Data requirements and data sources for biodiversity priority area selection

P H WILLIAMS[†], C R MARGULES* and D W HILBERT

CSIRO Sustainable Ecosystems, Tropical Forest Research Centre and the Rainforest Co-operative Research Centre, PO Box 780, Atherton Queensland 4883, Australia

[†]*Department of Entomology, The Natural History Museum, Cromwell Road, London SW7 5BD, UK*

**Corresponding author (Fax, 61-7-4091-8888; Email, chris.margules@csiro.au)*

The data needed to prioritize areas for biodiversity protection are records of biodiversity features – species, species assemblages, environmental classes – for each candidate area. Prioritizing areas means comparing candidate areas, so the data used to make such comparisons should be comparable in quality and quantity. Potential sources of suitable data include museums, herbariums and natural resource management agencies. Issues of data precision, accuracy and sampling bias in data sets from such sources are discussed and methods for treating data to minimize bias are reviewed.

[Williams P H, Margules C R and Hilbert D W 2002 Data requirements and data sources for biodiversity priority area selection; *J. Biosci.* (*Suppl. 2*) **27** 327–338]

1. Introduction

Biodiversity priority areas include protected areas as defined by IUCN (1994) but could also include other areas, which may be important for off-reserve protection. Systematic methods for identifying biodiversity priority areas require two separate but interdependent activities: compiling good data on the distribution and abundance patterns of the features to be conserved, the biodiversity surrogates; and the development of appropriate procedures for using those data to determine priorities (see Margules *et al* 2002). Both are necessary but neither alone is sufficient. Biodiversity surrogates might be species, species assemblages such as vegetation or habitat types, environmental domains, or combinations of these. Data on the distribution patterns, abundance or spatial extent of features such as species and species assemblages can be compiled from collections of field records held in museums and herbariums, or they can be gathered from new surveys designed specifically for such purposes. Environmental data can be compiled from maps and increasingly in electronic form from, for example, meteorological agencies. All available data have to be evaluated critically to assess the geographical and temporal bias, as well as their suitability as biodiversity surrogates. In this paper we examine both the values and the limitations for systematic conservation planning of data sets derived from collections that already exist; what we call here 'existing data'. The acquisition of new data is always desirable, but not always feasible given time and cost constraints so existing data are widely used. Appropriate methods for the design of surveys to collect new data have been developed by Gillison and Brewer (1985), and Austin and Heyligers (1989) and tested recently by Wessels *et al* (1998).

Existing data such as museum collections or published maps are often compiled from many different field collections. The details of sampling methods are often unrecoverable and each collection might have its own peculiar biases. Field records are often taken from the places that the collector expected to find what he or she was looking for, or collected opportunistically. Records of koalas (Margules

Keywords. Biodiversity priority areas; data accuracy; data precision; sample bias; spatial models

and Austin 1994) and elapid snakes (Longmore 1986) in parts of Australia, for example, map the road network, and records of many trees in the Amazon map the river network (Williams *et al* 1996b). Figure 1 is a map of the records of tree species on the Yucatan Peninsula in Mexico, which clearly maps the road network. Given that prioritizing areas for biodiversity conservation using complementarity (Margules *et al* 2002) is essentially a matter of comparison, this spatial bias is significant. It is difficult to make valid comparisons when not all areas have records taken from them. The challenge in using existing data is to devise appropriate treatments so that the effects of these sampling biases are minimized.

The reality is that existing data sets are generally far from ideal. In the face of numerous and probably irreversible planning and policy decisions affecting biodiversity every day, it is necessary to make full use of existing data. Nonetheless, it is also necessary to acknowledge their limitations and establish ideal data requirements, not only as an aspiration, but also because identifying the ideal helps to capitalize on the useful information content of existing data.

The values and limitations of existing data are considered here under four headings. First, the general form of data required for selecting biodiversity priority areas, including a discussion of areas and the biodiversity features that characterize them. Second, the assumptions of the relationship of sample to 'population' (in the statistical sense) required for the use of these data. Third, the kinds of data that already exist, and sources of those data. Fourth, how existing data can be assessed and, if necessary, treated, so that they more nearly satisfy the assumptions required for a priority areas analysis. An important question is to establish when existing data should be rejected as being so strongly in violation of necessary assumptions as to be dangerously misleading.

Figure 1. A map of field records of tree species on the Yucatan Peninsula, Mexico. These records map the road network. The map is courtesy of Guillermo Ibarra Manríquez and Rafael Durán.

2. General form of data

The basic data requirement is for an areas by features matrix (figure 1 of Margules *et al* 2002). This describes a circumscribed geographical space, within which biodiversity priorities are to be determined. Decisions have to be taken as to which areas to include, and which features to use to describe those areas. These decisions should be governed by a precise specification of the conservation goal, but they will inevitably also be constrained by the quality and quantity of any existing data, and the resources available for compiling, evaluating, and analysing those data, or for collecting new data. The geographical space may be the entire globe, regions of the globe, nations, regions, or biomes across nations, or parts of nations, states within nations, regions, or biomes within nations, and so on. The methods for analysing the data matrix and identifying biodiversity priority areas described by Margules *et al* (2002 and see also Pressey *et al* 1993; Margules and Pressey 2000) are independent of scale.

2.1 *Areas*

The areas in the matrix are the priority area candidates. They may be mapped as irregular polygons, such as catchments, habitat remnants, or units of tenure, or as regular polygons, such as grid cells. They may cover the entire geographic space or, as in the case of habitat remnants in cropland for example, only part of the space. They may vary in size or be uniform in size.

Variation in size does not affect the priority area selection procedures because those methods depend on the identity of the features present, not on their density or on how big is the area in which they occur. Most measures of diversity are intimately linked up with area extent because they are properties of spatially defined sets of objects. For example, species richness within a habitat is positively correlated with size of areas in the well-known species-area relationship. In the absence of any other information, richness has been used as a measure of biodiversity value. However, summarizing data as counts of the number of features, such as species, loses the identity of those features, and thus makes it impossible to use complementarity. As Higgs and Usher (1980) first noted, it is the number of species in common (or, more appropriately, the number of species not shared by two areas) which should determine relative value for biodiversity protection. The more species that are res-

tricted to an area, the higher its value. A related problem is that areas of richness for different groups do not necessarily coincide (Prendergast *et al* 1993; Williams and Gaston 1994, 1998; Prendergast and Eversham 1997). Thus, selecting rich areas is often inefficient and may miss many species (Pressey and Nicholls 1989), particularly those that have small range sizes and occur in relatively species-poor assemblages (Williams *et al* 1996a).

