level of precaution to be exercised when employing these predictions in regional conservation planning and decision-making. The evaluation work also assessed the relative performance of a wide range of alternative modelling techniques and strategies, and therefore provided information that may inform the selection of appropriate techniques for use in other regions. Performance of species modelling was evaluated in terms of both: (1) the accuracy with which a model for a given species predicts the actual distribution of that species, and (2) the effectiveness with which species models for a given biological group (e.g. trees, birds) function as a general surrogate for spatial pattern in biodiversity within that and other groups. The general approach adopted in the former of these evaluations resembled that employed in many other studies of the performance of species models (e.g. Flather and King 1992; Edwards et al. 1996; Fielding and Bell 1997; Beard et al. 1999; Elith 2000), while the approach adopted in the latter evaluation appears to be unique.

Evaluating predictive accuracy of species models

The predictive accuracy of species models was evaluated by comparing actual observations of occurrence, at a set of surveyed sites, with predicted probabilities (or relative likelihoods) of occurrence generated by a model. Wherever possible, models were evaluated using independent data obtained from sites other than those used to develop the model. Northeast NSW provided a good test-bed for this type of evaluation because of its relatively long and staged history of biological data collection. Models developed at a given point in this history could be evaluated using data collected during later surveys. In many cases these later surveys were purposely designed so that sites were located well away from previously surveyed sites, thereby ensuring spatial independence between the development and evaluation datasets. If no independent dataset was available for a particular evaluation then statistical resampling was applied to the development dataset to reduce bias in the assessment of performance. This was generally achieved by jackknifing (or 'leave one out' cross-validation), in which each site is withheld in turn from the development dataset, and a model is fitted to the remaining sites. The prediction obtained from this model for the withheld site is then compared to the actual observation for that site. This procedure is repeated for each site, thereby producing a set of 'independent' predictions and observations for all sites in the development dataset. We note, however, that for any future evaluations the independence of predictions and observations might be better achieved by applying cross-validation to coarser subdivisions of the dataset (e.g. 5–10 groups; T.J. Hastie, personal communication).

The evaluation work conducted in northeast NSW was based on a conceptual framework proposed by Murphy and Winkler (1992) for assessing predictions from probabilistic models, and adapted for application to species modelling by Pearce and Ferrier (2000a). This framework identifies two major components of performance: (1) discrimination capacity – the ability of a model to correctly distinguish between occupied and unoccupied sites, and (2) calibration – the agreement between predicted probabilities of occurrence and observed proportions of sites occupied. The discrimination capacity of models developed for northeast NSW was assessed in terms of the area under a relative operating characteristic (ROC) curve relating proportions of correctly and incorrectly classified predictions over a continuous range of probability thresholds. The calibration of models was assessed using a technique originally proposed by Cox (1958) and later refined by Miller et al. (1991), in which logistic regression is used to analyse the relationship between predicted probability of occurrence and observed proportion of sites occupied, and to quantify individual components of calibration error. The discrimination capacity and calibration of models was also assessed graphically using a series of three plots, as illustrated in Figure 6. A detailed description of the overall conceptual framework, and the specific techniques used to assess discrimination capacity and calibration, is provided by Pearce and Ferrier (2000a).

Results of the various evaluation studies conducted using data from northeast NSW have been reported in detail elsewhere (Ferrier and Watson 1997; Pearce and Ferrier 2000a,b, 2001; Pearce et al. 2001a,b). Major findings of this work were: (1) models developed for northeast NSW exhibited a reasonably high level of predictive accuracy. When models for a representative sample of 153 species of vascular plants and vertebrates were evaluated using independent survey data, 89% of the models performed significantly better than random in terms of discrimination capacity, while

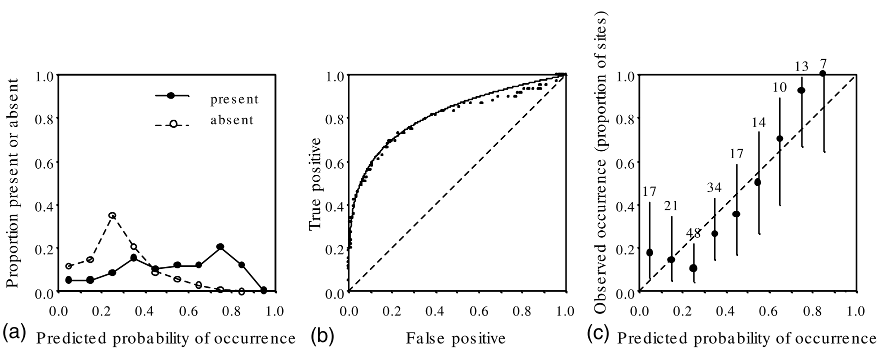

Figure 6. An example (for the skink *Calyptotis scutirostrum*) of three diagnostic plots used to assess the predictive accuracy of species models, employing independent data. (a) The frequency distribution of predicted probabilities for occupied and unoccupied evaluation sites, providing an indication of discrimination capacity. (b) The ROC curve depicting the relationship between false positive and true positive fractions as the probability threshold is varied between 0 and 1, thereby providing another representation of discrimination capacity. (c) The relationship between predicted probability of occurrence and the observed proportion of evaluation sites occupied, for 10 0.1 probability intervals between 0 and 1, providing an indication of model calibration.

