



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

SIMON FERRIER^{1,*}, MICHAEL DRIELSMA¹, GLENN MANION¹
and GRAHAM WATSON^{1,2}

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, University of New England, P.O. Box 245, Armidale, NSW 2351, Australia; *Author for correspondence (e-mail: simon.ferrier@npws.nsw.gov.au)*

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

Abstract. Regional conservation planning can often make more effective use of sparse biological data by linking these data to remotely mapped environmental variables through statistical modelling. While modelling distributions of individual species is the best known and most widely used approach to such modelling, there are many situations in which more information can be extracted from available data by supplementing, or replacing, species-level modelling with modelling of communities or assemblages. This paper provides an overview of approaches to community-level modelling employed in a series of major land-use planning processes in the northeast New South Wales region of Australia, and evaluates how well communities and assemblages derived using these techniques function as surrogates in regional conservation planning. We also outline three new directions that may enhance the effectiveness of community-level modelling by: (1) more closely integrating modelling with traditional ecological mapping (e.g. vegetation mapping); (2) more tightly linking numerical classification and spatial modelling through application of canonical classification techniques; and (3) enhancing the applicability of modelling to data-poor regions through employment of a new technique for modelling spatial pattern in compositional dissimilarity.

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

Introduction

Social and economic factors often place severe limits on the total area of land that can be set aside, or otherwise managed, for conservation of biodiversity within any given region. Care must therefore be taken to direct conservation effort to those parts of a region that are assessed as being of highest conservation priority – a process referred to here as ‘regional conservation planning’. The past two decades have seen increasing interest in the application of ‘systematic’ approaches to such planning (Margules and Pressey 2000). The aim of these approaches is to design a system of conservation areas that is representative of the diversity encompassed by a region, and to configure these areas to promote long-term persistence of the elements of biodiversity they contain.

Systematic conservation planning requires information on the spatial distribution of biodiversity. Yet such information is often grossly incomplete, particularly at the species and genetic levels. Planning therefore commonly employs those entities (e.g. species) for which we do have reasonable distributional information as 'surrogates' for biodiversity as a whole (Noss 1990). Even for these surrogate species, or groups of species, distributional data may be confined to records obtained at a small set of survey or collection sites separated by extensive areas of unsurveyed land (Maddock and du Plessis 1999). Statistical modelling of biological survey data in relation to remotely mapped environmental variables – e.g. terrain, climate, substrate or land-cover variables – provides a powerful means of filling geographical gaps in the coverage of data. When coupled with geographical information system (GIS) technology, such modelling enables biological distributions to be extrapolated across large regions, thereby providing geographically complete information for conservation planning and other environmental applications (see reviews by Franklin 1995; Austin 1998; Guisan and Zimmermann 2000).

Modelling distributions of individual species is the best known and most widely used approach to modelling and extrapolating relationships between biological and environmental data. Much less attention has been given in the conservation planning literature to other potential modelling strategies. This is unfortunate because, while modelling of individual species performs well in some situations, there are other situations in which it may work less effectively than alternative modelling approaches (Ferrier 2002). An individual species can be modelled successfully only if sufficient locational data (or expert knowledge of habitat requirements) are available for that species. Modelling of individual species is also a relatively time-consuming and expensive process. For these reasons, use of modelled distributions in regional conservation planning is often restricted to a relatively small number of fine-filter (*sensu* Noss 1987) or priority species – e.g. selected vertebrate or vascular plant species of special conservation concern. Even if sufficient data and resources are available to model a larger number of species, it may be difficult for planning and decision-making processes to cope with the quantity and complexity of information generated by such analyses. While modelling individual species can provide a good basis for considering the particular needs of selected species in conservation planning, it may not be the most effective way of addressing spatial pattern in biodiversity as a whole. The latter might often be addressed more effectively by integrating spatial modelling with numerical classification techniques designed to analyse patterns within large, complex datasets.

Such integration allows better use to be made of all available data for all surveyed species, not just selected priority species. The approach also alleviates potential problems associated with considering a large number of individual species in conservation planning and decision-making, by reducing this complexity to a much smaller set of higher-level entities – i.e. communities or assemblages. The principal disadvantage of modelling communities or assemblages instead of species is that the approach

may not allow planning to give due consideration to the needs of individual species of particular conservation concern – e.g. threatened, focal or flagship species. The distributions of such species may still warrant modelling on an individual basis, allowing them to be considered as discrete entities in planning, alongside modelled distributions of communities or assemblages. By further combining these two modelling approaches with more traditional data types (e.g. vegetation mapping), the coarse-filter/fine-filter strategy (Noss 1987) is effectively extended to make maximum use of all available information. Biodiversity as a whole is addressed using some combination of traditional land-class mapping and modelled distributions of communities or assemblages, while individual species of particular concern are addressed using a combination of locational records and modelled species distributions.

This paper provides an overview of community-level modelling work conducted in the northeast New South Wales (NSW) region of Australia. It is the second in a two-paper series describing approaches to deriving models for use in conservation planning within this region. The first paper (Ferrier et al. 2002) focused on species-level modelling. Both species-level and community-level modelling have played a pivotal role in a series of government-led planning processes in northeast NSW during the past 6 years, resulting in major additions to the region's protected area system. The extensive biological and environmental datasets established for northeast NSW have also served as a test-bed for research on the performance of alternative modelling approaches.

The paper starts by describing four basic strategies for integrating numerical classification and statistical modelling to predict distributions of communities or assemblages. We provide examples of how these strategies have been applied in northeast NSW, and summarise the results of a recent study evaluating the effectiveness with which communities or assemblages derived using these techniques function as surrogates in regional conservation planning. We then outline three new directions in community-level modelling that may enhance the effectiveness of the general approach, in northeast NSW and elsewhere, by: (1) more closely integrating modelling with traditional ecological mapping (e.g. vegetation mapping); (2) more tightly linking numerical classification and spatial modelling through application of canonical classification techniques; and (3) enhancing the applicability of modelling to data-poor regions through application of a new technique for modelling spatial pattern in compositional dissimilarity.

Basic approaches to community-level modelling

Imagine a region in which one or more components of biodiversity (e.g. vascular plants, birds, beetles) have been surveyed at a number of sites. At each site the presence or absence (or relative abundance) of species has been recorded, thereby generating a sites-by-species matrix. One possible strategy for analysing such data can be

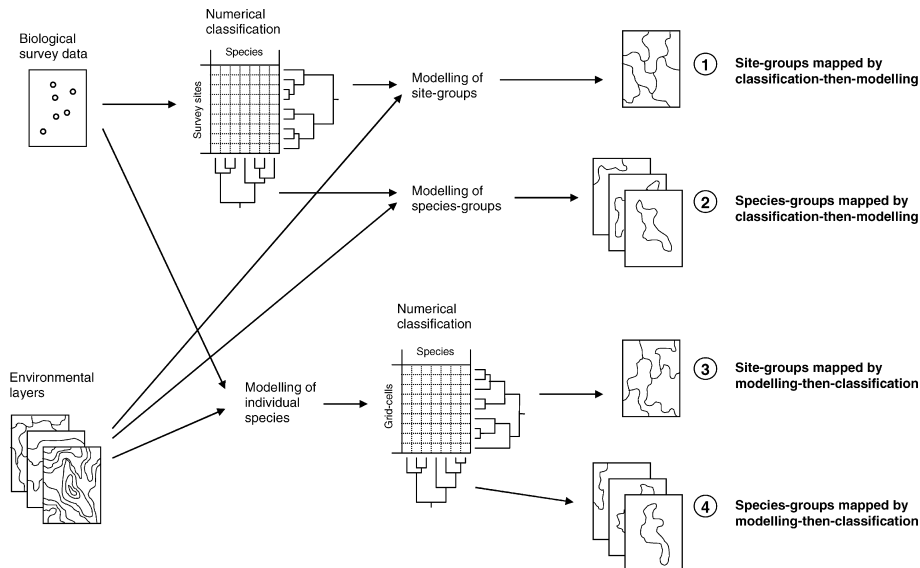


Figure 1. Major strategies for integrating numerical classification and modelling of communities or assemblages.

referred to as ‘classification-then-modelling’. In this strategy the data matrix is first subjected to numerical classification to derive either groups of sites containing similar species, or groups of species occurring at similar sites (Figure 1). The terminology used to describe such groups is notoriously inconsistent. In conservation planning applications within Australia, groups of sites derived from this type of analysis are often referred to as ‘communities’, while groups of species are referred to as ‘species assemblages’. However, given the inconsistency in usage of such terms within the wider scientific community, we will here refer explicitly to the two types of groups as ‘site-groups’ and ‘species-groups’.

Site-groups generated by numerical classification of survey data can be modelled and extrapolated using techniques similar to those employed for modelling species distributions (Ferrier et al. 2002). For example, generalised linear modelling (GLM, McCullagh and Nelder 1989) or generalised additive modelling (GAM, Hastie and Tibshirani 1990; Yee and Nelder 1989) can be used to model the probability of each site-group occurring at a given location, as a function of mapped environmental variables (including, where available, land-cover attributes or classes mapped from aerial photography or satellite imagery). Distributions of these site-groups are then extrapolated across the entire region of interest. Each location (e.g. grid cell) in the region is assigned to the site-group predicted to have the highest probability of occurrence at that location. Alternatively, decision-tree modelling (Brieman et al. 1984; Moore et al. 1991) can be used to generate discrete environmental rules for allocating grid cells to site-groups. Regardless of the particular modelling technique, maps derived

by modelling numerically classified site-groups can be employed in regional conservation planning in the same manner as any other land classification (e.g. a traditional vegetation map).

Species-groups generated by numerical classification of survey data can also be modelled and extrapolated across an entire region of interest. In this case, the modelled response for each species-group is the proportion of species (belonging to the group) that occur, or are predicted to occur, at a given location. For example, if an assemblage contains eight species then a modelled value of 0.5 for a location means that four of the eight species are predicted to occur at that location. Unlike site-groups, members of two or more species-group can occur at the same location. Modelling of species-groups therefore generates a series of mapped layers (one for each group), that can be employed in regional conservation planning in the same manner as modelled distributions of individual species. Examples of previous applications of the classification-then-modelling strategy include Moore et al. (1991) and Keith and Bedward (1999) for modelling of site-groups, and McKenzie et al. (1989) for modelling of species-groups.