Standardized regular grids may make comparisons between data sets more meaningful and data sets themselves less subject to change if political boundaries change, particularly at broader spatial scales. Because measures of diversity are correlated with areal extent, near-equal-area grid systems have been used in many studies. They include a rectangular grid based on intervals of 2° latitude (McAllister *et al* 1994) and a hexagonal grid (White *et al* 1992) used as a basis for the Gap Analysis program in the United States (Scott *et al* 1993). Another popular grid system is based on the Universal Transverse Mercator (UTM) projection, though a substantial proportion of the grid cells are not equal in area. It is used for the European Invertebrate Survey (EIS) (Leclercq 1979; Pekkarinen *et al* 1981) and for *Atlas Florae Europaeae* (Jalas and Suominen 1972–1994). Summaries of regional patterns of bumblebees world-wide (Williams and Gaston 1998) used a cylindrical equal area projection with minimum shape distortion at latitudes 46° north and south (Maling 1974), where bumblebee records are particularly numerous (figure 2). The bumblebee grid was calculated from intervals of 10° longitude and appears on this projection as equal-area squares (ca 611,000 km^2). The area of the Earth's land surface within each cell still varies greatly, as does the proportion of different kinds of habitat.

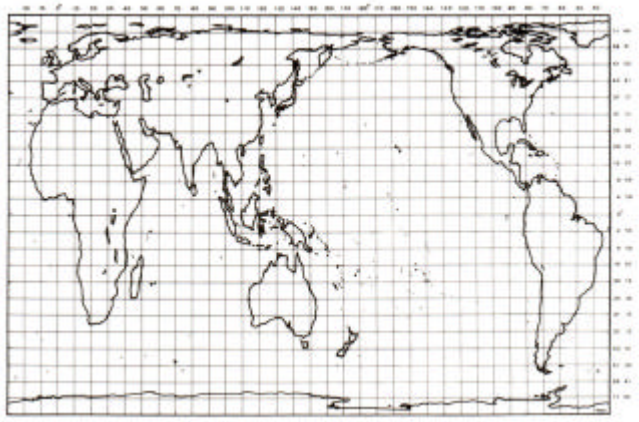

Figure 2. Map of the world (excluding Antarctica) using a cylindrical equal-area projection, orthomorphic (minimum shape distortion) at 46° north and south (where bumblebee records are particularly plentiful). Intervals of 10° longitude (top of map) were used to calculate intervals of latitude (right of map) to provide equal-area grid cells of approximately 611,000 km^2.

Nevertheless, variation in the size of areas does have some important implications. Smaller areas may have smaller populations of the target taxa and this may cause serious management problems. The area selection process can address this problem in two ways (and see Gaston *et al* 2002). One is to ensure that areas in the matrix are sufficiently large in the first place. Unfortunately, this may not always be possible, for example with fragmented habitat, in which case another option is to run a priority areas analysis regardless of area size, and then adjust the boundaries of any small areas chosen for management or acquisition purposes after analysis (e.g. Bedward *et al* 1992). Area boundaries have to be as flexible as possible. Areas in the matrix are units to work with, not objects that are fixed in size forever more.

The size of areas used in the selection process may not correspond to the size of areas actually acquired and managed. A lack of spatial correspondence between the two can cause three problems. First, when selection areas are much larger than management areas it may not be possible to manage all the features of a chosen priority area. A finer resolution of the spatial distribution patterns of the features will be required at the management or acquisition stage. Second, when selection areas are similar in size to management areas but do not correspond precisely; different features may be found in management areas, thus impeding the selection goal. Third, if selection areas are smaller than management areas the total complement of features found in a management area will be divided among several selection areas, making it difficult, at the selection stage, to know when the conservation goal has been achieved.

Analyses could be repeated at progressively finer spatial scales. In this way, areas might be analysed at a broad scale, using grid cells or administrative areas, without aiming to identify specific local conservation areas. Once a priority region at this broad scale has been identified, more geographically restricted, higher resolution analyses could be conducted using realistic land management units.

2.2 *Features*

Features are used to characterize areas. The contribution of different areas to the overall goal of representing biodiversity is measured by the list, or the list and spatial extent or abundance, of the biodiversity features within those areas. Features may be taxa (or, in principle, the characters taxa represent), environmental variables, species

assemblages, or environmental classes, or various combinations of these. All of these features are surrogates in the sense that they are supposed to stand in for overall biodiversity and are therefore abstractions of biodiversity to a greater or lesser extent. Running counter to the scale of precision, progressively higher levels have the advantage of being more practical for measurement. Thus for sectoral analyses of small well-known taxa (e.g. fruit bats, Mickleburgh *et al* 1992) the more direct taxonomic measures may be feasible. However, for analyses of a larger portion of overall biodiversity, a more remote surrogate such as vegetation type or environmental diversity, or some combination of biotic and environmental measures is likely to be the best practical solution (Margules and Pressey 2000; Nix *et al* 2000). Higher scales also integrate ecological functions, including the processes that help maintain ecosystem viability (McKenzie *et al* 1989).

3. Assumptions in the use of data

3.1 *Spatial consistency*

All biological data sets are (better or worse) samples of the geographical space they were taken from and inevitably incorporate some degree of spatial bias. Yet, setting biodiversity priorities means comparing areas with one another, and valid comparisons cannot be made unless the same relationship between sample and population can be assumed to hold for all areas being compared. This is a universal problem in land use planning and decision-making, and is especially acute in biodiversity planning because the relevant information is expensive to acquire.

Often, field observations or samples are taken in an opportunistic way. The taxa recorded are the ones of interest to the collector, and the field sites where they were looked for are in places where she or he might expect to find those taxa, or are conveniently accessible. Sample sites with records tend to be subsets of sites where the species actually occur and there are usually few, if any, records of where they were looked for but not found; that is, sites with recorded absences (Margules and Austin 1994). One way to improve the consistency of coverage of existing data sets is to model the wider spatial distribution patterns of taxa or assemblages recorded during surveys or collecting expeditions. Some common spatial modelling techniques, which utilize environmental variables as predictors, are described below. However, if spatial models are to be used to improve existing data sets, the degree of bias, both geographical and environmental, needs to be determined.

3.2 *Data precision and accuracy*

Errors come in many forms. Misidentification of both attributes and areas may be common (e.g. species or areas may have been misidentified, or the same name may be shared by two or more species or areas, or may have changed over time). Some assessment of the accuracy of existing data sets is essential, and this is likely to be a time-consuming and tedious process. A thorough treatment of data standards is beyond the scope of this paper. However, it is possible to minimize errors by checking new data for consistency with existing data and implementing routines for detecting unexpected outliers.