70% achieved high levels of discrimination – i.e. an ROC area greater than 0.7 (Pearce et al. 2001a). (2) Models derived from presence/absence data performed better than those derived from presence-only data for the same species (Ferrier and Watson 1997). (3) Discrimination capacity exhibited a positive correlation with the incidence of species in the dataset (i.e. the proportion of sites at which each species was recorded) (Pearce and Ferrier 2000b). (4) Discrimination capacity was also positively correlated with the total number of sites (presence and absence) used to derive a model (Pearce and Ferrier 2000b). (5) Models derived using fine-scaled environmental data performed better than models derived using coarse-scaled data. (6) In comparisons of alternative modelling techniques, logistic regression models fitted using GAM generally out-performed those fitted using GLM (Ferrier and Watson 1997; Pearce and Ferrier 2000b), which in turn performed better on average than decision-tree models and models based on simple profile matching (e.g. BIOCLIM) (Ferrier and Watson 1997). (7) In a further comparison of GAM-based logistic regression modelling with techniques for modelling relative abundance rather than presence/absence (GLM- and GAM-based Poisson regression and zero-inflated negative binomial regression), predictions from the latter techniques performed no better as a relative index of abundance than predicted probabilities of occurrence generated by logistic regression (Pearce and Ferrier 2001). (8) The strategy used in northeast NSW to select predictors for inclusion in species models produced models with better discrimination capacity than models produced by alternative selection strategies (e.g. forward–backward selection of variables) (Pearce and Ferrier 2000b). (9) The discrimination capacity of faunal models was generally improved by incorporating predictors describing spatial context (Pearce et al. 2001b). (10) The incorporation of expert opinion into statistical modelling of faunal distributions resulted in small but statistically insignificant gains in predictive accuracy (although see Pearce et al. (2001b) for a discussion of caveats applying to this particular study).

Evaluating species modelling as a biodiversity surrogate

The performance of species modelling in northeast NSW was also evaluated as part of a more general research project that assessed the effectiveness with which various forms of environmental mapping and modelling function as surrogates for spatial pattern in biodiversity as a whole (Ferrier and Watson 1997). The focus here was not on how well modelling can predict distributions of individual species of interest (i.e. fine-filter species), but rather on how well modelling of multiple species within a biological group might perform as a general (coarse-filter) surrogate for spatial pattern in that and other groups. This question is of fundamental relevance to any attempt to employ species modelling as a primary basis for regional conservation planning. By selecting conservation areas to maximise representation of a set of modelled species, to what extent are we in turn maximising representation of biodiversity as a whole?

The biological and environmental datasets assembled for northeast NSW were used to compare the performance of species modelling with a wide range of other surrogate mapping approaches, including various types of vegetation mapping, abiotic environmental classification and ordination, and canonical ordination. These surrogates were evaluated by treating biological survey sites as candidate areas for conservation, and selecting sites in the order that maximised representation of diversity within a given surrogate (without any reference to the actual biological survey data for those sites). The site selected at each step was that which provided the greatest improvement in representation of diversity within the surrogate. In the case of species modelling, representation was measured by predicting the number of modelled species occurring in at least one of the selected sites. This number was estimated from predicted probabilities of occurrence for each of the modelled species using standard probability theory:

$$\text{Predicted no. of species represented} = \sum_{i=1}^{\text{species}} \left[1 - \prod_{j=1}^{\text{sites}} (1 - p_{ij}) \right]$$

where p_{ij} is the predicted probability of modelled species i occurring at selected site j.

This approach to measuring representation based on modelled probabilities of species occurrence has also been proposed independently by two more recent studies (Polasky et al. 2000; Williams and Araújo 2000).