A second strategy for modelling distributions of communities or assemblages can be referred to as 'modelling-then-classification'. The first step in this strategy is to model and extrapolate the distributions of individual species included in the sites-by-species matrix. This generates a mapped layer for each species indicating the predicted probability of that species occurring within each and every grid cell in the region. By restructuring these predictions as a large cells-by-species matrix they can be subjected to numerical classification, thereby deriving either groups of grid cells with similar predicted species or groups of species with similar predicted distributions (Figure 1). This strategy generates maps of the same form as those produced by the classification-then-modelling strategy. The robustness of maps of site-groups or species-groups derived by modelling-then-classification clearly depends on the robustness of the underlying species models. Examples of previous applications of the modelling-then-classification strategy include Lenihan (1993), Austin (1998) and Cawsey et al. (2002) for modelling of site-groups, and Nix (1991) and Lehmann et al. (2002) for modelling of species-groups.

Applications in northeast NSW

The northeast NSW study region was described in the first paper of this series (Ferrier et al. 2002). Much of the modelling work to date has focused on the eastern section of the region – an area of 79 000 km² that encompasses most of the region's taller eucalypt forest and rainforest (see Figure 1 in Ferrier et al. 2002). The NSW National Parks and Wildlife Service established an environmental GIS database for northeast NSW in the late 1980s, and this has since undergone continual refinement. The database contains a wide range of mapped and modelled layers pertaining to topography, climate, substrate, vegetation cover and disturbance, most of which are

stored at a 1 ha (100 m × 100 m) grid-cell resolution (Ferrier et al. 2000a, 2002). Extensive field surveys of vascular flora, vertebrate fauna and selected components of invertebrate fauna (mainly ground-dwelling arthropods) were conducted during the early 1990s, with survey sites stratified to sample the environmental variation of the region (Hines et al. 2000; Ferrier et al. 2002). By 1994, distributional models had been derived for 1684 vascular plant species and 713 vertebrate animal species (Ferrier et al. 2000c, 2002). These were GAM-based logistic regression models fitted using the S-PLUS statistical package (MathSoft). Between 1995 and 1998, models for individual species of conservation concern were combined with community-level modelling to provide information for a series of government-led planning processes designed to resolve a long-running conflict between the needs of commercial forest harvesting and the protection of biodiversity, old growth and wilderness values (Ferrier et al. 2002). The last of these processes – the Comprehensive Regional Assessment (1996–1998) – resulted in extensive additions to the region's reserve system.

Both of the community-level modelling strategies described in the preceding section – classification-then-modelling and modelling-then-classification – were applied in various ways to the biological and environmental data for northeast NSW, generating maps for both site-groups and species-groups. Particular attention was given to developing automated software tools for implementing the modelling-then-classification strategy (Ferrier et al. 1999b). This was achieved by linking ArcView to the pattern analysis package PATN (Belbin 1995). Special-purpose scripts developed in ArcView constructed a cells-by-species matrix from a specified set of modelled species distributions and passed this matrix to PATN. If the region of interest contained too many grid cells for efficient processing in PATN then a random sample of cells was extracted. The data matrix was then subjected to numerical classification by calling appropriate PATN functions from ArcView. The parameters required to run these functions (e.g. the number of site-groups or species-groups to be generated) were specified using dialog boxes within ArcView. Relevant outputs from PATN (i.e. the membership of each derived group, and the dendrogram describing the hierarchical relationship between these groups) were automatically imported back into ArcView. This information was then used to map the spatial distribution of each group identified by PATN. In the case of site-group classifications a single grid layer was generated in which each grid cell in the region was assigned to a single group. If a grid cell had not been included in the sample of cells analysed within PATN then the compositional dissimilarity between this cell and each group centroid was used to assign the cell to its nearest group. In the case of species-group classifications a separate grid layer was generated for each group. Each of these layers was derived by averaging the modelled probability surfaces for all species included in the relevant species-group.

Two examples of the application of this approach to modelling-then-classification in northeast NSW are presented in Figures 2 and 3. In the first example (Figure 2)

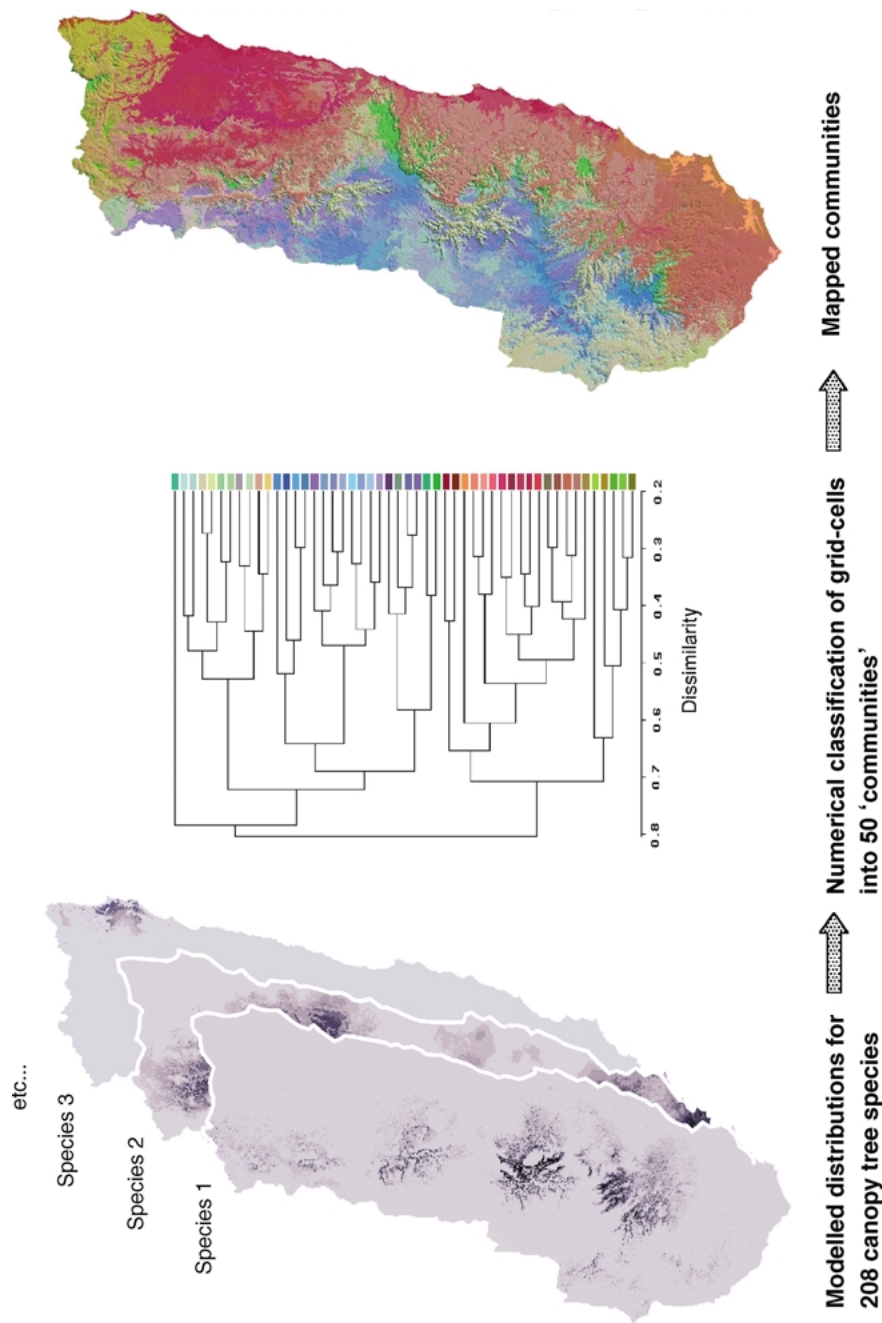


Figure 2. An example application of the modelling-then-classification strategy to derive and map site-groups (or 'communities') from modelled distributions of canopy tree species in northeast NSW.