In Australia, the Environmental Resources Information Network (ERIN) has developed a number of such procedures and routines (Chapman and Busby 1994). All names are checked against a master file as they are loaded into the database. Records that do not match are returned to the custodian of the data for checking. Non-current names, when they are recognized, are changed to current names, and any remaining unrecognized names that cannot be resolved are flagged to indicate a taxonomic problem. The option exists, then, to exclude these from future analyses.

Checking geographic locations is more problematic. ERIN adopted the innovative method of detecting outliers using climate profiles modelled for each species. Field records are matched with climatic attributes generated with the BIOCLIM software (Busby 1986; Nix 1986; Busby 1991; and see below). Any records that lie on or beyond the margins of the climatic profile are queried and re-checked.

Errors of omission are inevitable. Usually, it has to be assumed that the broader patterns detected in data sets are not artefacts, but represent real patterns that are sufficiently robust to show through the effects of confounding factors.

3.3 *Sampling bias*

The effort used to record attributes in the field varies with different collections and is another major source of bias (Rich 1997). The relationship between number of species recorded and sampling effort, the species-discovery curve, generally increases steeply at first, but with progressively larger samples becomes less steep as new discoveries become less frequent (Magurran 1988; Colwell and Coddington 1994). However, with movement of individuals and populations, species' ranges shift continuously so that there may be no asymptote to an absolute fixed value for local species number. Consequently, any count has to be measured relative to sampling effort. An extreme example was given by Grinnell (1922,

p. 375), who calculated that by sampling just within California, every species of North American bird could be recorded over a period of 410 years.

Ideally, sampling effort should be perfectly uniform, so that all recorded variations in distribution or abundance patterns are real, and not an artefact of variation in sampling effort. This is especially important if the number of samples is small and so falls on the first, steep, part of the species-discovery curve. Species of small body size, low density or low apparency to collecting methods (e.g. organisms that burrow deep in the soil) are especially likely to be missed. Sampling can also be confounded by differences between species in seasonal (phenological) or daily (diurnal) patterns of migration, activity, or apparency. In order to minimize this effect; Gibbons *et al* (1993) used mean counts from time-limited sub-samples (2 km × 2 km squares) for estimating the number of British birds in 10 km × 10 km squares. Potential sources of bias due to characteristics of taxa should be taken into account when choosing biodiversity surrogates from among existing data sets. Choosing taxa that are less susceptible to sampling variation can help minimize this bias.

The assumption of uniform sampling effort is almost always violated to a greater or lesser extent. In practice, the multitude of possible contingent variables cannot be adequately controlled for in surveys, particularly in voluntary recording schemes. This problem is compounded in compilations of existing data. Therefore, some adjustment for variation in sampling effort is almost always desirable when dealing with existing data sets.

3.4 *Turnover in space and time*

Field records from surveys and collecting expeditions are 'snap-shots' in time. Many existing data sets were recorded over one brief time period. Compilations of data recorded at different times can increase the temporal range of records, but records taken at different times may not necessarily have been taken from the same place. Compilations from point records to broader mapping units such as grid cells may also obscure spatial turnover. These two factors compound at least three related problems of estimating distribution patterns: spatial heterogeneity, or species turnover, at a fine scale may be masked when occupancy is recorded at a broader spatial scale; field records may include vagrants as well as breeding populations; and patterns of occupancy, even for breeding populations, may change with time.

Distributions of species, populations, and assemblages may be dynamic on a range of time scales. The immediate need for priority areas selection is to estimate the current distribution patterns, not historical patterns. Nevertheless, information on historical and predicted future distributions, where available, should be used in priority setting because distribution patterns do change with environmental variation, both seasonally and over longer time periods. Even in the short term, occupancy of an area by a population may be both discontinuous in time and dependent on the security of other suitable areas (Hanski 1994; Gaston *et al* 2002). This may depend at least in part on certain kinds and levels of disturbance, such as fire, flood, tree fall, etc. Very mobile species, such as some fish and birds, may breed in one area but depend for their survival on distant feeding and overwintering grounds.

The viability of populations within any given area can be regarded as a management problem rather than a data selection problem. Indeed it has to be, if sets of presence data are to be used for biodiversity priority setting. However, in choosing and treating existing data sets, temporal bias and the dynamics of populations should be borne in mind. At the very least, breeding and non-breeding records should be distinguished whenever possible (Williams *et al* 1996a).

4. Sources of existing data

There are many organizations throughout the world holding data that can validly be used in an analysis of biodiversity priorities. Existing biological data can be extracted from collections in museums and herbariums, from various departments of government such as natural resource management agencies, and from non-government organizations. Examples of these agencies include the World Conservation Monitoring Centre (WCMC), the Australian Environmental Resources Information Network (ERIN), the UK Biological Records Centre (BRC), the British Trust for Ornithology (BTO), BirdLife International, The Nature Conservancy (USA), the US Gap Analysis Project, etc.

Even before bias can be assessed and decisions taken either to proceed with existing data, model expected data (see below), collect new data, or reject the data, those data have to be extracted from a source. This is usually a time-consuming and labour-intensive first step in biodiversity priority area selection. There is no way to avoid this step and there are no short cuts, although, fortunately, digitized data sets are becoming more widely available.

Ready access to existing data sets is not always possible. Custodians may place restrictions on access to their data, or charge for the use of data. In some cases, access is restricted to protect the locations of rare or threatened species (e.g. Gibbons *et al* 1993). In other cases, custodians may wish to protect ownership for research purposes. Belbin *et al* (1994), in a review of

ERIN data holdings in Australia, recognized four categories of data access: unrestricted access, formal acknowledgement required, permission required (selective or incomplete access), and confidential. They found that 51% were restricted, confidential, or requiring permission, and 49% were unrestricted.

Environmental data are more widely available, more accessible, and generally exist in a more consistent form than most biological data. Environmental data alone can be used as surrogates for biodiversity and they are required for any formal modelling of wider distribution patterns of species or populations from the point records that field collections represent (see below). Environmental data fall into the three broad categories of terrain, climate and substrate. Terrain refers to surface morphology and includes parameters such as elevation, slope, relief, and aspect. These parameters can be recorded from topographic maps if the resolution is sufficiently fine, but in most cases it will be more appropriate to interpolate them from Digital Elevation Models (DEMs). DEMs are not yet routinely available or as accurate as topographic maps. The construction of a DEM can be time-consuming and is technically demanding (Hutchinson 1991). However, DEMs allow consistent and repeatable interpolation across whole regions and should constitute the necessary first step in generating environmental surfaces (Hutchinson 1993). Climate data are available from national meteorological bureau(s), but may have to be digitized for spatial modelling. Climate data can be interpolated spatially with the aid of DEMs by fitting surfaces as smooth tri-variate functions of latitude, longitude, and elevation (Hutchinson 1995). Climate data can sometimes be augmented with data collected by forestry, conservation, agriculture, and water resource agencies. Physical and chemical substrate data may be the most difficult to obtain. However, substrate mapping has been completed in a number of countries, regions, and biomes. For example, thematic maps of lithological substrate, soils, and landforms may be available.