Once sites had been selected according to a particular surrogate, the biological survey data for these sites were then used to assess the number of species actually represented (i.e. included in the hypothetical set of conservation areas) after each selection. This procedure generated a species accumulation curve describing the relationship between the cumulative number of sites selected (*X* axis) and the cumulative number of species represented (*Y* axis). A 'species accumulation index' was then derived by scaling the area under the accumulation curve, obtained using the surrogate, in relation to areas under two other accumulation curves (see example in Figure 7): (1) a 'mean random curve' estimated by averaging a large number of individual random curves, each derived by selecting sites in random order without reference to either the surrogate or the biological survey data, and (2) an 'optimum curve' derived by selecting sites using the actual biological survey data in place of the surrogate (i.e. at each step selecting the site that most improved the number of species represented). The index can range from 1 (one) for a perfect surrogate down to 0 (zero) or less for

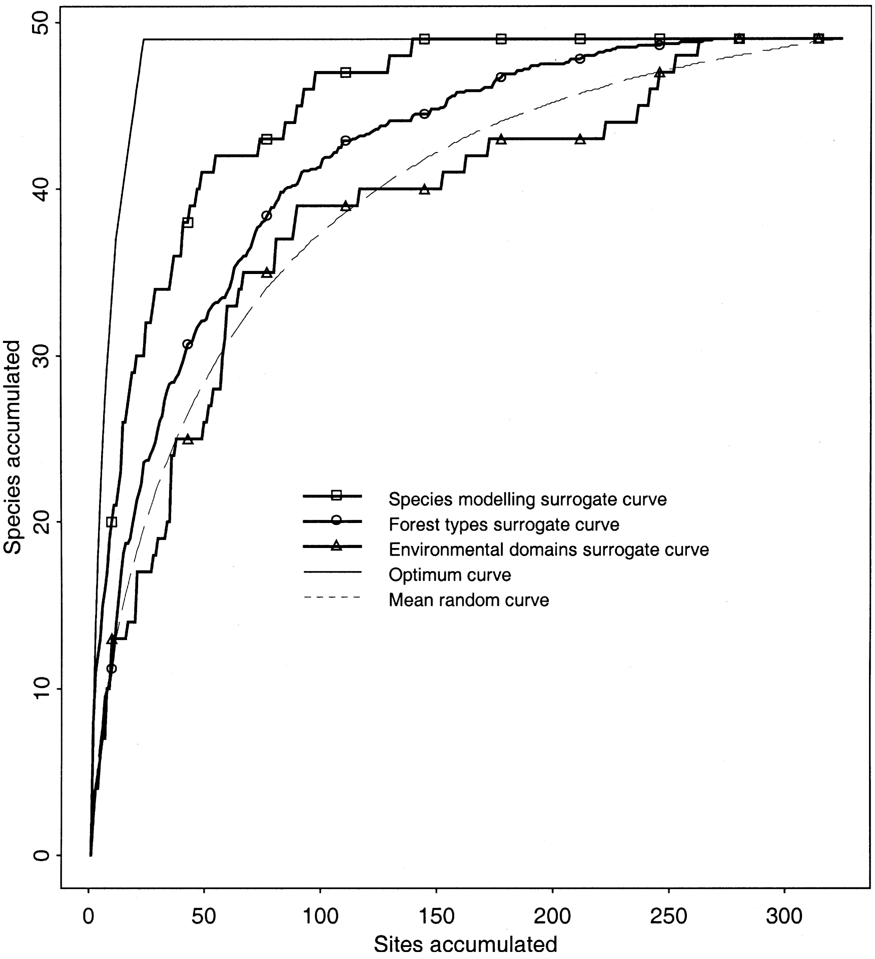

Figure 7. An example (for reptiles) of species accumulation curves derived to evaluate the performance of alternative biodiversity surrogates.

a surrogate that performs no better than a random selection of sites. Bootstrapping of the site data was used to estimate confidence limits for observed values of the index, and to test the statistical significance of differences in performance between surrogates. A more detailed explanation of this evaluation technique is provided by Ferrier and Watson (1997) and Ferrier (2002).

In the northeast NSW study, surrogates were evaluated using survey data for 10 biological groups: ants, beetles, spiders, reptiles, birds, bats, rainforest canopy trees, rainforest understorey plants, open-forest canopy trees and open-forest understorey plants. Independence between the biological data used to derive some of the surrogates (including species modelling) and the data used to evaluate those surrogates was achieved by randomly splitting the survey sites into two sets of equal size – a model-development set and an evaluation set. The study evaluated numerous combinations of surrogates and biological groups, and presentation of full results is therefore beyond the scope of this paper (interested readers should access Ferrier and Watson 1997). Selected results are summarised graphically in Figure 8. The six surrogates for which results are presented are: (1) 'species models (within group)', in which GAM was used to derive models for all species within a given biological group (e.g. reptiles) and these models were then evaluated using actual biological survey data for the same group; (2) 'modelled canopy trees', in which modelling of canopy tree species was evaluated as a surrogate for each of the other biological groups; (3) 'forest type mapping', an existing fine-scaled (1:25 000) vegetation map derived from