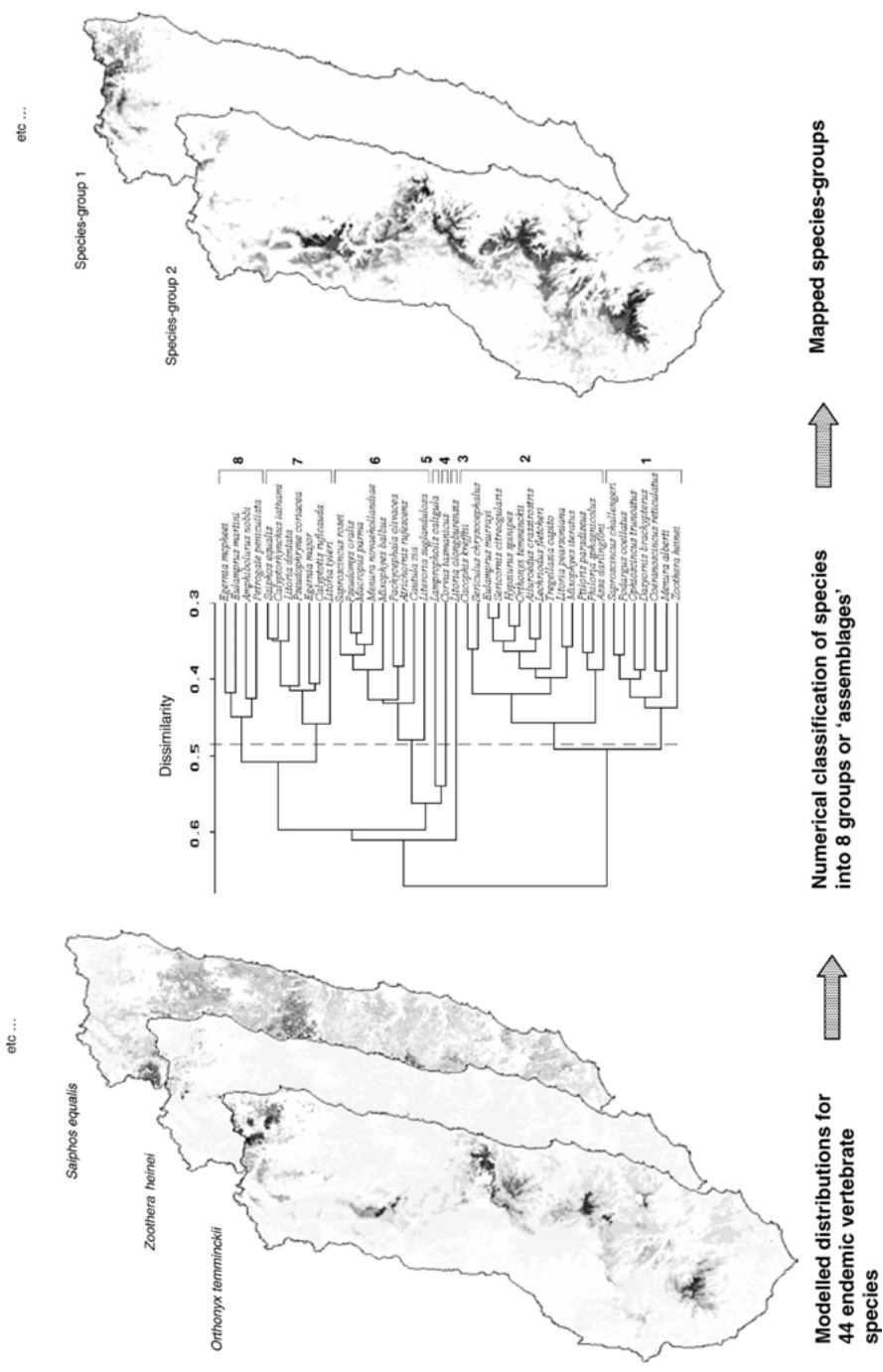


Figure 3. An example application of the modelling-then-classification strategy to derive and map species-groups (or 'assemblages') from modelled distributions of endemic vertebrate species in northeast NSW.

modelled distributions for 208 individual species of canopy trees were used to derive and map 50 site-groups (or 'communities'). The original analysis on which this example is based mapped 198 groups, but these have been amalgamated to a 50-group level for the purpose of the example. The classification was performed using the ALOB non-hierarchical clustering algorithm within PATN (Bray–Curtis dissimilarity measure), with the derived groups then subjected to hierarchical clustering to produce the dendrogram. This particular classification played a vital role in an Interim Forest Assessment conducted in 1995 and 1996, by generating estimates of the original (i.e. pre-clearing) extent of vegetation communities within northeast NSW. These estimates provided the primary basis for setting and evaluating vegetation protection targets.

In the second example (Figure 3) modelled distributions for 44 species of vertebrates endemic to northeast NSW were used to derive and map eight species-groups (or 'assemblages'). This classification was performed using a hierarchical clustering algorithm based on the Two-Step dissimilarity measure (Austin and Belbin 1982). This particular classification played an important role in the Comprehensive Regional Assessment, where it was used to identify centres of endemism in northeast NSW requiring special protection. Other classifications of vertebrate and vascular plant assemblages have been employed more recently to help identify key habitats and corridors within the region, as a basis for regional vegetation planning across all land tenures within the region.

Evaluating modelled communities and assemblages as surrogates for biodiversity

In the preceding sections we have described four basic strategies for modelling distributions of communities or assemblages: (1) classification-then-modelling of site-groups, (2) classification-then-modelling of species-groups, (3) modelling-then-classification of site-groups, and (4) modelling-then-classification of species-groups. All of these approaches have been, or could be, used to derive coarse-filter biodiversity surrogates for use in regional conservation planning. To help provide some guidance as to the relative efficacy of these alternatives, an earlier study evaluating the performance of individual species models as surrogates for biodiversity (Ferrier and Watson 1997; Ferrier et al. 2002) was more recently extended to evaluate approaches to community-level modelling (Ferrier et al. 1999b).

This extended study employed the same 10 biological datasets as those used in the original evaluation study – ants, beetles, spiders, reptiles, birds, bats, rainforest canopy trees, rainforest understorey plants, open-forest canopy trees, and open-forest understorey plants (Ferrier and Watson 1997; Ferrier et al. 2002). The sites for each biological group were again split randomly into a development sample and an evaluation sample. Four new types of surrogates were derived for each biological dataset: (1) site-groups derived by classification (Bray–Curtis dissimilarity measure and UPGMA clustering) then modelling (GAM-based logistic regression modelling

of each site-group, followed by assignment of each and every grid cell in the region to the group with the highest predicted probability of occurring in that cell). Site-group maps were derived at four levels of classification resolution: 5, 15, 50 and 150 groups. (2) Species-groups derived by classification (Two-Step dissimilarity measure and UPGMA clustering) then modelling (GAM-based logistic regression modelling of the proportion of species from a given species-group occurring at each site, with a separate model produced for each group). Species-group maps were derived at three levels of classification resolution: 5, 10 and 20 groups. (3) Site-groups derived by modelling of individual species (GAM-based logistic regression) then classification of grid cells into 5, 15, 50 and 150 groups. (4) Species-groups derived by modelling of individual species then classification of these species into 5, 10 and 20 species groups.

As in the original evaluation study (Ferrier and Watson 1997; Ferrier et al. 2002), each of these surrogates was evaluated by selecting sites in the order that maximised representation of diversity within the surrogate, and then using the actual survey data for the sites to derive a species accumulation curve. For site-group surrogates the site selected at each step was drawn from the group that had the lowest proportion of its sites already selected. If two or more sites met this criterion then one of these was selected at random. For species-group surrogates, sites were selected using the approach described by Ferrier et al. (2002) for individual species modelling, except in this case predicted probabilities of occurrence for individual species were replaced by predicted proportions of species within each species-group. A 'species accumulation index' for each combination of surrogate and evaluation dataset was derived by scaling the area under the accumulation curve (obtained using the surrogate) in relation to two other curves – a 'mean random curve' estimated by averaging a large number of individual random curves, each derived by selecting sites in random order, and an 'optimum curve' derived by selecting sites using the actual biological data in place of the surrogate. Bootstrapping was used to estimate confidence limits for each observed value of the index. A more detailed explanation of this evaluation technique is provided by Ferrier and Watson (1997) and Ferrier (2002).

Presentation of full evaluation results for all combinations of surrogate approaches and biological groups is beyond the scope of this paper (interested readers should access Ferrier et al. 1999b). Selected results are summarised graphically in Figure 4. This figure depicts results for the best-performing surrogate within each combination of surrogate type (e.g. site-groups derived by classification-then-modelling) and biological group. In other words, while several different versions (i.e. different classification resolutions) of each surrogate type were evaluated, only the result for the best-performing version is shown in Figure 4. The clearest trend exhibited by these results is that modelling of species-groups consistently performed better than modelling of site-groups as a basis for selecting conservation areas. Furthermore, the performance of species-group modelling was generally not far below that of individual spe-

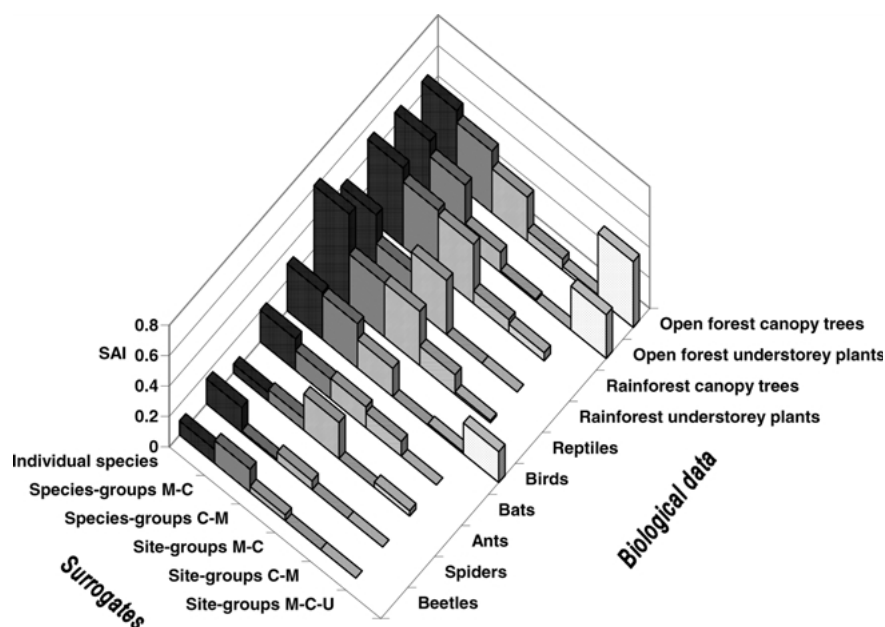


Figure 4. A summary of the performance of approaches to mapping communities or assemblages by integrating numerical classification and modelling. The approaches were evaluated as potential surrogates for biodiversity, using independent survey data for 10 biological groups in northeast NSW. SAI – Species Accumulation Index; M-C – modelling-then-classification; C-M – classification-then-modelling; M-C-U – modelling-then-classification, but incorporating information on compositional dissimilarity between groups into the evaluation. The absence of a bar for a given combination of surrogate and biological group indicates that this combination was not evaluated.

cies modelling. The order in which classification and modelling were conducted (i.e. classification-then-modelling vs. modelling-then-classification) had a less consistent effect on performance.

The poor performance of site-group surrogates in this study may reflect inadequacies in the site selection algorithm employed in the evaluation, rather than inadequacies in the surrogates themselves. The algorithm treated the site-groups within each classification as discrete entities. No information on the relationships (i.e. dissimilarities in biological composition) between these groups was employed in selecting sites. While this approach reflects that commonly used in real-world conservation planning, there is clearly room for improvement. Faith and Walker (1996a) and Woinarski et al. (1996) have demonstrated the importance of considering information on compositional dissimilarities between site-groups when prioritising areas for conservation. As an experiment we re-evaluated the performance of the modelling-then-classification site-group approach for three biological groups (birds, open-forest canopy trees and open-forest understorey plants) using an alternative site selection algorithm based on that advocated by Faith and Walker (1996a). This algorithm uses information on compositional dissimilarities between site-groups to select sites in the order that maximises

representation of overall diversity. As depicted in Figure 4, this approach greatly improved the performance of site-group modelling, yielding results roughly equivalent to those of species-group modelling. This type of consideration of relationships between site-groups or species-groups deserves greater attention in conservation planning. There is also potential to extend the measures of dissimilarity employed in such analyses to incorporate information on the taxonomic or phylogenetic relationships of the species involved (Faith 1996; Clarke and Warwick 1998).