Maps are a popular and efficient (measured both as information/ink ratio and speed of communication) way of summarizing and communicating existing data (Tufte 1990). Two classes of map relevant to computer aided (GIS) mapping are referred to as vector maps and raster maps (Burrough 1986). The former defines areas by joining boundary points as polygons, often of irregular shape and size. The latter defines a regular array or grid of areas of similar shape and size. It has been claimed that vector maps allow data to be presented at any scale. However, they merely give the impression of lacking scale, which is limited, as always, by the number (resolution) and precision of data points. What is important is that data are registered with precision, either as 'points' with small errors, or on fine grids, so that they can be used subsequently in analyses at a broad range of spatial scales depending on the questions to be addressed. In this regard, there is no fundamental difference between vector maps and raster maps.

5. Data treatments to reduce bias

Extrapolation of feature richness (e.g. Sanders 1968; Hurlbert 1971; Margules *et al* 1987; Palmer 1990; Prendergast *et al* 1993; Colwell and Coddington 1994), procedures for smoothing feature richness across neighbourhoods (e.g. Eversham *et al* 1992; Lawton *et al* 1994; Williams and Gaston 1998) and procedures for the treatment of spatial autocorrelation (e.g. Pearson and Carroll 1998) lose information on feature identity. Consequently, they are not suitable for use with the priority area selection methods described by Margules *et al* (2002). Fortunately, a number of analytical procedures which retain feature identity are available, to address the problems of spatial bias in existing data sets. An exhaustive review is beyond the scope of this paper. However, some of the more common methods are described briefly below (with references to detailed examples of applications taken from Austin 1994 and Austin *et al* 1994). Many spatial modelling techniques require spatial models of environmental variables, commonly the predictors, to be constructed first.

Procedures for estimating wider spatial patterns from point records assume that each species occupies a unique niche (Hutchinson 1957), which may not be easily predicted from that of other species. Thus species distribution patterns are most accurately defined in multi-dimensional environmental (niche) space, with the resultant spatial pattern showing high or dense populations in scattered locations representing the most favourable habitat, and lower, more sparse, populations in areas of less favourable habitat. Problems may arise with these models when organisms are restricted to just part of their potentially suitable habitat by barriers to dispersal and low dispersal abilities (e.g. vicariance model of biogeography: Nelson and Platnick 1981), or by non-equilibrium dynamics of metapopulations. Plant species in the hyper-diverse Cape region, South Africa (Cowling 1992), and in south-western Western Australia, may be cases in point. Species may also be restricted to sub-optimal habitat by practices such as land clearing, grazing, and altered fire regimes. It is usually difficult to measure the extent of such changes and incorporate them in spatial models.

5.1 *Heuristic models*

One of the more widely used spatial modelling techniques is BIOCLIM (Nix 1986; Busby 1991). BIOCLIM

works by generating climatic indices for the locations of field records of species and then finding other locations with the same or similar indices, which thus map potential distribution patterns. The first step is to model monthly mean minimum and maximum temperature and precipitation for the entire region of interest. This is done by fitting a function of three independent variables, latitude, longitude, and elevation, to climatic values measured at climate stations. The function is a tri-variate spline, with spatial interpolation made by use of a DEM (Hutchinson 1991, 1995). The next step is to derive climatic indices from these surfaces. The current published version of BIOCLIM contains 16 indices (table 1, Busby 1991). Other versions have been developed using a larger number of indices, including solar radiation. The climatic indices for sites at which a species (or other feature such as vegetation type) has been recorded can then be aggregated into a climatic profile for that species. This profile is expressed as the extremes, and the 5 and 95 percentile values. If an unrecorded site falls within the climate profile of all 16 indices for a species, then that site has a potentially suitable climate for that species.

Because this method uses presence only data, it is restricted to estimating the geographic range of a species. It cannot predict absences within that range. Large areas may be included as potential occurrence space due to the presence of a single extreme observation (Walker and Cocks 1991; Carpenter *et al* 1993). Thus the method is limited to estimating potential distributions. There are no quantitative predictions or statistical tests and confidence limits.

On the other hand, presence only data are by far the most common form of existing biological field records. BIOCLIM is an innovative, technically sophisticated tool for adding value to existing data of this kind and reducing its inherent spatial bias. Published BIOCLIM applications include the spatial estimation of the distribution patterns of elapid snakes (Nix 1986), C_3 and C_4 grasses (Prendergast and Hattersley 1985), temperate rainforest tree species (Busby 1986; Hill et al 1988), weeds (Panetta and Dodd 1987), *Eucalyptus* trees (Williams 1991), and rainforest vertebrates (Nix and Switzer 1991).

Table 1. The sixteen climate profile parameters employed in BIOCLIM Version 2·0 (from Busby 1991).

Parameter	Unit
Annual mean temperature	°C
Minimum temperature of coolest month	°C
Maximum temperature of warmest month	°C
Annual temperature range (3–2)	°C
Mean temperature of coolest quarter	°C
Mean temperature of warmest quarter	°C
Mean temperature of wettest quarter	°C
Mean temperature of driest quarter	°C
Annual mean precipitation	mm
Precipitation of wettest month	mm
Precipitation of driest month	mm
Coefficient of variation of monthly precipitation	–
Precipitation of wettest quarter	mm
Precipitation of driest quarter	mm
Precipitation of coolest quarter	mm
Precipitation of warmest quarter	mm

HABITAT (Walker and Cocks 1991) is also an heuristic method. It can work with a variety of environmental data, including climatic indices derived from BIOCLIM. In addition, it can use categorical attributes such as soil type and geological substrate. HABITAT is similar in concept to BIOCLIM in that it generates an 'environmental envelope' from which potential distribution is determined. Unlike BIOCLIM, it treats the attributes as interdependent influences on potential distribution. It uses linear programming to describe the envelope occupied as a function of the linear combination of the set of attributes.

Another heuristic method is DOMAIN (Carpenter *et al* 1993), which differs from BIOCLIM in that it measures the climatic similarity between candidate sites and the nearest sites with records, using the Gower metric (Gower 1971). A continuous function is generated so that degrees of similarity can be selected (Austin *et al* 1994).