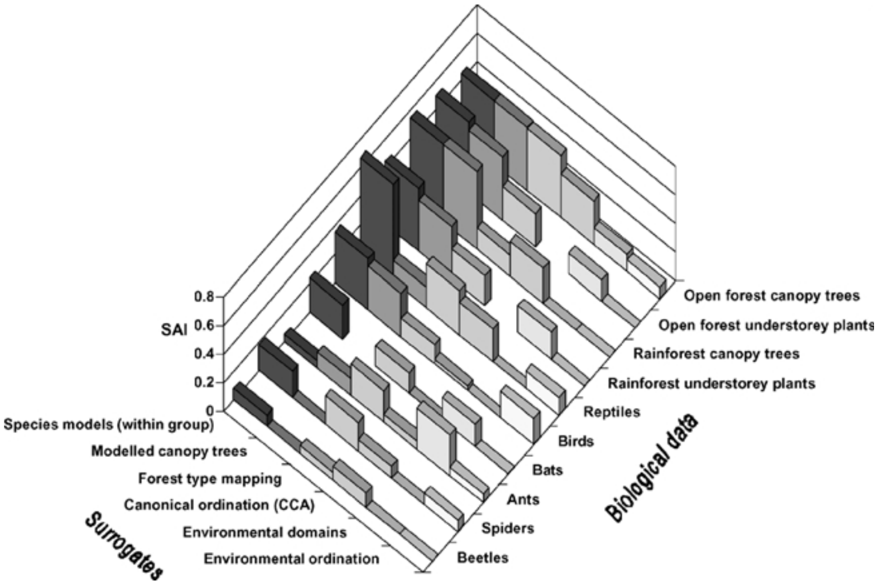

Figure 8. A summary of the performance of six surrogate approaches, evaluated using independent survey data for 10 biological groups in northeast NSW. SAI – Species Accumulation Index. The absence of a bar for a given combination of surrogate and biological group indicates that this combination was not evaluated.

aerial photograph interpretation, and containing 130 types; (4) 'canonical ordination (CCA)' (ter Braak 1986), in which biological survey data for the group of interest was modelled in relation to environmental variables using canonical correspondence analysis, with four axes; (5) 'environmental domains', in which numerical classification was used to group sites into 125 environmental classes (domains) based purely on abiotic environmental variables; and (6) 'environmental ordination' in which these same environmental variables were used to derive a hybrid multidimensional-scaling ordination (Belbin 1991), with four axes.

Three trends are worth noting in Figure 8. First, species modelling within each group of interest achieved the best overall performance of the evaluated surrogates. Second, modelling of canopy trees performed reasonably well as a surrogate for understorey plants and, to a lesser extent, vertebrates. Finally, the ground-dwelling arthropod groups (ants, spiders and beetles) were, in general, not served well by any of the surrogate approaches, including species modelling. Possible explanations for this last result are explored by Ferrier et al. (1999a, 2002). The overall results of the evaluation work are also discussed further in Ferrier and Watson (1997) and Ferrier (2002).

Conclusions

Statistical modelling of species distributions has made an enormous contribution to regional conservation planning in northeast NSW during the past 20 years, culminating in extensive additions to the region's protected area system. Evaluation of the predictive accuracy of a subset of models has suggested that such modelling can provide a reasonably reliable basis for planning at a regional scale. Further evaluation has also indicated that species modelling within selected biological groups can perform well as a surrogate for biodiversity as a whole, compared with other surrogates commonly employed in conservation planning. Ongoing refinement of the models for northeast NSW is likely to be achieved by incorporating additional biological data, and by refining the analytical methodology to better address spatial autocorrelation and better integrate land-cover attributes derived from fine-scaled remote-sensing.

The modelling approaches developed in this study are potentially applicable to other regions with reasonable coverage of biological and environmental data. However, we recognise that the level of investment in data collection and analysis in northeast NSW is unusually high relative to that for a large proportion of the world's regions, particularly those identified as global priorities for conservation action. Modelling approaches developed or applied in relatively data-rich regions may not necessarily work effectively elsewhere. There is an urgent need to adapt existing approaches, and perhaps develop new approaches, to better cope with sparse and biased datasets in regions encompassing high levels of diversity. These approaches may need to further extend the capability of existing techniques to accommodate presence-only data and to integrate expert opinion into model fitting. Relatively data-rich regions such as northeast NSW can play an important role in such development by serving as test-beds for evaluating the performance of alternative modelling approaches against independent biological datasets.