This evaluation of approaches to modelling communities and assemblages in northeast NSW has only scratched the surface of a very large and complex issue. More work of this kind in other regions is required if we are to better understand the relative strengths and weaknesses of possible modelling approaches.

New directions in community-level modelling

The community-level modelling strategies discussed so far, and depicted in Figure 1, are those most commonly employed in regional conservation planning. However, they are not the only possible strategies for community-level modelling, nor are they necessarily the most appropriate strategies in all situations. We believe that there is still considerable room for improvement in the development of approaches to community-level modelling, and in the application of these approaches to regional planning. Our research group is currently investigating a number of new directions in community-level modelling that we feel show particular promise. We outline three of these developments below.

Integrating statistical modelling with traditional ecological mapping

Derivation of mapped communities through statistical modelling may be viewed as an alternative to more traditional forms of ecological mapping. In the latter, land-classes such as 'vegetation types', 'ecosystems' or 'habitats' are often mapped by interpretation of aerial photography or satellite imagery, supplemented by varying levels of ground-truthing (Alexander and Millington 2000). Statistical modelling and traditional mapping approaches each have strengths and weaknesses. A major strength of the statistical modelling approach is that both the classification of communities, and the linking of these communities to remotely mapped variables, are performed in a quantitative, explicit and repeatable manner. However, a potential weakness in the approach may be an over-reliance on abiotic environmental predictors (terrain, climate, substrate) to model and extrapolate distributions of communities. As in the case of species-level modelling, discussed in the first paper of this series (Ferrier et al. 2002), predictions based purely on abiotic variables may be unreliable if these variables are not mapped at a sufficient level of spatial resolution and accuracy, or if key variables are not considered in the modelling. Particular problems may arise

if the current distribution of communities is a function not only of current environmental conditions, but also of past conditions or disturbance patterns. For example, the distribution of rainforest in northeast NSW can be only partly predicted by current environmental conditions, because it has also been shaped by relatively stochastic fire events occurring over many thousands of years (Ferrier et al. 2002). Yet the distribution of rainforest in this region can be mapped accurately by interpretation of aerial photography or satellite imagery.

As in our work on species-level modelling (Ferrier et al. 2002) we view abiotic environmental variables, and attributes or classes discerned from aerial photography or satellite imagery, as complementary rather than competing sources of information for use in mapping community distributions. Aerial photography and satellite imagery may reveal patterns of community distribution that cannot be predicted using abiotic variables alone. On the other hand these same abiotic variables may help to delineate (or separate) communities that cannot be reliably discriminated using only aerial photography or satellite imagery. The real challenge is to find more effective ways of integrating all available sources of information – biological survey data, abiotic environmental variables, aerial photography, satellite imagery – to map communities within any given region. One means of achieving this is to incorporate attributes or classes derived from aerial photography or satellite imagery as additional predictors, alongside abiotic environmental variables, in statistical modelling of community distributions. These additional predictors can be employed in both the modelling-then-classification and classification-then-modelling strategies described earlier in this paper (they can also be employed in the canonical strategy described in the next section). Where the additional predictors are satellite image spectral bands (or some transformation of these bands) then the resultant modelling process can also be viewed, from an image processing perspective, as an extended form of supervised image classification (Lees and Ritman 1991; Franklin 1995).

In many situations, particularly when employing information extracted from aerial photography rather than satellite imagery, the additional predictors will not be continuous quantitative variables but will instead consist of nominal-scaled categories – e.g. vegetation types or photo-pattern classes. If the number of such categories is reasonably small then they may be simply treated as classes of a factor variable in statistical modelling. This approach has been employed extensively in northeast NSW as a means of incorporating broad vegetation types mapped from satellite imagery and aerial photography (e.g. rainforest, moist eucalypt forest, dry eucalypt forest) into both species-level and community-level modelling (Ferrier et al. 2000c, 2002). However, such an approach is rendered impracticable if the number of mapped classes is increased to a point where many of the classes contain few, or no, survey sites. In this situation the modelling may fail to detect real associations between communities and mapped classes, and will therefore exhibit a high Type II error-rate and low statistical power.

As we have done for species-level modelling (Ferrier et al. 2002), we are currently investigating an approach to using expert opinion to more effectively integrate complex land classifications (containing many classes) into modelling of community distributions. The approach is based on the classification-then-modelling strategy described earlier in this paper. The existing land classification is excluded from the initial modelling – i.e. this modelling is performed using abiotic environmental variables alone. It is assumed that, as a result of the initial modelling, each community will be predicted to have a certain probability of occurring in each grid cell within the region of interest. In the original classification-then-modelling approach each grid cell would then be assigned to the community for which this predicted probability of occurrence is highest. In the extended approach, however, the assignment of grid cells to communities is further constrained (or conditioned) by expert opinion as to the likelihood of each community occurring within each class of an existing land classification. This expert information is stored in a matrix, or look-up table, in which each row corresponds to an existing land-class (e.g. a mapped vegetation type) and each column corresponds to a community (see Figure 5). The values in the cells of the table can be specified in one of two ways: (1) as zeros and ones, with a 0 indicating that experts believe a given community cannot occur within a given land-class and

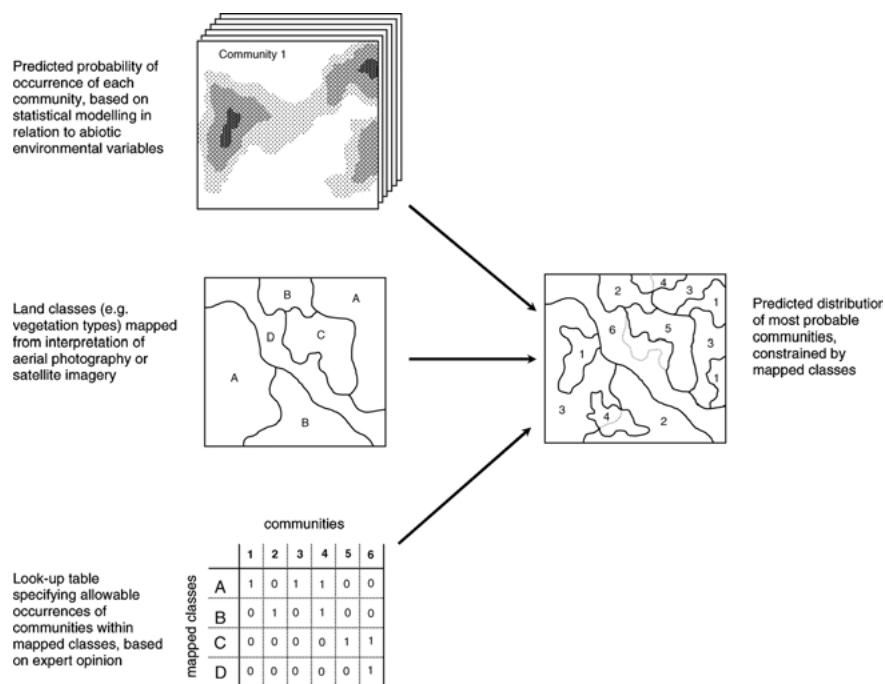


Figure 5. A diagrammatic illustration of an approach to using expert opinion to integrate statistical modelling of community distributions (in relation to abiotic variables) with an existing land classification – e.g. a vegetation map – derived from independent interpretation of aerial photography or satellite imagery.

a 1 indicating that the community can occur within the land-class; or (2) as values between 0 and 1, with each value indicating the proportion of a given land-class that experts believe should be occupied by a given community (the values across each row sum to 1).

When employing the first of these options the look-up table determines which communities are allowed to occur in each grid cell of interest (according to the land-class in which the cell is located), and the cell is then assigned to the community from this set that has the highest predicted probability of occurrence (according to the initial statistical modelling) (see also Figure 5). With the second option the proportions in the table corresponding to the grid cell of interest are treated as prior probabilities of each community occurring within that cell. Bayes' Rule is used to derive posterior probabilities of occurrence by combining these prior probabilities with predicted probabilities of occurrence from the statistical modelling, in a similar manner to that described by Strahler (1980) and De Bruin and Gorte (2000). Each cell is then assigned to the community with the highest posterior probability of occurrence.

Regardless of which option is employed, this general approach provides a simple and highly flexible means of employing expert opinion to integrate statistical modelling of community distributions (in relation to abiotic variables) with an existing land classification – e.g. a vegetation map – derived from independent interpretation of aerial photography or satellite imagery. The approach not only can handle existing classifications containing a large number of classes but also can be applied to situations in which different parts of a region of interest are covered by different classifications. For example, in northeast NSW we are currently trialing the approach as a way of extrapolating vegetation communities (derived from numerical classification of floristic plot data and modelled in relation to abiotic environmental variables) across areas covered by three different levels of existing vegetation mapping: (1) detailed mapping of more than 100 vegetation types derived from interpretation of 1:25 000 aerial photography; (2) coarser mapping of less than 20 vegetation types, also derived from aerial photograph interpretation; and (3) areas without any existing vegetation mapping (including cleared land). Expert opinion is being used to specify which communities can occur within each of the mapped vegetation types. These communities are assigned a value of 1 in the relevant row of the look-up table, while the other communities are assigned a value of 0. The table contains rows for all of the vegetation types from the different classifications. Areas without any existing vegetation mapping are treated as falling within a single 'unmapped' class. All communities are assigned a value of 1 in the row corresponding to this class, which means that grid cells in unmapped areas are assigned to communities purely on the basis of predicted probabilities of occurrence from the statistical modelling. Rows corresponding to mapped vegetation types contain a mixture of zeros and ones and these types therefore play a role in constraining the assignment of grid cells to communities. The number of zeros in the row corresponding to a given type will typically increase

with increasing classification resolution and/or reliability of the available mapping. This in turn increases the extent to which existing mapping will constrain predictions from the statistical modelling.

A canonical classification approach to community-level modelling

We have recently developed and applied an alternative approach to community-level modelling in northeast NSW that integrates community classification and spatial modelling within a single analysis, rather than performing these steps sequentially. The approach employs a form of constrained or 'canonical' numerical classification (*sensu* De Sarbo and Mahajan 1984; Gordon 1996). In the same way that canonical ordination techniques (e.g. Canonical Correspondence Analysis; ter Braak 1986) fit an ordination to biological data such that the ordination axes are functions of environmental variables, the classification technique employed here uses biological data to divide sites into groups such that this grouping is also defined in terms of decision rules based on mapped environmental variables.