5.2 *Regression models*

Regression models are appropriate for data of the presence/absence kind. That is, if a species (or community, assemblage) is recorded in the areas × features matrix as absent, it has been looked for in that area and not found. Regression models are statistical correlations of the observed presence or absence of a feature with variables which predict wider spatial distribution patterns. They use the same conceptual framework (broad scale distribution controlled by environmental variables) as BIOCLIM and similar heuristic models. Environmental variables such as rainfall, temperature, lithological substrate, and aspect (to estimate solar radiation), are fitted to presence and absence records of features such as species recorded from field sites. The probability (along with confidence limits) of finding the feature at unrecorded sites is calculated for different combinations of these predictor variables from the parameters of the model.

The most commonly used regression models in ecology and biogeography are Generalised Linear Models (GLM) (McCullagh and Nelder 1989; Crawley 1993; and see Austin *et al* 1990; Lindenmayer *et al* 1991a,b; and Nicholls 1989, 1991, for examples of ecological applications), and Generalised Additive Models (GAM)

(Hastie and Tibshirani 1990; see Yee and Mitchell 1991 for an ecological application). Austin and Meyers (1996) compared the performance of these two methods in predicting the distribution patterns of Eucalypts in south-eastern Australia. They concluded that, while there was little to choose between them, and that application of ecological knowledge and experience in interpreting the models is essential for both, GAM had the technical advantage of greater flexibility in fitting response surfaces.

Almost all existing data sets in museums and herbariums are unsuitable for regression modelling because the lack of a record in any particular area does not necessarily signify a real absence. Data sets have to be examined carefully for sampling bias before being subjected to regression analysis. If new data are to be collected in a survey to assist with biodiversity priority setting, then the survey should be designed and conducted in such a way that the records obtained can be spatially interpolated with regression models.

5.3 *Multivariate interpolation*

Multivariate clustering to define and map communities or assemblages is a tool that can be used to interpolate sparse data sets, but with some loss of detail. Many data sets throughout the world are sparse in their geographic coverage because they come from very large areas. The total number of records of many species may be too low for the species to be modelled individually with any confidence. Information on co-occurrence is used to group sample sites (e.g. quadrats), or the species, and the resulting communities or assemblages can be mapped. Clustering is a tool for reducing complexity in large data sets for the practical purposes of description and communication. Anderberg (1973), Sneath and Sokal (1973) and Jongman *et al* (1987) are comprehensive standard texts. Faith (1991) reviewed pattern analysis methods for nature conservation, and Belbin (1987) provides an algorithm and an example of non-hierarchical clustering for very large data sets.

5.4 *Computational methods*

There are at least three computer induction methods which may prove to be useful for modelling spatial distribution patterns of species or assemblages from presence/absence data: decision trees, neural nets, and genetic algorithms. They differ from the methods above primarily in that they are algorithms that are not easily

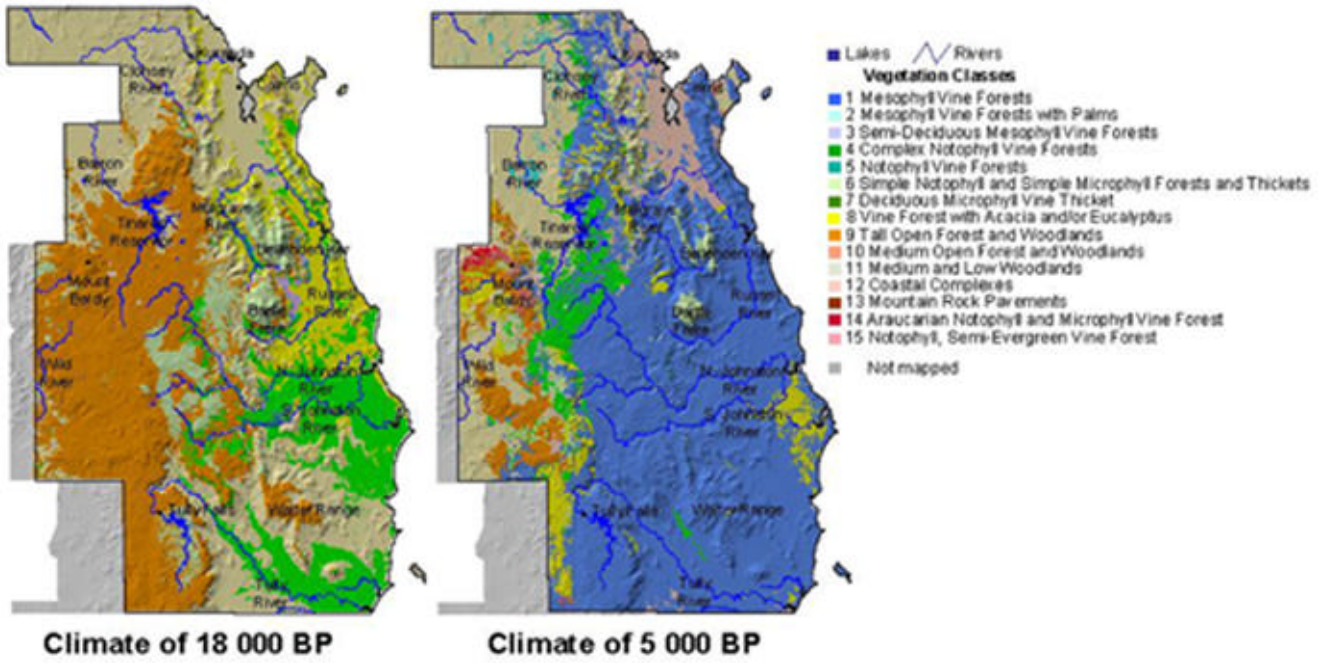

Figure 3. Potential distributions of fifteen forest structural classes (Hilbert and van den Muyzenberg 1999) in the central Wet Tropics of Australia, modelled using an artificial neural network and estimates of two past climates. Note the great expansion of potential rainforest distributions in the warm wet climate of 5000 BP compared to the dry, open woodland and tall open forests that dominated at the last glacial maximum (c. 18000 BP).