To date, most research and development work on statistical modelling of spatial pattern in biodiversity has focused on modelling of individual species. There is now a large literature describing, and debating the relative merits of, different statistical techniques for modelling species distributions. While modelling of individual species clearly has an important role to play in conservation planning, there are many situations in which this approach might be more appropriately supplemented or replaced by modelling of higher-level biodiversity entities (e.g. communities or assemblages) or collective properties of biodiversity (e.g. species richness or compositional dissimilarity). Application of these alternative approaches needs to be supported by research that looks beyond formulating and evaluating relatively minor variations in statistical methodology for modelling species distributions, to focus more broadly on formulating and evaluating other potential strategies for modelling spatial pattern in biodiversity as a whole. The second paper of this series (Ferrier et al. 2002) describes various such approaches to community level modelling that have been developed and applied in northeast NSW.

Acknowledgements

The collection, entry and checking of the biological data used to model species distributions in northeast NSW was made possible only through the dedication and sheer hard work of over 100 individuals spread across several State and Commonwealth funded projects. Particular assistance in deriving and interpreting models was provided by Saskia Raaphorst, Harry Hines, Dianne Brown, Jill Smith, Sandy Gilmore, David Scotts, John Hunter and Stephen Griffith. While the development of our modelling approaches benefited greatly from numerous stimulating discussions with many individuals, the early support and encouragement provided by Mike Austin deserves special mention. Glenn Manion assisted in preparing a number of the figures. Some of the work on evaluating the performance of models was funded by Environment Australia.

References

Anderson R.S. and Ashe J.S. 2000. Leaf litter inhabiting beetles as surrogates for establishing priorities for conservation of selected tropical montane cloud forests in Honduras, Central America (Coleoptera; Staphylinidae, Curculionidae). Biodiversity and Conservation 9: 617–653.

Anonymous 1992. National Forest Policy Statement: a New Focus for Australia's Forests. Australian Government Publishing Service, Canberra, Australia.

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

Austin M.P. 1998. An ecological perspective on biodiversity investigations: examples from Australian eucalypt forests. Annals of the Missouri Botanical Garden 85: 2–17.

Austin M.P. and Meyers J.A. 1996. Current approaches to modeling the environmental niche of eucalypts: implications for management of biodiversity. Forest Ecology and Management 85: 95–106.

Austin M.P., Cunningham R.B. and Fleming P.M. 1984. New approaches to direct gradient analysis using environmental scalars and statistical curve-fitting procedures. Vegetatio 55: 11–27.

Beard K.H., Hengartner N. and Skelly D.K. 1999. Effectiveness of predicting breeding bird distributions using probabilistic models. Conservation Biology 13: 1108–1116.

Belbin L. 1991. Semi-strong hybrid scaling, a new ordination algorithm. Journal of Vegetation Science 2: 491–496.

Bennett S., Watson G. and Barratt D. 1997. Species Distribution Modelling Toolkit (SPMODEL). User's Manual. Environment Forest Group, Environment Australia, Canberra, Australia.

Bio A.M.F., Alkemade R. and Barendregt A. 1998. Determining alternative models for vegetation response analysis: a non-parametric approach. Journal of Vegetation Science 9: 5–16.

Boyce M.S., Vernier P.R., Nielsen S.E. and Schmiegelow F.K.A. 2002. Evaluating resource selection functions. Ecological Modelling 157: 279–298.

Brieman L., Friedman J.H., Olshen R.A. and Stone C.J. 1984. Classification and Regression Trees. Wadsworth International Group, Belmont, California.

Caro T.M. and O'Doherty G. 1999. On the use of surrogate species in conservation biology. Conservation Biology 13: 805–814.

Cleveland W.S., Grosse E. and Shyu W.M. 1992. Local regression models. In: Chambers J.M. and Hastie T.J. (eds) Statistical Models in S. Wadsworth and Brooks, Pacific Grove, California, pp. 309–376.

Commonwealth of Australia 1997. Nationally Agreed Criteria for the Establishment of a Comprehensive, Adequate and Representative Reserve System for Forest in Australia. A Report by the Joint ANZECC/MCFFA National Forest Policy Statement Implementation Sub-committee. Commonwealth of Australia, Canberra, Australia.

Cowling R.M., Pressey R.L., Lombard A.T., Desmet P.G. and Ellis A.G. 1999. From representation to persistence: requirements for a sustainable system of conservation areas in the species-rich mediterranean-climate desert of southern Africa. Diversity and Distributions 5: 51–71.

Cox D.R. 1958. Two further applications of a model for binary regression. Biometrika 45: 562–565.