The analysis begins by considering the complete set of survey sites and searching for a binary environmental rule that splits the sites into two groups, such that the biological difference between these groups is maximised relative to the biological variation within the groups. Each binary environmental split is defined in terms of a cutpoint, such as 'mean annual rainfall = 850 mm' – which would divide the sites into two groups, those with rainfall <850 mm and those with rainfall greater than or equal to 850 mm. The biological difference between the two groups formed by a split is evaluated by first estimating the compositional dissimilarity (in the occurrence of species) between all pairs of sites involved – e.g. using the Bray–Curtis index. These dissimilarities are then used to calculate the statistic:

$$D = \bar{d}_B - \bar{d}_W$$

where \bar{d}_B is the average dissimilarity between pairs of sites on opposite sides of the environmental split (i.e. one site in group 1 and the other in group 2) and \bar{d}_W is the average dissimilarity between pairs of sites on the same side of the split (i.e. both sites in group 1 or both sites in group 2). The statistical significance of D is estimated using a Monte Carlo randomisation procedure (Manly 1991), in which the observed value is compared to a distribution of D values obtained by randomly permuting the assignment of sites to the two groups. This approach to significance testing closely resembles the ANOSIM technique described by Clarke (1993), and the earlier multi-response permutation procedure described by Mielke et al. (1976).

In the simplest form of the significance test, the null and alternative hypotheses (H_0 and H_1) are:

$$H_0: D = 0$$

$$H_1: D > 0$$

In our application we have generalised this test to allow evaluation of a wider range of hypotheses of the form:

$$H_0: D \leq D_t$$

$$H_1: D > D_t$$

where D_t is a specified threshold with which to compare the observed value of D . Assigning D_t a value greater than 0 helps to alleviate the problem that sites may otherwise be split into groups exhibiting a difference significantly greater than 0 yet this difference is inconsequentially small in biological terms.

If the best available split is statistically significant the sites are partitioned into two groups. This splitting procedure is then applied recursively to the resulting groups. Each group of sites formed by a split is itself evaluated as a candidate for further splitting. The process stops when none of the best available splits for the resulting groups is statistically significant. The technique generates a hierarchical partitioning of sites into groups (communities), in which each division in the hierarchy is defined in terms of an environmental decision rule (see Figures 6 and 7). The strategy can therefore be viewed as combining elements of polythetic divisive classification

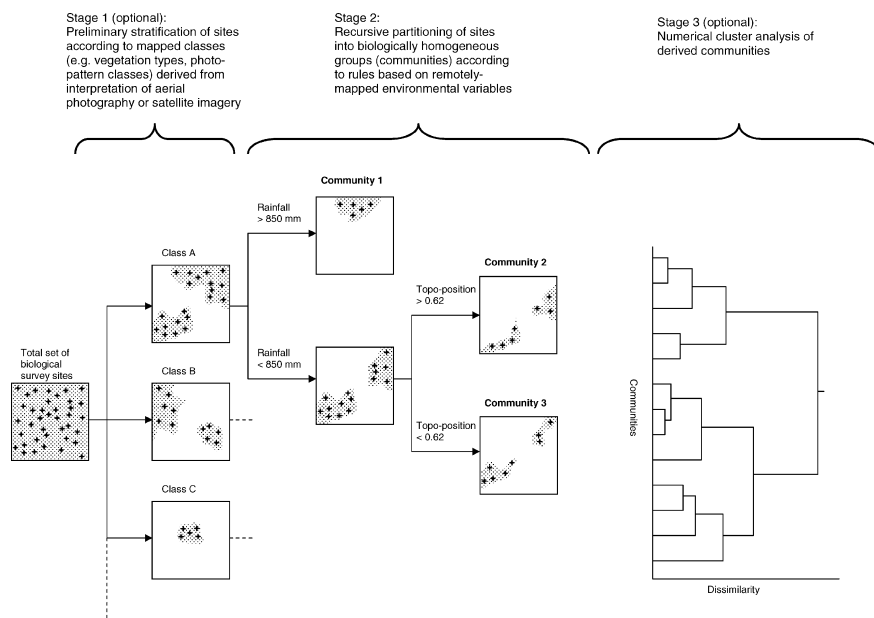


Figure 6. A diagrammatic illustration of a canonical classification approach to community-level modelling that integrates community classification and spatial modelling within a single analysis.

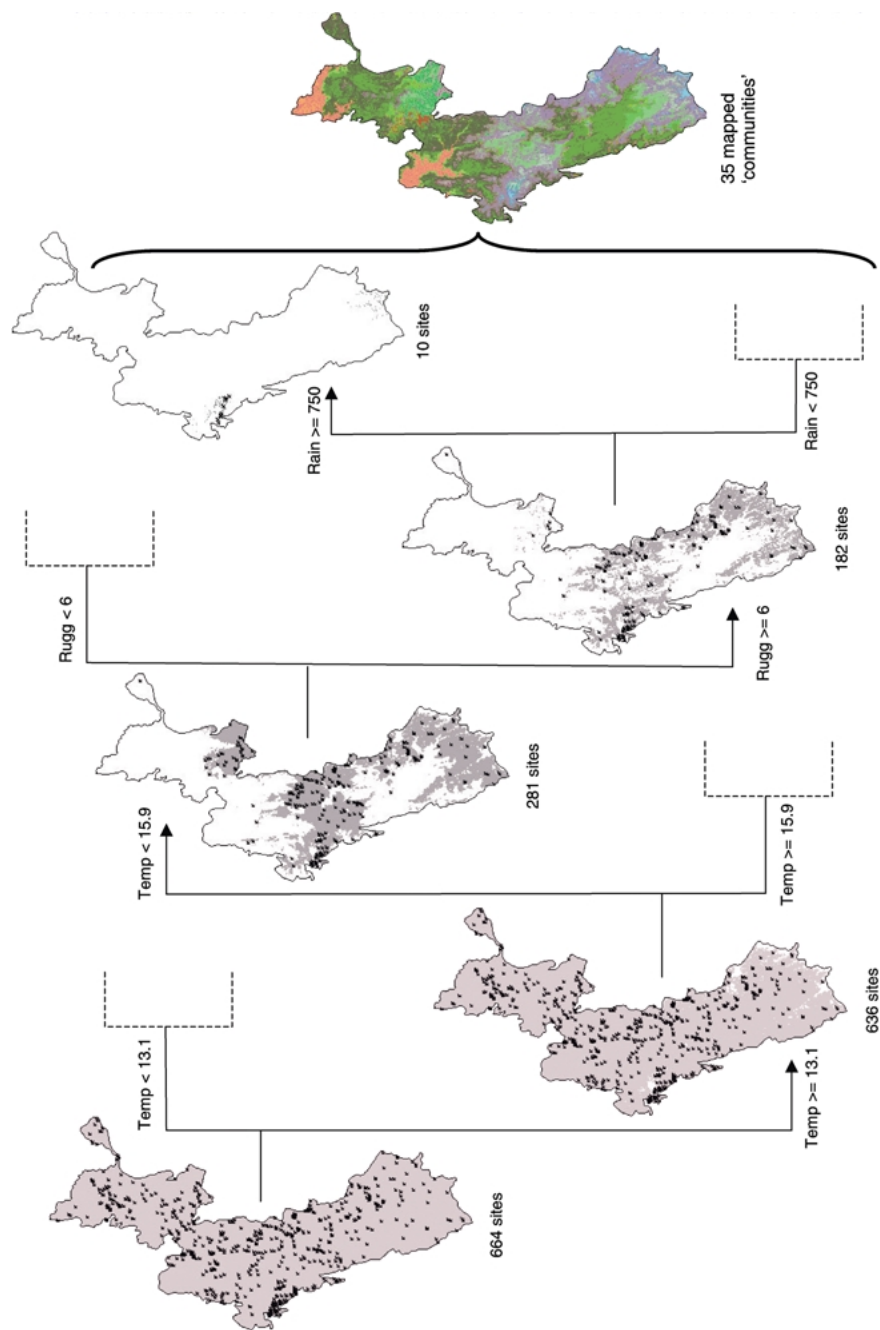


Figure 7. An example application of the new canonical classification technique for deriving and mapping communities, using canopy tree data and abiotic environmental layers for the Nandewar Bioregion. Survey sites are recursively partitioned into communities by identifying environmental splits that maximise the biological difference between groups, in terms of compositional dissimilarity, relative to the biological variation within groups. The environmental rules associated with the splits are then used to map the predicted distribution of each derived community.

(Lance And Williams 1975; Hill 1979) with elements of decision-tree modelling (Brieman et al. 1984; Michaelsen et al. 1987; Moore et al. 1991). Where appropriate, sites may be initially stratified according to some existing classification for a region – e.g. classes from an existing vegetation map, or photo-pattern classes from a preliminary interpretation of aerial photography or satellite imagery. The recursive partitioning process can then be applied separately to the sites within each of these classes (see Figure 6).

As a final step, groups of sites generated by the recursive partitioning can themselves be subjected to numerical cluster analysis – e.g. using a hierarchical agglomerative procedure such as UPGMA, with each initial group treated as an object in the classification (see Figure 6). This provides additional information on the biological relationships (i.e. levels of dissimilarity) between groups generated by the initial partitioning. It also provides a basis for identifying any groups, generated in different parts of the initial partitioning, that appear to be biologically similar and may therefore be candidates for amalgamation into a single community. Decisions about whether or not to amalgamate such groups can be based on significance testing similar to that employed in the initial partitioning.

A major advantage of the canonical classification approach described above is that, because rules defining the environmental distribution of each community are built into the classification, these distributions can be readily extrapolated within a GIS without requiring any additional spatial modelling (we have implemented the approach as an ArcView software extension). Because the approach considers biological and environmental data simultaneously within a single integrated analysis, emphasis is placed on ensuring that derived communities are meaningful not only in terms of biological variation but can also be defined in terms of environmental variation, and thereby mapped. In this way the technique effectively filters out biological variation that cannot be accounted for in terms of the available environmental variables, and therefore cannot be mapped. Some of this unexplained variation may relate to fine-scaled environmental variation not captured by the available environmental GIS layers while some may be simply noise, reflecting sampling error associated with small survey plots, short observation periods or variability between observers.