described by specific mathematical functions. Lees (1994) and Austin *et al* (1994) provide introductory reviews. Decision trees are described by Breiman *et al* (1984). Data sets are split recursively into sub-sets so as to maximize the prediction of a dependent variable in a sub-set, creating a binary tree of decision rules. Examples of ecological and biogeographic applications can be found in Moore *et al* (1991), Stockwell *et al* (1990), and Walker and Cocks (1991). Neural nets are described by Rumelhart and McClelland (1986) and several introductory texts exist (e.g. Aleksander and Morton 1990; Caudill 1990). They are a form of artificial intelligence, which learns to predict a pattern from a training set. Useful properties of neural networks are that they do not assume any *a priori* distribution (Brown *et al* 1998), they are not seriously biased by outliers (Haung and Lippman 1988), and they can approximate an arbitrary function to any desired degree of accuracy (Hecht–Nielson 1991). They have frequently been used to classify remotely sensed images (Fitzgerald and Lees 1993; Lees 1994) and have been applied to various other purposes in ecology (e.g. Paruelo and Tomasel 1997; Deadman and Gimblett 1997). Hilbert and van den Muyzenberg (1999) used an artificial neural network to model the distributions of fifteen forest classes in a tropical landscape with 75% accuracy at a one hectare resolution. Using a variety of techniques (Hilbert and Ostendorf 2001), this model has been used to estimate the pre-clearing distribution of forest types, the distribution of rainforests from the last glacial maximum (c. 18000 BP) to the present, analyse the sensitivity of rainforests to climate change in the future (Hilbert *et al* 2001), the locations of Pleistocene rainforest refugia, and interglacial refugia of wet sclerophyll forests (Hilbert *et al* 2000). An example of modelled potential distributions of forest types in the drier climate of the late glacial maximum and the wetter climate of 5000 BP is shown in figure 3. Genetic algorithms are described by Holland (1992). They use the idea that evolution solves the problem of survival by constantly testing and re-testing the fitness of individuals through mutation and genetic recombination. Thus, they are essentially a minimization method that can be applied to any optimization problem. There are few ecological examples, but Stockwell and Noble (1992) describe the use of genetic algorithms to model animal distributions.

5.5 *Rejecting data*

Compilations of existing data that meet the assumptions required for a priority areas analysis can be used directly. Data that can be made to meet or approximate those assumptions, as outlined above, can be used after treatment. However, the end value of any treatment of existing data will always be limited, especially if the data were collected for other purposes, such as taxonomic description. If existing data contain a strong bias of unknown direction, then no amount of treatment can remove it. For example, if sampling effort among areas is very uneven, an empirical or statistical model relating field samples to environmental variables is likely to be unrepresentative and misleading. Data, which still fail to meet the necessary assumptions, must be rejected. In that case, either the task at hand has to be abandoned, a different question has to be asked of the data, or new data have to be collected.

6. Summary

Existing data should be evaluated in the light of a precisely defined goal. For identifying biodiversity priority areas, the basic requirement is an areas × features matrix. Choices have to be made concerning which areas to include and which features (the biodiversity surrogates) to use to describe those areas. Existing data should then be examined to see if they are appropriate. The choice of features will generally be a compromise from among a range of possible biodiversity surrogates, because it is not possible to measure biodiversity directly. Many institutions throughout the world hold relevant data referenced to areas. However, these data were often collected for purposes other than biodiversity conservation planning. Treatment of compilations of existing data is likely to be necessary before they meet the assumptions required for a priority areas analysis.

References

Aleksander I and Morton H 1990 *An introduction to neural computing* (London: Chapman and Hall)

Anderberg M R 1973 *Cluster analysis for applications* (New York: Academic Press)

Austin M P 1994 *Modelling of landscape patterns and processes using biological data. Sub-project 3: data capability* (Consultancy report for ERIN) (Canberra: CSIRO Division of Wildlife and Ecology)

Austin M P and Heyligers P C 1989 Vegetation survey design for conservation: gradsect sampling of forests in north-eastern New South Wales; *Biol. Conserv.* **50** 13–32

Austin M P and Meyers J A 1996 Current approaches to modelling the environmental niche of eucalyptus: implications for management of forest biodiversity; *For. Ecol. Manag.* **85** 95–106

Austin M P, Meyers J A and Doherty M D 1994 *Modelling of landscape patterns and processes using biological data. Sub-project 2: predictive models for landscape patterns and processes* (Consultancy report for ERIN) (Canberra: CSIRO Division of Wildlife and Ecology)

Austin M P, Nicholls A O and Margules C R 1990 Measurement of the realized qualitative niche: environmental niches of five *Eucalyptus* species; *Ecol. Monogr.* **60** 161–177

Bedward M, Pressey R L and Keith D A 1992 A new approach for selecting fully representative reserve networks: addressing efficiency, reserve design and land suitability with an iterative analysis; *Biol. Conserv.* **62** 115–125

Belbin L 1987 The use of non-hierarchical allocation methods for clustering large sets of data; *Aust. J. Ecol.* **19** 32–39

Belbin L, Austin M P, Margules C R, Cresswell I D and Thackway R 1994 *Modelling of landscape patterns and processes using biological data. Sub-project I: data suitability* (Consultancy report for ERIN) (Canberra: CSIRO Division of Wildlife and Ecology)

Brieiman L, Friedman J H, Olshen R A and Stone C J 1984 *Classification and regression trees* (Belmont, California: Wadsworth International Group)

Brown D G, Lusch D P and Duda K A 1998 Supervised classification of types of glaciated landscapes using digital elevation data; *Geomorphology* **21** 233–250

Burrough P A 1986 *Principles of Geographic Information Systems for Land Resources Assessment* (Oxford: Clarendon Press)

Busby J R 1986 A biogeoclimatic analysis of *Nothofagus cunninghamii* (Hook.) Oerst. in southeastern Australia; *Aust. J. Ecol.* **11** 1–7

Busby J R 1991 BIOCLIM – a bioclimatic analysis and prediction system; in *Nature conservation: cost effective biological surveys and data analysis* (eds) C R Margules and M P Austin (Melbourne: CSIRO) pp 64–68

Carpenter G, Gillison A N and Winter J 1993 DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals; *Biodiver. Conserv.* **2** 667–680

Caudill M 1990 *A1 EXPERT: neural networks primer* (San Francisco: Miller Freeman)

Caughley G 1994 Directions in conservation biology; *J. Anim. Ecol.* **63** 215–224

Chapman A D and Busby J R 1994 Linking plant species information to continental biodiversity, inventory, climate modelling and environmental monitoring; in *Mapping the diversity of nature* (ed.) R I Miller (London: Chapman and Hall) pp 179–195

Colwell R K and Coddington J A 1994 Estimating terrestrial biodiversity through extrapolation; *Philos. Trans. R. Soc. London* **B345** 101–118

Cowling R M (ed.) 1992 *The ecology of fynbos* (Cape Town: Oxford University Press)

Crawley M J 1993 *GLIM for ecologists* (Oxford: Blackwell)

Deadman P J and Gimblett H R 1997 Applying neural networks to vegetation management plan development; *AI Appl.* **11** 107–112

Eversham B C, Harding P T, Loder N, Arnold H R and Fenton R W 1992 Research applications using data from species surveys in Britain; in *Faunal inventories of sites for cartography and nature conservation* (eds) J L van Goethem and P Grootaert (Brussels: Proceedings of the 8th International Colloquium of the European Invertebrate Survey) pp 29–40