Dinerstein E., Powell G., Olson D., Wikramanayake E., Abell R., Loucks C., Underwood E., Allnutt T., Wettengel W., Ricketts T., Strand H., O'Connor S. and Burgess N. 2000. A Workbook for Conducting Biological Assessments and Developing Biodiversity Visions for Ecoregion-Based Conservation. Part 1: Terrestrial Ecoregions. Conservation Science Program, World Wildlife Fund, Washington, DC.

Edwards Jr, T.C., Deshler E.T., Foster D. and Moisen G.G. 1996. Adequacy of wildlife habitat relation models for estimating spatial distributions of terrestrial vertebrates. Conservation Biology 10: 263–270.

Elith J. 2000. Quantitative methods for modeling species habitat: comparative performance and an application to Australian plants. In: Ferson S. and Burgman M. (eds) Quantitative Methods for Conservation Biology. Springer-Verlag, New York, pp. 39–58.

Environment Australia 1999. Response to Disturbance of Forest Species, Upper North East and Lower North East Regions. A Project Undertaken As Part of the NSW Comprehensive Regional Assessments. Resource and Conservation Division, Department of Urban Affairs and Planning, Sydney, Australia.

Ferrier S. 1984. The Status of the Rufous Scrub-Bird *Atrichornis Rufescens*: Habitat, Geographical Variation and Abundance. Ph.D. Thesis. University of New England, Armidale, Australia.

Ferrier S. 1991. Computer-based spatial extension of forest fauna survey data: current issues, problems and directions. In: Lunney D. (ed) Conservation of Australia's Forest Fauna. Royal Zoological Society of NSW, Sydney, Australia, pp. 221–227.

Ferrier S. 1992a. Development of a Predictive Modelling Module for E-RMS. Unpublished consultancy report prepared by NSW National Parks and Wildlife Service. Australian National Parks and Wildlife Service, Canberra, Australia.

Ferrier S. 1992b. Environmental Resource Mapping System. User's Manual. New South Wales National Parks and Wildlife Service, Sydney, Australia.

Ferrier S. 1997. Biodiversity data for reserve selection: making best use of incomplete information. In: Pigram J.J. and Sundell R.C. (eds) National Parks and Protected Areas: Selection, Delimitation and

Management. Centre for Water Policy Research, University of New England, Armidale, Australia, pp. 315–329.

Ferrier S. 2000. Applications and directions post-1995. In: Brown D., Hines H., Ferrier S. and McKay K. (eds) Establishment of a Biological Information Base for Regional Conservation Planning in Northeast New South Wales, Phase 1 (1991–1995). Occasional Paper no. 26, New South Wales National Parks and Wildlife Service, Sydney, Australia, pp. 151–154.

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

Ferrier S. and Smith A.P. 1990. Using geographical information systems for biological survey design, analysis and extrapolation. Australian Biologist 3: 105–116.

Ferrier S. and Watson G. 1994. Modelling the spatial distribution of forest fauna and flora: approaches to reducing and evaluating prediction error. In: Anonymous (ed) Proceedings of International Biodiversity Conference: Conserving Biological Diversity in Temperate Forest Ecosystems – Towards Sustainable Management. Centre for Resource and Environmental Studies, Australian National University, Canberra, Australia, pp. 81–82.

Ferrier S. and Watson G. 1997. An Evaluation of the Effectiveness of Environmental Surrogates and Modelling Techniques in Predicting the Distribution of Biological Diversity. Environment Australia, Canberra, Australia.

Ferrier S., Gray M.R., Cassis G.A. and Wilkie L. 1999a. Spatial turnover in species composition of ground-dwelling arthropods, vertebrates and vascular plants in northeast New South Wales: implications for selection of forest reserves. In: Ponder W. and Lunney D. (eds) The Other 99%. The Conservation and Biodiversity of Invertebrates. Royal Zoological Society of New South Wales, Sydney, Australia, pp. 68–76.

Ferrier S., Pearce J., Drielsma M., Watson G., Manion G., Whish G. and Raaphorst S. 1999b. Evaluation and Refinement of Techniques for Modelling Distributions of Species, Communities and Assemblages. Unpublished report prepared for Environment Australia. New South Wales National Parks and Wildlife Service, Sydney, Australia.

Ferrier S., Brown D. and Hines H. 2000a. Environmental GIS database. In: Brown D., Hines H., Ferrier S. and McKay K. (eds) Establishment of a Biological Information Base for Regional Conservation Planning in Northeast New South Wales, Phase 1 (1991–1995). Occasional Paper no. 26, New South Wales National Parks and Wildlife Service, Sydney, Australia, pp. 33–76.

Ferrier S., Brown D., Hines H., Scotts D., Griffiths S. and Hunter J. 2000b. Introduction. In: Brown D., Hines H., Ferrier S. and McKay K. (eds) Establishment of a Biological Information Base for Regional Conservation Planning in Northeast New South Wales, Phase 1 (1991–1995). Occasional Paper no. 26, New South Wales National Parks and Wildlife Service, Sydney, Australia, pp. 15–28.