During the Comprehensive Regional Assessment in northeast NSW, this canonical classification approach was used to subdivide broad forest types (mapped from aerial photography) into finer-scaled floristic communities, employing full vascular plant data recorded at plots scattered throughout each mapped type, in conjunction with abiotic environmental variables (NSW NPWS 1999b). Latitude and longitude were also included as predictor variables in the analysis, thereby allowing geographical splitting of communities to account for biogeographical patterns (e.g. dispersal barriers, range limits) not explainable in terms of environmental variables alone. In some cases, communities generated within different forest types were subsequently amalgamated if they did not exhibit a significant difference in terms of the *D*

statistic. The communities derived through this process were employed as primary coarse-filter surrogates in all land-use planning and decision-making arising out of the Comprehensive Regional Assessment. The performance of these derived communities as biodiversity surrogates was shown to be significantly better than that of the original forest types, based on results of an extended application of the surrogate evaluation study described earlier in this paper (NSW NPWS 1999a). More recently, the approach has also been applied in the Nandewar Bioregion (for a locational map see Figure 1 in Ferrier et al. 2002) to divide this entire region into vegetation communities, using only survey data for canopy trees and remotely mapped abiotic variables – no existing vegetation mapping was available in this case (Ferrier et al. 2000b; see also Figure 7).

Modelling compositional dissimilarity – an alternative strategy for mapping spatial pattern in biodiversity

All of the approaches we have discussed to this point have one thing in common – they require reasonably large quantities of survey data for the entities being modelled, whether these be species, communities or assemblages. These approaches are therefore applicable only to regions with extensive survey datasets. Yet many of the world's regions identified as being of high priority for conservation action (Olson and Dinerstein 1998; Myers et al. 2000) are relatively data-poor – particularly those in the tropics. Available data in these regions may provide only a very sparse and biased sample of a region's biodiversity – in terms of both spatial and taxonomic coverage.

Even in northeast NSW, probably one of the most thoroughly surveyed regions in the world, modelling of biological distributions becomes problematic once we move beyond vertebrates and vascular plants and attempt to consider the other 99% of biodiversity. This is demonstrated by the relatively poor performance of traditional modelling approaches when evaluated using ground-dwelling arthropod data for this region (Figure 4, and Ferrier et al. 2002). A recent study by Ferrier et al. (1999a) suggests that this poor performance might result from ground-dwelling arthropods exhibiting stronger patterns of species turnover or replacement than vertebrates and vascular plants, particularly between historically isolated areas of the same environment or habitat type. Because the density of survey sites is sparse relative to the spatial grain of species turnover, modelled species are probably being predicted to occur in areas where they are, in reality, replaced by other (as yet unsurveyed, or even unknown) species. In this situation it also makes little sense to attempt to classify and model communities, given that the communities detected by surveys may represent only a fraction of all communities occurring within the region.

Traditional approaches to modelling spatial pattern in biodiversity are likely to be challenged in coming years by two current shifts in global conservation focus (Ferrier 2002). One is the increasing emphasis being placed on achieving conservation outcomes in highly diverse, yet relatively data-poor, regions such as many of

those in the tropics. The other is a broadening of the taxonomic focus of conservation effort beyond flagship vertebrate species, and vascular plant communities, to better accommodate all the other groups (e.g. invertebrates) that comprise the greater bulk of biodiversity. Addressing these challenges through modelling of individual entities of biodiversity (e.g. species, communities) alone would be a daunting task. There are simply too many entities and too little locational data for each entity.

A possible alternative in such situations is to shift the focus of modelling from individual entities to collective properties (*sensu* Austin 1999) of biodiversity. Modelling of species richness is probably the best known manifestation of this approach, exemplified by a growing literature describing regional-scaled regression models relating field-surveyed species richness to remotely mapped environmental variables (e.g. Heikkinen and Neuvonen 1997; Leathwick et al. 1998; Lwanga et al. 1998; Wohlgemuth 1998). Regional patterns of species richness extrapolated from such modelling may however be of limited value in regional conservation planning. Conserving only areas of highest species richness will not necessarily maximise the total number of species conserved in a region, as these areas will probably fail to include species that occur only in areas of lower richness. To maximise the total number of species, and therefore diversity, represented in a set of conservation areas we require additional information on the complementarity or compositional dissimilarity of these areas – i.e. how dissimilar two areas are in terms of the species they contain (Pressey et al. 1993).

Compositional dissimilarity can be viewed as another collective property of biodiversity, alongside species richness. In terms of Whittaker's (1972, 1977) original conceptual framework for partitioning species diversity into components, local species richness equates to the α diversity component of inventory diversity, while compositional dissimilarity is an amalgam of the β and γ (*sensu* Caddy 1993) diversity components of differentiation diversity. β diversity is the difference in composition of species between different habitats or environments. γ diversity is the difference in composition of species between geographically isolated occurrences of the same habitat or environment.

We have recently developed a new approach to modelling the level of compositional dissimilarity between pairs of survey sites as a function of the environmental and geographical separation of these sites. The basic analytical strategy is derived from that of permutational matrix regression (Manly 1986; Smouse et al. 1986; Legendre et al. 1994), itself an extension of the matrix correlation approach widely used to test correlations between distance matrices (Burgman 1987; Legendre 1993). In its original form, matrix regression simply employs multiple linear regression to predict the dissimilarities (or distances) in one matrix (the response) as a function of the distances in one or more independent (explanatory) matrices. In this particular application the response matrix contains compositional dissimilarities between all pairs of survey sites calculated, for example, using the Bray–Curtis measure (Poulin and Morand 1999; Ferrier et al. 1999a). A sites-by-sites matrix is also prepared for each of the explanatory variables.

For example, if one of these variables is mean annual rainfall, then a matrix is prepared in which each value is the difference in rainfall between a given pair of sites. Linear matrix regression produces models of the form:

$$d_{ij} = a + b_1(|x_{1i} - x_{1j}|) + b_2(|x_{2i} - x_{2j}|) + \dots$$

where d_{ij} is the compositional dissimilarity between sites i and j ; a , the intercept (the expected compositional dissimilarity when two sites have the same values for all explanatory variables); x_{1i} , the value for the first explanatory variable (e.g. rainfall) at site i ; b_1 the regression coefficient for the difference in the first explanatory variable.

Significance testing in matrix regression is performed by Monte Carlo permutation to overcome the problem of lack of independence between site pairs. This approach has been used in northeast NSW to compare regional patterns of compositional dissimilarity in ground-dwelling arthropods with patterns exhibited by vertebrates and vascular plants (Ferrier et al. 1999a).

Matrix regression, in the form described above, assumes that all relationships between the biological and environmental variables are linear. This assumption is unlikely to be satisfied in many real-world datasets. We have therefore recently extended the technique of matrix regression to address two common types of non-linearity encountered in ecological data. We refer to the resulting approach as ‘generalised dissimilarity modelling’ (GDM) (Ferrier et al. 1999b; Ferrier 2002; Ferrier and Manion in preparation; see also Figure 8). The first type of non-linearity addressed by GDM relates to the well-established fact that the relationship between ecological separation of sites and observed compositional dissimilarity is not linear, but instead curvilinear (Gauch 1973; Faith et al. 1987). Most measures of compositional dissimilarity, including the Bray–Curtis measure, are constrained between 0 and 1. As ecological separation increases, and sites share progressively fewer species, the observed dissimilarity will approach, but cannot exceed, a value of 1. The relationship between ecological separation and compositional dissimilarity is therefore asymptotic. By assuming a linear relationship between ecological separation and compositional dissimilarity the original matrix regression approach is unlikely to perform well when applied to datasets exhibiting a high level of β and/or γ diversity, and therefore containing compositional dissimilarities close to, or equal to, 1.

With GDM this problem is addressed by fitting models using generalised linear modelling instead of ordinary linear regression. This allows the curvilinear relationship between ecological separation and compositional dissimilarity to be accommodated through specification of appropriate link and variance functions. The link function currently being trialed for use with the Bray–Curtis dissimilarity measure is:

$$\eta = -\ln(1 - \mu)$$

which is designed to approximate the asymptotic behaviour of compositional dissimilarity with increasing ecological separation. Other link functions may be more

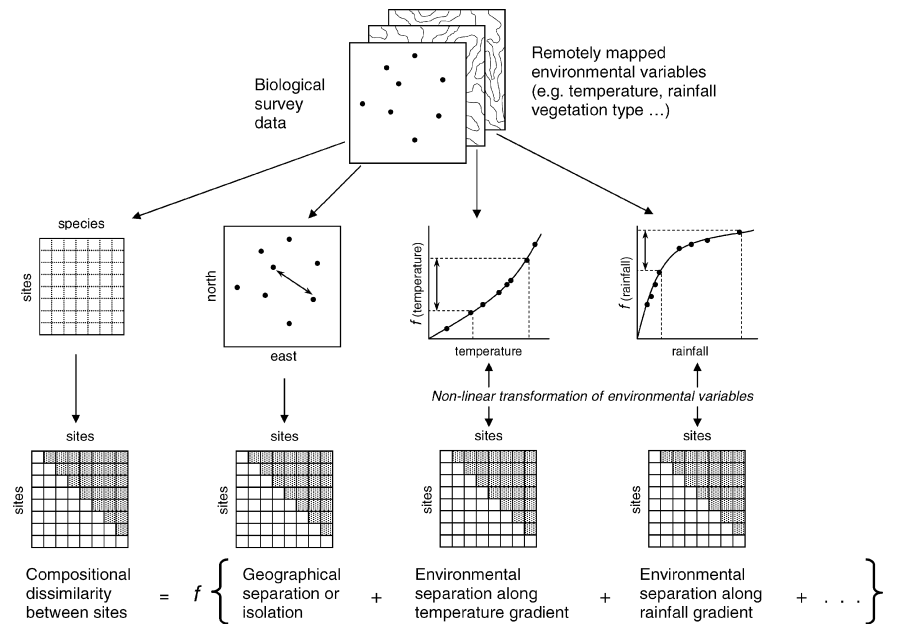


Figure 8. A diagrammatic illustration of the GDM approach to modelling compositional dissimilarity (based on Ferrier 2002).

appropriate for particular datasets and dissimilarity measures, and therefore worthy of further investigation. The variance function currently being trialed for use with the Bray–Curtis measure (applied to presence/absence data) is:

$$\mu(1 - \mu)$$

The second type of non-linearity addressed by GDM relates to the rate of compositional change, or ‘turnover’, along environmental gradients. In ordinary matrix regression it is assumed that this rate remains constant across the entire range of each environmental variable. However, violations of this assumption may be common in real-world datasets (Whittaker 1977; Wilson and Mohler 1983; McNaughton 1994; Oksanen and Tonteri 1995; Simmons and Cowling 1996). Environmental gradients are measured on essentially arbitrary scales that may not concord well with real patterns of compositional turnover (e.g. log-transformed mean annual rainfall may provide better concordance than untransformed rainfall).