Faith D P 1991 Effective pattern analysis methods for nature conservation; in *Nature conservation: cost effective biological surveys and data analysis* (eds) C R Margules and M P Austin (Melbourne: CSIRO) pp 47–53

Fitzgerald R W and Lees B G 1993 Assessing the classification accuracy of multicourse remote sensing data; *Remote Sensing Environ.* **47** 1–25

Gaston K J, Pressey R L and Margules C R 2002 Persistence and vulnerability: retaining biodiversity in the landscape and in protected areas; *J. Biosci.* (*Suppl. 2*) **27** 361–384

Gibbons D W, Reid J B and Chapman R A 1993 *The new atlas of breeding birds in Britain and Ireland: 1988–1991* (London: Poyser)

Gillison A N and Brewer K R W 1985 The use of gradient directed transects or gradsects in natural resource surveys; *J. Environ. Manag.* **20** 103–127

Gower J C 1971 A general coefficient of similarity and some of its properties; *Biometrics* **27** 857–871

Grinnell J 1922 The role of the 'accidental'; *Auk* **39** 373–380

Hanski I 1994 Patch-occupancy dynamics in fragmented landscapes; *Trends Ecol. Evol.* **9** 131–135

Hastie T and Tibshirani R 1990 *Generalised additive models* (London: Chapman and Hall)

Haung W Y and Lippman R P 1988 Comparisons between neural net and conventional classifiers; *Proc. Int. Joint Conf. Neural Networks, San Diego* **4** 485–493

Hecht-Nielson R 1991 *Neurocomputing* (Reading, Massachusetts: Addison-Wesley)

Higgs A J and Usher M B 1980 Should nature reserves be large or small?; *Nature* (*London*) **258** 586

Hilbert D W, Graham A and Parker T 2000 *Forest and woodland habitats of the Northern Bettong* (*Bettongia tropica*) *in the past, present, and future* (Report prepared for the Queensland Parks and Wildlife Service) (Atherton: CSIRO)

Hilbert D W and Ostendorf B 2001 The utility of empirical, artificial neural network approaches for modelling the distribution of regional to global vegetation in past, present and future climates; *Ecol. Modelling* **146** 311–327

Hilbert D W, Ostendorf B and Hopkins M S 2001 Sensitivity of tropical forests to climate change in the humid tropics of North Queensland; *Aust. Ecol.* **26** 590–603

Hilbert D W and van den Muyzenberg J 1999 Using an artificial neural network to characterise the relative suitability of environments for forest types in a complex tropical vegetation mosaic; *Diversity Distrib.* **5** 263–274

Hill R S, Read J and Busby J R 1988 The temperature-dependence of photosynthesis of some Australian temperate rainforest trees and its biogeographical significance; *J. Biogeogr.* **15** 431–439

Holland J H 1992 Genetic algorithms; *Sci. Am.* **267** 44–50

Hurlbert S H 1971 The non-concept of species diversity: a critique and alternative parameters; *Ecology* **52** 577–586

Hutchinson G E 1957 Concluding remarks; *Cold Spring Harbor Symp. Quant. Biol.* **22** 415–427

Hutchinson M F 1991 Application of thin plate smoothing splines to continent-wide data assimilation; in *Data assimilation systems* (ed.) J D Jasper (Melbourne: Bureau of Meteorology) pp 104–113

Hutchinson M F 1993 Development of a continent-wide DEM with applications to terrain and climate analysis; in *Environmental Modelling with GIS* (eds) M F Goodchild, B O Parks and L T Steyaert (New York: Oxford University Press) pp 392–399

Hutchinson M F 1995 Interpolating mean rainfall using thin plate smoothing splines; *Int. J. Geogr. Inform. Syst.* **9** 385–403

IUCN (The World Conservation Union) 1994 *Guidelines for protected area management categories* (Gland, Switzerland: IUCN)

Jalas J and Suominen J (eds) 1972, 1973, 1976, 1979, 1980, 1983, 1986, 1989, 1991, 1994 *Atlas Florae Europaeae* Volumes 1–10 (The Committee for Mapping the Flora of Europe and Societas Biologica Fennica Vanamo)

Jongman R T H, ter Braak C J F and van Tongeren O F 1987 *Data analysis in community and landscape ecology* (Wagenigen: Pudoc)

Lawton J H, Prendergast J R and Eversham B C 1994 The numbers and spatial distributions of species: analyses of British data; in *Systematics and conservation evaluation* (eds) P L Forey, C J Humphries and R I Vane-Wright (Oxford: Oxford University Press) pp 177–195

Leclercq J 1979 Tous ces atlas, toutes ces cartes, c'est pour quoi faire?; *Notes Fauniques Gembloux* **2** 1–22

Lees B G 1994 Decision trees, artificial neural networks and genetic algorithms for classification of remotely sensed and ancillary data; in *Proceedings of the 7th Australian Remote Sensing Conference* (Floreat, Western Australia: Remote Sensing and Photogrammetry Association Australia) pp 51–60

Lindenmayer D B, Cunningham R B, Nix H A, Tanton M T and Smith A P 1991a Predicting the abundance of hollow-bearing trees in montane forests of southeastern Australia; *Aust. J. Ecol.* **16** 91–98

Lindenmayer D B, Cunningham R B, Tanton M T, Nix H A and Smith A P 1991b The conservation of arboreal marsupials in the montane ash forests of the Central Highlands of Victoria, south-east Australia. III. The habitat requirements of Lead-beater's possum, *Gymnobelideus leadbeateri*, and models of the diversity and abundance of arboreal marsupials; *Biol. Conserv.* **56** 295–315

Longmore R (ed.) 1986 *Atlas of elapid snakes of Australia* (Canberra: Australian Government Publishing Service)

Magurran A E 1988 *Ecological diversity and its measurement* (London: Croom Helm)

Maling D H 1974 Personal projections; *Geograph. Mag.* **46** 599–600

Margules C R and Austin M P 1994 Biological models for monitoring species decline: the construction and use of data bases; *Philos. Trans. R. Soc. London* **B344** 69–75

Margules C R, Nicholls A O and Austin M P 1987 Diversity of *Eucalyptus* species predicted by a multi-variable environmental gradient; *Oecologia* **71** 229–232

Margules C R and Pressey R L 2000 Systematic conservation planning; *Nature* (*London*) **405** 243–253

Margules C R, Pressey R L and Williams P H 2002 Representing biodiversity: data and procedures for identifying priority areas for conservation; *J. Biosi.* (*Suppl. 2*) **27** 309–326

Margules C R, Redhead T D, Faith D P and Hutchinson M F 1995 *BioRap: Guidelines for using the BioRap methodology and tools* (Canberra: CSIRO)