Ferrier S., Pressey R.L. and Barrett T.W. 2000c. A new predictor of the irreplaceability of areas for achieving a conservation goal, its application to real-world planning, and a research agenda for further refinement. Biological Conservation 93: 303–325.

Ferrier S., Watson G., Hines H. and Brown D. 2000d. Predictive modelling of biological data. In: Brown D., Hines H., Ferrier S. and McKay K. (eds) Establishment of a Biological Information Base for Regional Conservation Planning in Northeast New South Wales, Phase 1 (1991–1995). Occasional Paper no. 26, New South Wales National Parks and Wildlife Service, Sydney, Australia, pp. 97–130.

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

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

Finkel E. 1998. Ecology: software helps Australia manage forest debate. Science 281: 1789–1791.

Flather C.H. and King R.M. 1992. Evaluating performance of regional wildlife habitat models: implications to resource planning. Journal of Environmental Management 34: 31–46.

Franklin J. 1995. Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. Progress in Physical Geography 19: 474–499.

Franklin J. 1998. Predicting the distribution of shrub species in southern California from climate and terrain-derived variables. Journal of Vegetation Science 9: 733–748.

Gioia P. and Pigott J.P. 2000. Biodiversity assessment: a case study in predicting richness from the potential distributions of plant species in the forests of southwestern Australia. Journal of Biogeography 27: 1065–1078.

Groves C., Valutis L., Vosick D., Neely B., Wheaton K., Touval J. and Runnels B. 2000. Designing a Geography of Hope: a Practitioner's Handbook for Ecoregional Conservation Planning, 2nd ed. The Nature Conservancy, Washington, DC.

Guisan A. and Zimmermann N.E. 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135: 147–186.

Hanski I. 1999a. Habitat connectivity, habitat continuity, and metapopulations in dynamic landscapes. Oikos 87: 209–219.

Hanski I. 1999b. Metapopulation Ecology. Oxford University Press, New York.

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

Hines H., Brown D. and Scotts D. 2000. Biological surveys. In: Brown D., Hines H., Ferrier S. and McKay K. (eds) Establishment of a Biological Information Base for Regional Conservation Planning in Northeast New South Wales, Phase 1 (1991–1995). Occasional Paper no. 26, New South Wales National Parks and Wildlife Service, Sydney, Australia, pp. 77–96.

Humphries C.J., Williams P.H. and Vane-Wright R.I. 1995. Measuring biodiversity value for conservation. Annual Review of Ecology and Systematics 26: 93–111.

Hutchinson M.F., Belbin L., Nicholls A.O., Nix H.A., McMahon J.P. and Ord K.D. 1997. BioRap Rapid Assessment of Biodiversity, Vol 2. Spatial Modelling Tools. Australian BioRap Consortium, CSIRO, Canberra, Australia.

Kremen C. 1994. Biological inventory using target taxa: a case study of the butterflies of Madagascar. Ecological Applications 4: 407–422.

Lambeck R.J. 1997. Focal species: a multi-species umbrella for nature conservation. Conservation Biology 11: 849–856.

Lawes M.J. and Piper S.E. 1998. There is less to binary maps than meets the eye: the use of species distribution data in the Southern African Sub-region. South African Journal of Science 94: 207–210.

Leathwick J.R. 1995. Climatic relationships of some New Zealand forest tree species. Journal of Vegetation Science 6: 237–248.

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

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

Lehmann A., Overton J.McC. and Leathwick J.R. 2002. GRASP: generalized regression analysis and spatial predictions. Ecological Modelling 157: 187–205.

Maddock A. and du Plessis M.A. 1999. Can species data only be appropriately used to conserve biodiversity? Biodiversity and Conservation 8: 603–615.

Manly B.J.F., McDonald L.L. and Thomas D.L. 1993. Resource Selection by Animals: Statistical Design and Analysis for Field Studies. Chapman & Hall, London.

Margules C.R. and Austin M.P. 1994. Biological models for monitoring species decline: the construction and use of data bases. Proceedings of the Royal Society, London, B 344: 69–75.

Margules C.R. and Pressey R.L. 2000. Systematic conservation planning. Nature 405: 243–253.

Margules C.R. and Redhead T.D. 1995. Guidelines for Using the BioRap Methodology and Tools. CSIRO, Canberra, Australia.

McCullagh P. and Nelder J.A. 1989. Generalized Linear Models, 2nd ed. Chapman & Hall, London.