With GDM, variation in the rate of compositional turnover along gradients is accommodated through automated non-linear transformation of environmental variables. This is achieved using monotonic I-splines, in a similar fashion to that described by Winsberg and De Soete (1997). The splines are used to fit a transforming function f to each environmental variable that maximises the reduction in deviance achieved by including $|f(x_i) - f(x_j)|$ as a predictor in the model. By applying the link function described

above, in conjunction with monotonic transformation of environmental variables, models derived using GDM take the following form (see also Figure 8):

$$-\ln(1 - d_{ij}) = a + |f_1(x_{1i}) - f_1(x_{1j})| + |f_2(x_{2i}) - f_2(x_{2j})| + \dots$$

Matrix regression approaches, including GDM, provide a natural means of incorporating geographical separation as an additional predictor in modelling of compositional dissimilarity, thereby providing the potential to integrate modelling of both the β and γ components of differentiation diversity. An example of functions, fitted using GDM, relating compositional dissimilarity of sites surveyed in northeast NSW for ground-dwelling spiders to geographical separation and location on environmental gradients is provided in Figure 9. The potential applicability of GDM to modelling of γ diversity is discussed further by Ferrier (2002), including the possibility of replacing simple Euclidean distance with more sophisticated measures of biogeographical isolation in such analyses.

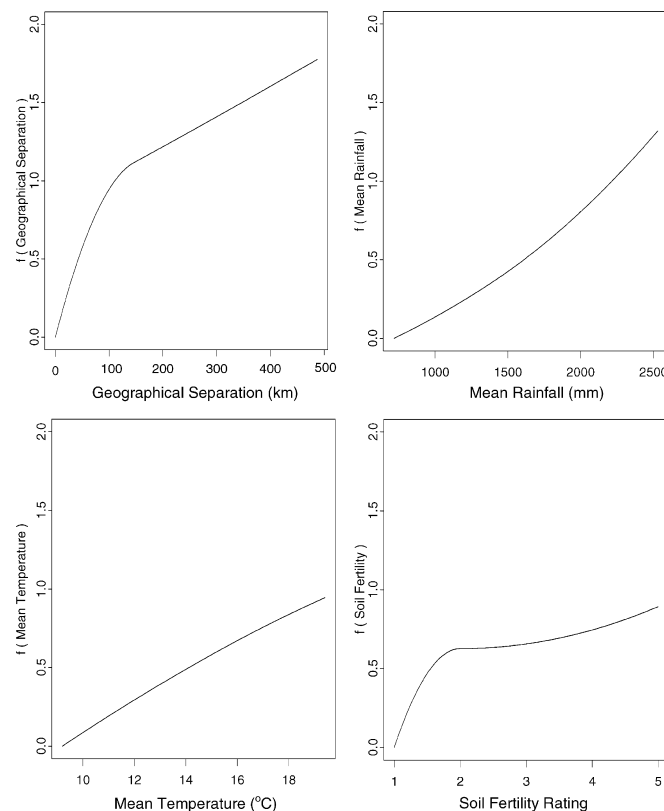


Figure 9. An example of functions fitted by applying GDM to data for ground-dwelling spiders in northeast NSW.

We are currently investigating two other possible refinements to the GDM approach that should further improve the technique's applicability to real-world conservation planning. The first of these involves extending the approach to better accommodate biological data collected by *ad hoc* surveys (i.e. presence-only data rather than presence/absence data). The second refinement involves adding a capability to model interactions between environmental predictors – e.g. the rate of compositional turnover relating to soil fertility might vary with mean annual rainfall.

In some respects GDM can be viewed as a form of constrained ordination and therefore bears a resemblance to other constrained ordination techniques such as canonical correspondence analysis (CCA) and redundancy analysis (RDA) (ter Braak 1986; Jongman et al. 1995). Advantages of GDM relative to CCA for regional conservation planning include: (1) GDM can potentially accommodate any measure of compositional (or even phylogenetic or genetic) dissimilarity, whereas CCA is based on a single χ^2 measure. (2) GDM can accommodate non-linear relationships between environmental distance and biological dissimilarity, whereas CCA assumes a linear relationship. (3) GDM can accommodate variation in the rate of compositional turnover along an environmental gradient, whereas CCA assumes a constant rate. (4) GDM provides a more natural and flexible means of incorporating geographical separation or isolation (γ diversity) into modelling of compositional dissimilarity.

Ferrier (2002) describes two ways in which modelling of compositional dissimilarity might be employed in regional conservation planning. In the first approach predictions from such models are used to constrain the derivation of environmental classifications. Numerical classification of all grid cells within a region into environmental classes or 'domains' is an increasingly popular technique for deriving coarse-filter surrogates for biodiversity in data-poor regions (e.g. Mackey et al 1989; Bernert et al. 1997; Fairbanks and Benn 2000; Nix et al. 2000). By basing such classification on predicted compositional dissimilarities between grid cells (e.g. from a model fitted using GDM), in place of dissimilarities derived from environmental variables alone, a surrogate can be derived that is likely to better reflect real patterns of β and γ diversity. Classes produced in this manner can then be employed in conservation planning in exactly the same way as any other land classification. Alternatively predictions from modelling of compositional dissimilarity could be used more directly to prioritise and select conservation areas, by employing techniques such as the environmental-diversity (ED) approach pioneered by Faith and Walker (1996a,b) in which conservation areas are assessed in relation to continuous environmental data rather than discrete land classifications.

Ferrier (2002) also describes how GDM can be further combined with new survey-design techniques to provide an integrated strategy for cost-effectively refining information on spatial pattern in biodiversity within data-poor regions. In this strategy predictions from an initial model, based on existing biological data, are used to strategically locate additional survey sites to maximise improvement in the environmental and geographical coverage of sampling. Data from these new sites are then used to

refine the initial model, thereby providing an improved basis for selecting any further sites. The process can be repeated iteratively, thereby allowing spatial information on biodiversity to be refined progressively over time, subject to availability of resources for survey work. This integrated approach is currently being trialed in a number of regions in Australia, and in collaborative projects in Guyana and South Africa.

Conclusions

Statistical modelling of biological survey data in relation to remotely mapped environmental variables provides a powerful means of making more effective use of sparse data in regional conservation planning. To date, most attention in this field has been directed towards modelling distributions of individual species. While species-level modelling has played, and should continue to play, an important role in conservation planning we feel that even greater benefit could be gained if this particular approach were more widely supplemented by community-level modelling approaches such as those described in this paper. Species-level modelling makes good sense when applied to selected species of particular conservation concern – e.g. threatened, focal or flagship species. However, the approach may not be an effective or efficient way of modelling spatial pattern in biodiversity as a whole, particularly when dealing with highly diverse taxa in poorly surveyed regions. There will simply be too many species, and too little data per species, for these to be modelled effectively on a species-by-species basis.

One of the main challenges of community-level modelling is deciding which of the many possible strategies is most appropriate for a given situation. Community-level modelling offers a far richer set of analytical strategies than does species-level modelling. With species-level modelling, the only real choice is between statistical model-fitting techniques (e.g. GLM, GAM, decision-trees or neural networks). However, with community-level modelling the choice of a particular statistical technique is often a relatively low-level concern, overshadowed by higher-level decisions about the overall analytical strategy – e.g. whether to adopt a modelling-then-classification or classification-then-modelling approach, or whether to derive and model site-groups or species-groups. There is little point in attempting to identify a single ‘best practice’ strategy for use in all situations. The approaches described in this paper each have particular strengths and weaknesses that need to be considered when evaluating their applicability to a given dataset and analytical objective. An approach that performs well in one situation – i.e. to achieve a particular objective using a particular dataset – may not perform well when applied to a different objective and/or dataset.

The choice of which community-modelling strategy to employ in any given situation needs to be informed by an awareness of the full range of alternatives. This paper has demonstrated that the set of possible approaches is potentially much richer than has been previously recognised. We hope this may engender increased interest in the application of new or neglected approaches, thereby adding further value to available biolog-

ical and environmental data and providing an improved basis for regional conservation planning. Four potential shifts in methodology that we feel are particularly worthy of further attention are: (1) more widespread use of species-group (assemblage) modelling as an alternative to site-group (community) modelling; (2) closer integration of community classification and spatial modelling through employment of canonical classification techniques; (3) increased incorporation of information from aerial photography and satellite imagery (including existing vegetation mapping and related land classifications) into community-level modelling, alongside abiotic environmental variables; and (4) increased use of modelling of collective properties of biodiversity (particularly compositional dissimilarity) in data-poor regions, as an alternative to modelling individual biodiversity entities (species, communities or assemblages).

Acknowledgements

The collection, entry and checking of data used to model biological 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. Jennie Pearce assisted in investigating alternative modelling solutions during the early stages of the development of GDM. Carmel Flint and Jill Smith assisted in trialing the application of the canonical classification approach to data from northeast NSW. Some of the work on evaluating the performance of different community-level modelling strategies, and developing GDM, was funded by Environment Australia.

References

- Alexander R. and Millington A.C. 2000. *Vegetation Mapping: From Patch to Planet*. Wiley, New York.
- 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. 1999. The potential contribution of vegetation ecology to biodiversity research. *Ecography* 22: 465–484.
- Austin M.P. and Belbin L. 1982. A new approach to the species classification problem in floristic analysis. *Australian Journal of Ecology* 7: 75–89.
- Belbin L. 1995. *PATN Users Guide and Technical Reference*. CSIRO Division of Wildlife and Ecology, Canberra, Australia.
- Bernert J.A., Eilers J.M., Sullivan T.J., Freemark K.E. and Ribic C. 1997. A quantitative method for delineating regions: an example for the Western Corn Belt Plains Ecoregion of the USA. *Environmental Management* 21: 405–420.
- Brieman L., Friedman J.H., Olshen R.A. and Stone C.J. 1984. *Classification and Regression Trees*. Wadsworth International Group, Belmont, California.
- Burgman M.A. 1987. An analysis of the distribution of plants on granite outcrops in southern Western Australia using Mantel tests. *Vegetatio* 71: 79–86.
- Cawsey E.M., Austin M.P. and Baker B.L. 2002. Regional vegetation mapping in Australia: a case study in the practical use of statistical modelling. *Biodiversity and Conservation* 11: 2239–2274 (this issue).
- Clarke K.R. 1993. Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology* 18: 117–143.