McAllister D E, Schueler F W, Roberts C M and Hawkins J P 1994 Mapping and GIS analysis of the global distribution of coral reef fishes on an equal-area grid; in *Mapping the diversity of nature* (ed.) R I Miller (London: Chapman and Hall) pp 155–175

McCullagh P and Nelder J A 1989 *Generalised linear models* 2nd edition (London: Chapman and Hall)

McKenzie N L, Belbin L, Margules C R and Keighery G J 1989 Selecting representative reserve systems in remote areas: a case study in the Nullarbor Region, Australia; *Biol. Conserv.* **50** 239–261

Mickleburgh S P, Hutson A M and Racey P A 1992 *Old World fruit bats: an action plan for their conservation* (Gland, Switzerland: IUCN)

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; *Environ. Manag.* **15** 59–71

Nelson G and Platnick N I 1981 *Systematics and biogeography*: *cladistics and vicariance* (New York: Columbia University Press)

Nicholls A O 1989 How to make biological surveys go further with Generalised Linear Models; *Biol. Conserv.* **50** 51–75

Nicholls A O 1991 Examples of the use of Generalised Linear Models in analysis of survey data for conservation evaluation; in *Nature conservation: cost effective biological surveys and data analysis* (eds) C R Margules and M P Austin (Melbourne: CSIRO) pp 54–63

Nix H A 1986 A biogeographic analysis of Australian elapid snakes; in *Atlas of elapid snakes of Australia* (ed.) R Longmore (Canberra: Australian Government Publishing Service) pp 4–15

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* (Canberra: Centre for Resource and Environmental Studies, Australian National University)

Nix H A and Switzer M A (eds) 1991 *Rainforest animals: atlas of vertebrates endemic to Australia's wet tropics Kowari* (1) (Canberra: Australian National Parks and Wildlife Service)

Palmer M W 1990 The estimation of species richness by extrapolation; *Ecology* **71** 1195–1198

Panetta F D and Dodd J 1987 Bioclimatic prediction of the potential distribution of skeleton weed, *Chondrilla juncea* L., in Western Australia; *J. Aust. Inst. Agric. Sci.* **53** 11–16

Paruelo J M and Tomasel F 1997 Prediction of functional characteristics of ecosystems: a comparison of artificial neural networks and regression models; *Ecol. Model.* **98** 173–186

Pearson D L and Carroll S S 1998 Global patterns of species richness: spatial models for conservation planning using bio-indicator and precipitation data; *Conserv. Biol.* **12** 809–821

Pekkarinen A, Teräs I, Viramo J and Paatela J 1981 Distribution of bumblebees (Hymenoptera, Apidae: *Bombus* and *Psithyrus*) in eastern Fennoscandia; *Not. Entomol.* **61** 71–89

Prendergast H D V and Hattersley P W 1985 Distribution and cytology of Australian *Neurachne* and its allies (Poaceae), a group containing C_3, C_4 and C_3–C_4 intermediate species; *Aust. J. Bot.* **33** 317–336

Prendergast J R and Eversham B C 1997 Species richness covariance in higher taxa: empirical tests of the biodiversity indicator concept; *Ecography* **20** 210–216

Prendergast J R, Wood S N, Lawton J H and Eversham B C 1993 Correcting for variation in recording effort in analyses of diversity hotspots; *Biodiver. Lett.* **1** 39–53

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 Ecol. Evol.* **8** 124–128

Pressey R L and Nicholls A O 1989 Efficiency in conservation evaluation: scoring versus iterative approaches; *Biol. Conserv.* **50** 199–218

Rich T C G 1997 Is *ad hoc* good enough?; *Trans. Suffolk Nat. Hist. Soc.* **33** 14–21

Rumelhart D E and McClelland J L 1986 *Parallel distributed processing: explorations in the microstructures of cognition* (Cambridge, Massachusetts: MIT Press)

Sanders H L 1968 Marine benthic diversity: a comparative study; *Am. Nat.* **102** 243–282

Scott J M, Davis F, Csuti B, Noss R, Butterfield B, Groves C, Anderson H, Caicco S, D'Erchia F, Edwards T C, Ulliman J

and Wright R G 1993 Gap analysis: a geographic approach to protection of biological diversity; *Wildl. Monogr.* **123** 1–41

Sneath P H A and Sokal R R 1973 *Numerical taxonomy* (San Francisco: W H Freeman)

Stockwell D R B and Noble I R 1992 Induction of sets of rules from animal distribution data: a robust and informative method of data analysis; *Math. Compu. Simul.* **33** 385–390

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; *A1 Appl.* **4** 33–43

Tufte E R 1990 *Envisioning information* (Cheshire, Connecticut: Graphics Press)

Walker P A and Cocks K D 1991 HABITAT: a procedure for modelling a disjoint environmental envelope for a plant or animal species; *Global Ecol. Biogeogr. Lett.* **1** 108–118

Wessels K J, van Jaarsveld A S, Grimbeek J D and van der Linde M J 1998 An evaluation of the gradsect biological survey method; *Biodiver. Conserv.* **7** 1093–1121

White D, Kimerling A J and Overton W S 1992 Cartographic and geometric components of a global sampling design for environmental monitoring; *Cartogr. Geogr. Inform. Syst.* **19** 5–22

Williams J E 1991 Biogeographic patterns of three sub-alpine eucalypts in south-east Australia with special reference to *Eucalyptus pauciflora*, Sieb ex Spreng; *J. Biogeogr.* **18** 223–230

Williams P H 1991 An annotated checklist of bumble bees with an analysis of patterns of description (Hymenoptera: Apidae, Bombini); *Bull. Br. Mus. Nat. Hist. Entomol.* **67** 79–152

Williams P H and Gaston K J 1994 Measuring more of biodiversity: can higher-taxon richness predict wholesale species richness?; *Biol. Conserv.* **67** 211–217

Williams P H and Gaston K J 1998 Biodiversity indicators: graphical techniques, smoothing and searching for what makes relationships work; *Ecography* **21** 551–560

Williams P, Gibbons D, Margules C, Rebelo A, Humphries C and Pressey R 1996a A comparison of richness hotspots, rarity hotspots and complementary areas for conserving diversity using British birds; *Conserv. Biol.* **10** 155–174

Williams P H, Prance G T, Humphries C J and Edwards K S 1996b Promise and problems in applying quantitative complementary areas for representing the diversity of some Neotropical plants (families Dichapetalaceae, Lecythidaceae, Caryocaraceae, Chrysobalanaceae and Proteaceae); *Biol. J. Linnean Soc.* **58** 125–157

Yee T W and Mitchell N D 1991 Generalised additive models in plant ecology; *J. Veg. Sci.* **2** 587–602