Miller J. and Franklin J. 2002. Predictive vegetation modelling with spatial dependence: vegetation alliances in the Mojave Desert. Ecological Modelling 157: 225–245.

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

Moore D.M., Lees B.G. and Davey S.M. 1991. A new method for predicting vegetation distributions using decision tree analysis in a geographic information system. Environmental Management 15: 59–71.

Murphy A.H. and Winkler R.L. 1992. Diagnostic verification of probability forecasts. International Journal of Forecasting 7: 435–455.

Myers N., Mittermeier R.A., Mittermeier C.G., da Fonseca G.A.B. and Kent J. 2000. Biodiversity hotspots for conservation priorities. Nature 403: 853–858.

Noss R.F. 1987. From plant communities to landscapes in conservation inventories: a look at the Nature Conservancy (USA). Biological Conservation 41: 11–37.

Noss R.F. 1990. Indicators for monitoring biodiversity: a hierarchical approach. Conservation Biology 4: 355–364.

NSW NPWS 1999. Modelling Areas of Habitat Significance for Vertebrate Fauna and Vascular Flora in North East NSW. A Project Undertaken As Part of the NSW Comprehensive Regional Assessments. Resource and Conservation Division, Department of Urban Affairs and Planning, Sydney, Australia.

Olson D.M. and Dinerstein E. 1998. The Global 200: a representation approach to conserving the Earth's most biologically valuable ecoregions. Conservation Biology 12: 502–515.

Pearce J. and Ferrier S. 2000a. Evaluating the predictive performance of habitat models developed using logistic regression. Ecological Modelling 133: 225–245.

Pearce J. and Ferrier S. 2000b. An evaluation of alternative algorithms for fitting species distribution models using logistic regression. Ecological Modelling 128: 127–147.

Pearce J. and Ferrier S. 2001. The practical value of modelling relative abundance of species for regional conservation planning. Biological Conservation 98: 33–43.

Pearce J., Ferrier S. and Scotts D. 2001a. An evaluation of the predictive performance of distributional models for flora and fauna in northeast NSW. Journal of Environmental Management 62: 171–184.

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

Polasky S., Camm J.D., Solow A.R., Csuti B., White D. and Ding R. 2000. Choosing reserve networks with incomplete species information. Biological Conservation 94: 1–10.

Pressey R.L. 1998. Algorithms, politics and timber: an example of the role of science in a public, political negotiation process over new conservation areas in production forests. In: Wills R. and Hobbs R. (eds) Ecology for Everyone: Communicating Ecology to Scientists. Surrey Beatty and Sons, Sydney, Australia, pp. 73–87.

Pressey R.L. 1999. Applications of irreplaceability analysis to planning and management problems. Parks 9(1): 42–51.

Pressey R.L., Humphries C.J., Margules C.R., Vane-Wright R.I. and Williams P.H. 1993. Beyond opportunism: key principles for systematic reserve selection. Trends in Ecology and Evolution 8: 124–128.

Simberloff D. 1998. Flagships, umbrellas, and keystones: is single-species management passé in the landscape era? Biological Conservation 83: 247–257.

Smith T.B., Bruford M.W. and Wayne R.K. 1993. The preservation of process: the missing element of conservation programs. Biodiversity Letters 1: 164–167.

Soberón J.M., Llorente J.B. and Oñate L. 2000. The use of specimen-label databases for conservation purposes: an example using Mexican Papilionid and Pierid butterflies. Biodiversity and Conservation 9: 1441–1466.

Stockwell D.R.B., Davey S.M., Davis J.R. and Noble I.R. 1990. Using induction of decision trees to predict Greater Glider density. AI Applications 4(4): 33–43.

ter Braak C.J.F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67: 1167–1179.

Thackway R. and Creswell I.D. 1997. A bioregional framework for planning the national system of protected areas in Australia. Natural Areas Journal 17: 241–247.

Vane-Wright R.I. 1996. Identifying priorities for the conservation of biodiversity: systematic biological criteria within a socio-political framework. In: Gaston K.J. (ed) Biodiversity: a Biology of Numbers and Difference. Blackwell, Oxford, UK, pp. 309–341.

Watson G. 1996. Predictive Species Modelling: User Manual. Unpublished report. Environment Australia, Canberra, Australia.

Williams P.H. and Araújo M.B. 2000. Using probability of persistence to identify important areas for biodiversity conservation. Proceedings of the Royal Society, London, B 267: 1959–1966.

Yee T.W. and Mitchell N.D. 1991. Generalized additive models in plant ecology. Journal of Vegetation Science 2: 587–602.

Zaniewski A.E., Lehmann A. and Overton J.McC. 2002. Predicting species distribution using presence-only data: a case study of native New Zealand ferns. Ecological Modelling 157: 259–278.