- Clarke K.R. and Warwick R.M. 1998. A taxonomic distinctness index and its statistical properties. *Journal of Applied Ecology* 35: 523–531.
- Cody M.L. 1993. Bird diversity components within and between habitats in Australia. In: Ricklefs R.E. and Schluter D. (eds) *Species Diversity in Ecological Communities: Historical and Geographical Perspectives*. University of Chicago Press, Chicago, Illinois, pp. 147–158.
- De Bruin S. and Gorte B.G.H. 2000. Probabilistic image classification using geological map units applied to land-cover change detection. *International Journal of Remote Sensing* 21: 2389–2402.
- De Sarbo W.S. and Mahajan V. 1984. Constrained classification: the use of a priori information in cluster analysis. *Psychometrika* 49: 187–215.
- Fairbanks D.H.K. and Benn G.A. 2000. Identifying regional landscapes for conservation planning: a case study from KwaZulu-Natal, South Africa. *Landscape and Urban Planning* 50: 237–257.
- Faith D.P. 1996. Conservation priorities and phylogenetic pattern. *Conservation Biology* 10: 1286–1289.
- Faith D.P. and Walker P.A. 1996a. Environmental diversity: on the best-possible use of surrogate data for assessing the relative biodiversity of sets of areas. *Biodiversity and Conservation* 5: 399–415.
- Faith D.P. and Walker P.A. 1996b. How do indicator groups provide information about the relative biodiversity of different sets of areas? On hotspots, complementarity and pattern-based approaches. *Biodiversity Letters* 3: 18–25.
- Faith D.P., Minchin P.R. and Belbin L. 1987. Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio* 69: 57–68.
- Ferrier S. 2002. Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Systematic Biology* 51: 331–363.
- 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., Manion G., Drielsma M. and Smith J. 2000b. Ecosystem Classification for the Nandewar Bio-region Derived From a Preliminary Analysis of Abiotic Environmental Variables and Floristic Survey Data. Unpublished report. NSW National Parks and Wildlife Service, Armidale, Australia.
- Ferrier S., Watson G., Hines H. and Brown D. 2000c. 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., Watson G., Pearce J. and Drielsma M. 2002. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling. *Biodiversity and Conservation* 11: 2275–2307 (this issue).
- Franklin J. 1995. Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. *Progress in Physical Geography* 19: 474–499.
- Gauch Jr H.G. 1973. The relationship between sample similarity and ecological distance. *Ecology* 54: 618–622.
- Gordon A.D. 1996. A survey of constrained classification. *Computational Statistics and Data Analysis* 21: 17–29.
- Guisan A. and Zimmermann N.E. 2000. Predictive habitat distribution models in ecology. *Ecological Modelling* 135: 147–186.

- Hastie T.J. and Tibshirani R. 1990. Generalised Additive Models. Chapman & Hall, London.
- Heikkinen R.K. and Neuvonen S. 1997. Species richness of vascular plants in the subarctic landscape of northern Finland: modelling relationships to the environment. *Biodiversity and Conservation* 6: 1181–1201.
- Hill M.O. 1979. TWINSpan: A FORTRAN Program for Arranging Multivariate Data in an Ordered Two-Way Table by Classification of the Individuals and Attributes. Department of Ecology and Systematics, Cornell University, Ithaca, NY.
- 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.
- Jongman R.H., ter Braak C.J.F. and van Tongeren O.F.R. 1995. Data Analysis in Community and Landscape Ecology. Cambridge University Press, Cambridge, UK.
- Keith D.A. and Bedward M. 1999. Native vegetation of the South East Forests region, Eden, New South Wales. *Cunninghamia* 6: 1–218.
- Lance G.N. and Williams W.T. 1975. REMUL: a new divisive polythetic classificatory program. *Australian Computer Journal* 7: 109–112.
- Leathwick J.R., Burns B.R. and Clarkson B.D. 1998. Environmental correlates of tree alpha-diversity in New Zealand primary forests. *Ecography* 21: 235–246.
- Lees B.G. and Ritman K. 1991. Decision-tree and rule-induction approach to integration of remotely sensed and GIS data in mapping vegetation in disturbed or hilly environments. *Environmental Management* 15: 823–831.
- Legendre P. 1993. Spatial autocorrelation: trouble or new paradigm? *Ecology* 74: 1659–1673.
- Legendre P., Lapointe F.-J. and Casgrain P. 1994. Modeling brain evolution from behavior: a permutational regression approach. *Evolution* 48: 1487–1499.
- Lehmann A., Leathwick J.R. and Overton J.McC. 2002. Assessing New Zealand fern diversity from spatial predictions of species assemblages. *Biodiversity and Conservation* 11: 2217–2238 (this issue).
- Lenihan J.M. 1993. Ecological response surfaces for North American boreal tree species and their use in forest classification. *Journal of Vegetation Science* 4: 667–680.
- Lwanga J.S., Balmford A. and Badaza R. 1998. Assessing fern diversity: relative species richness and its environmental correlates in Uganda. *Biodiversity and Conservation* 7: 1387–1398.
- Mackey B.G., Nix H.A., Stein J.A. and Cork S.E. 1989. Assessing the representativeness of the Wet Tropics of Queensland World Heritage Property. *Biological Conservation* 50: 279–303.
- 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.F.J. 1986. Randomization and regression methods for testing for associations with geographical, environmental and biological distances between populations. *Researches on Population Ecology* 28: 201–218.
- Manly B.F.J. 1991. Randomization and Monte Carlo Methods in Biology. Chapman & Hall, London.
- Margules C.R. and Pressey R.L. 2000. Systematic conservation planning. *Nature* 405: 243–253.
- McCullagh P. and Nelder J.A. 1989. Generalized Linear Models, 2nd Edition. Chapman & Hall, London.
- McKenzie N.L., Belbin L., Margules C.R. and Keighery G.J. 1989. Selecting representative reserve systems in remote areas. *Biological Conservation* 50: 239–261.
- McNaughton S.J. 1994. Conservation goals and the configuration of biodiversity. In: Forey P.L., Humphries C.J. and Vane-Wright R.I. (eds). Systematics and Conservation Evaluation. Clarendon Press, Oxford, UK, pp. 41–62.
- Michaelsen J., Davis F.W. and Borchert M. 1987. A non-parametric method for analyzing hierarchical relationships in ecological data. *Coenoses* 2: 39–48.
- Mielke P.W., Berry K.J. and Johnson E.S. 1976. Multi-response permutation procedures for a priori classifications. *Communications in Statistics part A - Theory and Methods* 5: 1409–1424.
- 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.
- 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.

- Nix H.A. 1991. Biogeography: pattern and process. In: Nix H.A. and Switzer M.A. (eds). *Rainforest Animals: Atlas of Vertebrates Endemic to Australia's Wet Tropics*. Australian National Parks and Wildlife Service, Canberra, Australia, pp. 11–39.
- Nix H.A., Faith D.P., Hutchinson M.F., Margules C.R., West J., Allison A., Kesteven J.L., Natera G., Slater W., Stein J.L. and Walker P. 2000. *The BioRap Toolbox: a National Study of Biodiversity Assessment and Planning for Papua New Guinea*. Consultancy Report to the World Bank. Centre for Resource and Environmental Studies, Australian National University, Canberra, Australia.
- 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 1999a. *Derived Forest Ecosystems: an Evaluation of Surrogacy Value and Internal Biological Variation*. A Project Undertaken As Part of the NSW Comprehensive Regional Assessments. Resource and Conservation Division, Department of Urban Affairs and Planning, Sydney, Australia.
- NSW NPWS 1999b. *Forest Ecosystem Classification and Mapping for Upper and Lower North East CRA Regions*. A Project Undertaken As Part of the NSW Comprehensive Regional Assessments. Resource and Conservation Division, Department of Urban Affairs and Planning, Sydney, Australia.
- Oksanen J. and Tonteri T. 1995. Rate of compositional turnover along gradients and total gradient length. *Journal of Vegetation Science* 6: 815–824.
- 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.
- Poulin R. and Morand S. 1999. Geographical distances and the similarity among parasite communities of conspecific host populations. *Parasitology* 119: 369–374.
- 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.
- Simmons M.T. and Cowling R.M. 1996. Why is the Cape Peninsula so rich in plant species? An analysis of the independent diversity components. *Biodiversity and Conservation* 5: 551–573.
- Smouse P.E., Long J.C. and Sokal R.R. 1986. Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Systematic Zoology* 35: 627–632.
- Strahler A.H. 1980. The use of prior probabilities in maximum likelihood classification of remotely sensed data. *Remote Sensing of Environment* 10: 1135–1163.
- ter Braak C.J.F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology* 67: 1167–1179.
- Whittaker R.H. 1972. Evolution and measurement of species diversity. *Taxon* 21: 213–251.
- Whittaker R.H. 1977. Evolution of species diversity in land communities. *Evolutionary Biology* 10: 1–67.
- Wilson M.V. and Mohler C.L. 1983. Measuring compositional change along gradients. *Vegetatio* 54: 129–141.
- Winsberg S. and De Soete G. 1997. Multidimensional scaling with constrained dimensions: CONSCAL. *British Journal of Mathematical and Statistical Psychology* 50: 55–72.
- Wohlgemuth T. 1998. Modelling floristic species richness on a regional scale: a case study in Switzerland. *Biodiversity and Conservation* 7: 159–177.
- Woinarski J.C.Z., Price O. and Faith D.P. 1996. Application of a taxon priority system for conservation planning by selecting areas which are most distinct from environments already reserved. *Biological Conservation* 76: 147–159.
- Yee T.W. and Mitchell N.D. 1991. Generalized additive models in plant ecology. *Journal of Vegetation Science* 2: 587–602.